Challenges in Clinical Metaproteomics Highlighted by the Analysis of Acute Leukemia Patients with Gut Colonization by Multidrug-Resistant Enterobacteriaceae

Julia Rechenberger [1],[†] , Patroklos Samaras [1],[†] , Anna Jarzab [1] , Juergen Behr [2] , Martin Frejno [1] , Ana Djukovic [3] , Jaime Sanz [4],[5] , Eva M. González-Barberá [4] , Miguel Salavert [4] , Jose Luis López-Hontangas [4] , Karina B. Xavier [6] , Laurent Debrauwer [7],[8] , Jean-Marc Rolain [9] , Miguel Sanz [4],[5] , Marc Garcia-Garcera [10] , Mathias Wilhelm [1],[*] , Carles Ubeda [3],[11],[*] and Bernhard Kuster [1],[2],[*]

[1] Chair of Proteomics and Bioanalytics, Technical University of Munich, 85354 Freising, Germany; julia.rechenberger@tum.de (J.R.); patroklos.samaras@tum.de (P.S.); anna.jarzab@tum.de (A.J.); martin.frejno@tum.de (M.F.)
[2] Bavarian Center for Biomolecular Mass Spectrometry, Technical University of Munich, 85354 Freising, Germany; juergen.behr@tum.de
[3] Centro Superior de Investigación en Salud Pública-FISABIO, 46020 Valencia, Spain; adjukovic@gmail.com
[4] Hospital Universitari i Politècnic La Fe, 46026 Valencia, Spain; sanz_jai@gva.es (J.S.); evamariagonzalezbarbera@gmail.com (E.M.G.-B.); salavert_mig@gva.es (M.S.); lopez_jlu@gva.es (J.L.L.-H.); sanz_mig@gva.es (M.S.)
[5] CIBERONC, Instituto Carlos III, 28029 Madrid, Spain
[6] Instituto Gulbenkian de Ciência, 2780 Oeiras, Portugal; kxavier@igc.gulbenkian.pt
[7] Toxalim, Université de Toulouse, INRA, INP-ENVT, INP-EI-Purpan, Université de Toulouse 3 Paul Sabatier, 31027 Toulouse, France; laurent.debrauwer@inra.fr
[8] Axiom Platform, UMR 1331 Toxalim, MetaToul-MetaboHUB, National Infrastructure of Metabolomics and Fluxomics, 31027 Toulouse, France
[9] Aix Marseille Univ, IRD, APHM, MEPHI, IHU-Méditerranée Infection, 13385 Marseille, France; jean-marc.rolain@univ-amu.fr
[10] Department of Fundamental Microbiology, University of Lausanne, 1015 Lausanne, Switzerland; marc.garcia.garcera@gmail.com
[11] Centers of Biomedical Research Network (CIBER) in Epidemiology and Public Health, 28029 Madrid, Spain
[*] Correspondence: mathias.wilhelm@tum.de (M.W.); ubeda_carmor@gva.es; (C.U.); kuster@tum.de (B.K.); Tel.: +49-8161-715659 (B.K.)
[†] These authors contributed equally to this work.

check for updates

Abstract: The microbiome has a strong impact on human health and disease and is, therefore, increasingly studied in a clinical context. Metaproteomics is also attracting considerable attention, and such data can be efficiently generated today owing to improvements in mass spectrometry-based proteomics. As we will discuss in this study, there are still major challenges notably in data analysis that need to be overcome. Here, we analyzed 212 fecal samples from 56 hospitalized acute leukemia patients with multidrug-resistant Enterobacticeae (MRE) gut colonization using metagenomics and metaproteomics. This is one of the largest clinical metaproteomic studies to date, and the first metaproteomic study addressing the gut microbiome in MRE colonized acute leukemia patients. Based on this substantial data set, we discuss major current limitations in clinical metaproteomic data analysis to provide guidance to researchers in the field. Notably, the results show that public metagenome databases are incomplete and that sample-specific metagenomes improve results. Furthermore, biological variation is tremendous which challenges clinical study designs and argues that longitudinal measurements of individual patients are a valuable future addition to the analysis of patient cohorts.

---

1. Introduction

Research over the past years has established the importance of the gut microbiota for human health and that a disturbed equilibrium is involved in the development of disease [1]. Therefore, scientists have begun characterizing the microbiome of the human gut in healthy and diseased states. Today, most microbiome studies rely on 16S rRNA or shotgun metagenome sequencing to provide a taxonomic description of the microbiome [2]. This does, however, not necessarily reflect proteomic and metabolic activity and, thus, may lack direct functional information. Other omic technologies, such as metaproteomics, metatranscriptomics or metabolomics, can supplement the genomic approaches by providing a molecular view on cellular processes at a more direct functional level [3]. The term metaproteomics was introduced by Wilmes and Bond [4] as well as Rodriguez-Valera [5] as "the large-scale characterization of the entire complement of environmental microbiota at a given point in time". The main conceptual advantage of metaproteomics is that it can add functional annotations to the description of the microbiome. In addition, metaproteomic can detect proteins from both the host and microbiota simultaneous and, thus, aid in the characterization of host-microbiome interactions [6].

So far, metaproteomic studies have mainly reported variations in the microbiome of healthy people [7,8], changes as a result of antibiotic treatment [9] or the role in chronic gut inflammation, such as Crohn's disease, inflammatory bowel disease or ulcerative colitis [10,11], as well as obesity [12,13] and diabetes [14]. Here, we present the first metaproteomic study of the gut microbiome in leukemia patients colonized with multidrug-resistant Enterobacteriaceae (MRE). Infections with multidrug-resistant pathogens during hospitalization are becoming critical. Leukemia patients especially are frequently affected since they have a compromised immune system and are regularly exposed to pathogens during extended periods in hospitals. In addition, leukemia patients are frequently treated with antibiotics altering the microbiome that confers resistance to intestinal colonization by exogenous bacteria [15]. Hence, leukemia patients frequently acquire secondary infections during hospitalization. Therefore, it is important to better understand how the gut microbiota may prevent colonization by pathogens and how such information may be utilized in the clinical management of patients.

Although sample preparation protocols in metaproteomics are becoming standardized for clinical studies [16] and the very high performance of liquid chromatography-mass spectrometry (LC-MS/MS) allows the efficient collection of metaproteomic data, the actual analysis of this data is still facing major challenges [17]. These include the lack of truly comprehensive bacterial sequence databases, the demand for considerable computational power, and a shortage of functional and taxonomic annotation [18]. Estimations for fecal samples suggest the potential presence of up to 1,000,000 possible unique proteins [19], leading to sequence databases that are enormous in terms of size. On top of requiring high computational power and large storage systems for data handling and processing, such excessive search spaces result in a significant loss of peptide identification sensitivity. While this issue can be partially addressed by using sample-specific databases generated by genomics or transcriptomics, the absence of annotations for these creates the need for large-scale sequence similarity (e.g., basic local alignment search tool (BLAST) [20]) searches to obtain taxonomic and functional information, which again requires great computational efforts. Furthermore, mapping peptides to proteins and taxa is not trivial due to the many (usually tryptic) peptides that are shared by homologous proteins [21]. In addition, metaproteomic analysis is further challenged by high levels of proteomic sample complexity, dynamic range of the species present and their protein expression levels and, importantly, by large inter- and intra-patient variability [22].

Despite these challenges, analysis of the metaproteome is important. Therefore, we embarked on the first gut metaproteome study of MRE gut colonized leukemia patients. We analyzed 212 fecal samples from 56 patients and provide, to our knowledge, one of the largest clinical metaproteomic datasets to date. In the present manuscript, we report on the analysis of this data, highlight the main challenges and draw some conclusions that may guide scientists and clinicians when designing and conducting metaproteomic projects.

2. Materials and Methods

2.1. Sampling Process

Fecal samples were collected from November 2013 until April 2015 from acute leukemia patients hospitalized at the Hospital La Fe (Valencia, Spain). All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved on the 1st of July 2013 by the Ethics Committee of CEIC Dirección General de Salud Pública y Centro Superior de Investigación en Salud Pública (20130515/08). A total of 802 fecal samples were collected from 133 patients. Samples were collected in weekly intervals during the hospitalization period and screened for the presence of MRE. A subset of 56 patients was included in the present study, where MRE colonization was detected in at least one sample. After the first MRE detection, one or more consecutive samples from that patient were included until MRE was not detectable or the patient was discharged. In the case of an MRE recolonization during the same or another hospitalization period samples from these new colonization process were also included following the same criteria. In total, 221 samples from 58 patients matched the inclusion criteria. Due to the limited sample amount, the metaproteome was analyzed for 212 samples from 56 patients. Fresh fecal samples were kept at 4 °C for less than 24 h. Subsequently, three aliquots of each sample were weighed and resuspended in 1 mL of autoclave-sterilized PBS 15% glycerol to preserve viability of bacteria upon freezing and kept at −80 °C until further processing. Three additional aliquots for proteomic and genomic analysis were weighed and directly frozen at −80 °C until further processing. Patient-related metadata was prospectively collected and recorded in a computerized database in Access®. This data included information about antibiotic treatments, pathogen presence, gender, age, and type of admission. MRE colonization was determined by culturing fecal samples with Brilliance ESBL agar plates (Oxoid), containing third-generation cephalosporin. Plates were incubated for 24 h at 37 °C. Taxonomic identification of the grown colonies was determined through matrix-assisted laser desorption/ionization - time of flight mass spectrometry (MALDI-TOF MS). If no growth was observed, the plate was left for an additional 24 h at 37 °C to confirm the negative result. In addition, the antibiotic resistant pattern was determined through the Vitek 2 system. Antibiotics tested included amikacin, amoxicillin-clavulanic acid, ampicillin, cefepime, cefotaxime, cefoxitin, ceftazidime, cefuroxime, ciprofloxacin, gentamicin, imipenem, ertapenem, piperacillin-tazobactam, tigecycline, and trimethoprim-sulfamethoxazole (co-trimoxazole). The susceptibility was determined according to the Clinical and Laboratory Standards Institute Guidelines (2016). In addition, resistance to meropenem was evaluated using ETEST antibiotic gradient strips (bioMérieux) in strains isolated from patients who had received meropenem. An isolate was considered multidrug resistant if it was non-susceptible to at least 1 agent in 3 or more antimicrobial categories defined by Magiorakos and co-workers [23].

2.2. Sample Preparation

For the proteomic analysis, feces were thawed on ice. For homogenization, 0.5 mL washing buffer (50 mM Na_2HPO_4/NAH_2PO_4, pH 8.0, 0.1% Tween20, 1× cOmplete Protease inhibitors (Sigma Aldrich, MO, USA)) and 2 glass beads (3 mm) were added to 50 mg feces, vortexed and sonicated in a water bath for 10 min. After homogenization, the sample was centrifuged for 15 min at 4 °C at 200× g and pellet and supernatant were kept. For further purification, this step was repeated three times.

The supernatant of the first washing step and the pellet was used for proteomic analysis. To process the bacterial pellet fraction, glass beads were removed, and the pellet was resuspended in 20 mM Tris/HCl and then further diluted in lysis buffer (20 mM Tris/HCl, pH 7.5, 2% SDS, 1× cOmplete Protease inhibitors (Sigma Aldrich)). Samples were heated to 60 °C for 10 min and ultrasonicated for 3× 1 min (0.5 amplitude, Sonopuls Mini20, Bandelin) on ice. After lysis, samples were centrifuged for 1 h at 4 °C at 20,000× *g* and the supernatant was reduced with 10 mM dithiothreitol (DTT) at 50 °C, 700 rpm for 40 min and alkylated with 55 mM chloroacetic acid (CAA) at room temperature for 20 min in the dark. Samples were mixed with 1× lithium dodecyl sulfate (LDS) buffer and run into an sodium dodecyl sulfate (SDS) gel (5 min, 200 V). The gel was stained with Coomassie and one single stained sample band was cut for in-gel digestion following standard procedures [24]. For the supernatant fraction of the fecal washing, 50 µL of each sample supernatant was reduced, alkylated, denatured, and in-gel digested as described above. Input material for LC-MS/MS analysis was normalized based on feces weight.

2.3. LC-MS/MS Analysis

LC-MS/MS measurements were performed using a Dionex Ultimate 3000 UHPLC+ system coupled to a Q Exactive HF mass spectrometer (Thermo Fisher Scientific, MA, USA). After reconstitution in 0.1% formic acid (FA), one half of the peptides were loaded on a trap column (75 µm × 2 cm, packed in-house with 5 µm C18 resin; Reprosil PUR AQ, Dr. Maisch) and washed using 0.1% FA at a flow rate of 5 µL/min for 10 min. Subsequently peptides were transferred to an analytical column (75 µm × 45 cm, packed in-house with 3 µm C18 resin; Reprosil PUR AQ, Dr. Maisch) with a flow rate of 300 nL/min and separated using a 60 min gradient from 4 to 32% LC solvent B (0.1% FA, 5% DMSO in acetonitrile) in LC solvent A (0.1% FA, 5% DMSO) [25]. The instrument was operated in data-dependent acquisition (DDA) and positive ionization mode. MS1 full scans were acquired from 360 to 1300 *m/z* at a resolution of 60 K, an automatic gain control (AGC) target value of 3e6 charges and a maximum injection time (maxIT) of 10 ms. Precursor ions for higher energy collisional dissociation (HCD) fragmentation were selected with a Top 20 method, and MS2 spectra were recorded from 200 to 2000 *m/z* at 30 K resolution using an isolation window of 1.7 *m/z*, an AGC target value of 2e5 charges, a maxIT of 50 ms (for pellets) and 100 ms (for supernatants), 25% normalized collisional energy (NCE), a dynamic exclusion of 20 ms (for pellets) and 25 ms (for supernatant) and with a fixed first mass of 100 *m/z*. Samples were randomized for measurement, and *E. coli* standards were measured as quality control every 30 samples. As a quality control for lysis and digestion, *E. coli* samples were processed along with every 9 samples (Supplementary Figure S1).

2.4. Data Processing and Analysis

For proteomic data analysis, raw files were searched using Maxquant/Andromeda (v. 1.5.7.4) [26] against four different sequence databases: the Integrated Genome Reference Catalog (IGC): 9,878,647 entries [27], SWISS-PROT bacteria: 333,480 entries, downloaded 24 April 2018, SWISS-PROT human canonical and isoform: 42,123 entries, downloaded 2 January 2016, Sample specific databases based on metagenomic sequencing, see below). For the human database, all 424 raw files were searched together, whereas for the bacterial databases each sample (pellet and supernatant combined) was searched separately. MaxQuant results were post-processed for 1% peptide spectrum match (PSM) and peptide false discovery rate (FDR) using Percolator [28]. Taxonomic and functional annotation was obtained using the Unipept Metaproteome Analyzer tool (v. 4.0) [29,30]. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral. proteomexchange.org) via the proteomics identifications (PRIDE) partner repository [31] with the dataset identifier PXD011515.

2.5. Metagenomic Sequencing and Data Processing

Part of the feces sample was used for the extraction of DNA to determine the microbiota composition. DNA was extracted using a QIAamp® Fast DNA Stool Mini kit (QIAGEN) with a previous step of mechanical disruption to improve cell lysis. Briefly, cells were resuspended in 1.4 mL of Inhibitex buffer and 500 μL of 0.1 mm glass beads and tubes were vortexed at maximum speed for 5 min prior to the lysis at 95 °C for 7 min. Subsequent steps of the DNA extraction followed the QIAamp kit protocol. DNA concentration was determined with a QubitTM fluorometer using the manufacturer's protocol. DNA was sequenced using the NextSeq platform from Illumina (high-output 300 cycle kit), following the manufacturer's protocol. An average of 11 M reads per sample and an average 47× coverage was obtained. For processing, adapter sequences were removed from raw data using Cutadapt (v. 1.12) [32]. Quality filtering was performed using Trimmomatic (v. 0.38) [33]. Only reads with a size of 101 bp or higher were further processed to avoid possible misclassification of short reads. Cleaned genomic data (on average 3.1 Gb per sample) were assembled with SPAdes (v. 3.7.1) [34] using the 'meta' algorithm to improve the metagenomics reconstruction. To maximize the contig size and reduce the probability of chimeric contig reconstruction, the following k-mer lengths were selected: 21, 33, 55, 77, 121, 251. For each sample, contigs with greater than 2500 bp were selected. From those, open-reading frames (ORFs) were identified and annotated using Prodigal (v. 2.6.2) [35]. Contigs from all samples were combined and clustered at 99% identity and 90% coverage to remove redundancy in the dataset, using VSEARCH (v. 2.9.0) [36]. Trimmed reads were then mapped against the database of non-redundant contigs using Bowtie2 (v. 2.3.4) [37]. Contig frequencies were used to reconstruct metagenomic-assembled genomes (MAGs) using MetaBAT (v. 2.10) [38]. Only MAGs with a completeness of, at least, 70% and a contamination of less than 10% were kept. MAGs were classified phylogenetically using CheckM (v. 1.10.13) [39]. Phylogenetically coherent MAGs were merged together using CheckM if their completeness increased more than 10% and the merged contamination was lower than 10%. For protein database generation ORFs were translated into amino acids.

2.6. 16S rRNA Sequencing and Data Processing

Extracted DNA was further used to determine the taxonomic distribution of each fecal sample. Therefore, the V3-V4 region of the 16S rRNA gene was amplified and sequenced using the MiSeq platform from Illumina, as described in the manual for "16S Metagenomic Sequencing Library Preparation" of the MiSeq platform (Illumina). Briefly, for each sample, a 25 μL reaction was prepared containing 12.5 ng of DNA, 12.5 μL 2× KAPA HiFi Hot Start Mix, and 0.2 mM of primers. Water was added to complete the volume of the reaction. In case there was not enough amount of DNA, the maximum volume (11.5 μL of DNA) was added to the reaction, but the number of amplification cycles was increased from 25 to 30. Cycling conditions were 95 °C for 3 min, and 25 cycles of 95 °C for 30 s, 55 °C for 30 s, and 72 °C for 30 s, and a final elongation cycle at 72 °C for 5 min. The amplification was confirmed through electrophoresis by loading 4 μL of the PCR reaction on a 1.6% agarose gel. Subsequently, the PCR product was purified with the AMPure XP beads as described in the Illumina protocol. Next, a limited-cycle PCR reaction was performed to amplify the DNA and add index sequences on both ends of the DNA, thus, enabling dual-indexed sequencing of pooled libraries. Index PCR consisted of a 50 μL reaction containing 5 μL of the DNA obtained from the previous PCR, 25 μL of 2× KAPA HiFi Hot Start Mix, and 5 μL of forward and reverse indexed primers. Temperature conditions were the same as for the first reaction, but the number of cycles was reduced to 8. The obtained PCR product was purified with the AMPure XP beads following the manufacturer's protocol. An equal amount of the purified DNA was taken from each sample for pooling. Each pool of samples (N = 96) was sequenced following Illumina recommendations. Sequences were processed using Mothur (v. 1.35) [40]. Initial trimming by quality was performed on paired ends of sequences before joining them into a single read. For this initial trimming the Prinseq Lute package was used. Parameters used for trimming included elimination of sequences shorter than 250 bp or that contained homopolymers longer than 8 bp or undetermined bases. Using the base quality scores, which range from 0 to 40

(0 being ambiguous base), sequences were trimmed using a sliding-window technique from the 3' end, such that the minimum mean quality score over a window of 50 bases never dropped below 25. Sequences were aligned to the 16S rRNA gene using the SILVA reference alignment as a template. Potential chimeric sequences were removed using the Uchime algorithm. To minimize the effect of pyrosequencing errors in overestimating microbial diversity, rare abundance sequences that differ in up to four nucleotides from a high abundant sequence were merged to the high abundant sequence using the pre.cluster option in Mothur. Since different numbers of sequences per sample could lead to a different diversity (i.e., more Operational Taxonomic Units-OTUs could be obtained in those samples with higher coverage), we rarefied all samples to the number of sequences obtained in the sample with the lowest number of sequences (10,095). In other words, 10,095 sequences were randomly selected from each sample for subsequent analysis: taxonomic characterization and OTUs identification. Sequences with distance-based similarity of 97% or higher were grouped into the same OTU using the VSEARCH [36] abundance based greedy clustering method. OTU-based microbial diversity was estimated by calculating the Shannon diversity index [41]. Each sequence was classified using the Bayesian classifier algorithm with a 60% bootstrap cutoff [42]. In most cases, classification could be assigned to the genus level. When it was not possible to classify a sequence to a certain taxonomic level, it was assigned as "Unclassified" followed by the upper taxonomic level.

2.7. Determination of 16S rRNA Counts with qPCR

To determine the total bacterial load of each fecal sample, qPCR of the 16S rRNA gene was performed. For this purpose, the KAPA SYBR FAST qPCR Kit was used. Briefly, for each sample, 20 μL PCR duplicates were prepared with each containing 2 μL of the DNA (see previous section) used as template, 10 μL of mix provided by the manufacturer, and 0.4 μL of forward and reverse primers at the final concentration of 0.2 mM. To complete the volume of the reaction, 7.2 μL of water was added. A PCR product of the 16S rRNA gene from *Enterococcus faecium* C68 strain was used for obtaining a standard curve. ENDMEMO program was used to determine the number of 16S rDNA molecules in the PCR product of *E. faecium* C68 based on sequence of 16S rRNA gene and concentration of the PCR product. A standard curve was obtained by making 5-fold dilutions of the PCR product. Cycling conditions of the qPCR were 94 °C for 5 min, and 45 cycles of 94 °C for 30 s, 56 °C for 30 s, and 68 °C for 30 s, and a final elongation cycle at 68 °C for 5 min. By extrapolation of the obtained results with the standard curve, the number of 16S rRNA gene copies was determined for each sample.

3. Results and Discussion

3.1. The Challenge of Experimental Design in Clinical Metaproteomics

In total, 212 fecal samples from 56 acute leukemia patients with MRE gut colonization were processed for metaproteomic analysis. Samples had been collected during hospitalization in approximately one-week intervals (Figure 1A). The cohort consisted of patients suffering from acute leukemia who were treated with chemotherapy or undergoing transplantation in addition to receiving antibiotics to treat their infections (Supplementary Table S1). As evident from Supplementary Table S1, the level of experimental control is much lower and the issue of confounding factors much higher in clinical studies than those conducted in animal or other model systems. Collecting multiple samples per patient and a large overall sample size are necessary to estimate inter- and intra-patient variability and to ensure statistical power.

The analytical workflow, however, from sample preparation to metaproteomic data is much easier to control. Here, samples were analyzed using a standardized proteomic workflow. Samples were divided into a supernatant and a pellet fraction to separate bacterial cells from secreted proteins of human and microbial origin. We opted for a sample preparation step that included gel electrophoresis and using in-gel trypsin digestion to generate peptides for LC-MS/MS analysis. In our experience, a short SDS gel electrophoresis step is a convenient and reproducible 'sample equalizer' because it is

compatible with harsh upstream protein extraction procedures (see methods) and generates peptides that are essentially devoid of non-peptidic contaminations, detergents, and insoluble particles, etc. (Figure 1B, Supplementary Figure S1). The employed 'single-shot' LC-MS/MS measurement strategy is also highly reproducible ensuring that the vast majority of the observed variation in the project data arises from biological rather than technical sources. The raw MS data were processed using four sequence databases of different sizes and content covering human and bacterial proteins. In this way, we were able to compare peptide identification results of every single database to the combination of all (Figure 1C). For the bacterial databases, SWISS-PROT annotated entries for bacteria and the large integrated reference gut microbiome catalog (IGC), both publicly available resources, were used. In addition, we performed shotgun metagenome sequencing for every sample and generated sample specific protein sequence databases for each.
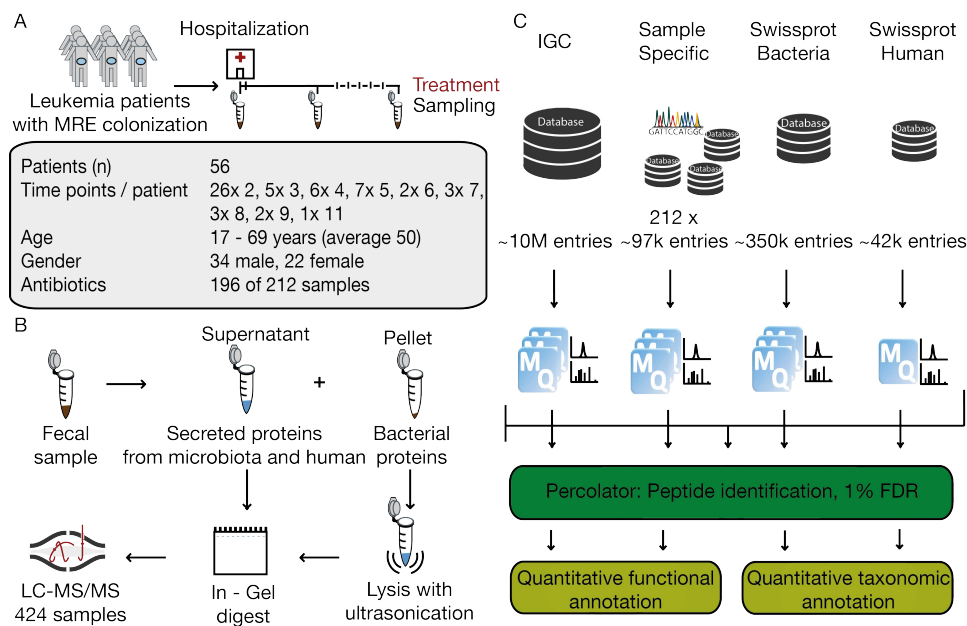

Figure 1. Study design, proteomic workflow and data processing pipeline. (**A**) Acute Leukemia patients were sampled in weekly interval during the time of hospitalization. In total 212 fecal samples of 56 patients with MRE gut colonization were analyzed, providing additional information about age, gender, and treatment conditions. (**B**) For the protein extraction, fecal samples were divided in supernatant and pellet fractions. Bacterial cells in the pellet fraction were lysed with ultrasonication and for both samples' proteins were digested in gel. Thereafter, samples were measured with LC-MS/MS. (**C**) Raw files were searched with four different databases in separate and combined MaxQuant searches and post-processed with Percolator and with quantitative functional and taxonomic annotation analyzed.

3.2. The Challenge of (the Lack of) a Comprehensive Sequence Search Space

To be able to convert LC-MS/MS data into identified peptide sequences, comprehensive, ideally complete, protein sequence collections must be available for the (many) species that are present in a metaproteomic sample. The selection of the sequence database is, therefore, a particularly important step for metaproteomic analyses [21,43]. To demonstrate this challenge, the following section investigates several theoretical aspects before we turn to the experimental data. To illustrate the size of the theoretical search space for the three bacterial and the human sequence databases used in this study, we digested each in silico to compare the number of resulting theoretical tryptic peptides. With almost 10 million protein entries, IGC is the largest database in our comparison. It was assembled as part of an international initiative that performed shotgun metagenome sequencing of 1070 individuals from around the world and is considered to be a comprehensive and high-quality gene catalog for the human gut microbiome. Somewhat surprisingly, it turns out that IGC only covers

half of the theoretical peptides derived from the sample-specific sequencing databases (Figure 2A) generated in our project. This means that even IGC is not nearly as comprehensive as one might expect. Again, in proteomics, peptide and, therefore, protein identification relies on the matching of spectra to peptide sequences from the protein database. If a peptide sequence in the database does not fit the actual sequence of the acquired spectrum, it cannot be identified. Even single amino acid changes can lead to missing and false positive identifications. Therefore, it is important that the protein database used for peptide identification matches the bacterial composition of a given sample as closely as possible [44]. Bacterial populations are evolving very fast to adapt to changing environments [45]. Therefore, each individual will likely contain unique strain compositions in the gut although particular bacterial species can be shared [46]. It is, therefore, not clear if or to what extent public metagenome collections can ever be comprehensive.

In comparison, the catalog of bacterial proteins in SWISS-PROT is much smaller, but its entries are manually validated and annotated which is why SWISS-PROT is perhaps the highest quality sequence database available. However, the vast majority of the bacterial entries in SWISS-PROT are not specific to the gut environment and, thus, one might not expect a major overlap to the genomic databases. Indeed, SWISS-PROT shares only 0.5% of the total peptides with either IGC or the sample-specific protein databases and is, therefore, not a useful source of sequences for gut metaproteomics.

By direct comparison, the bacterial databases combined are 200 times larger than the theoretical search space of human sequences in SWISS-PROT. Fortunately, only 10% of all theoretical human peptides are shared with the bacterial database entries allowing the identification of bacterial and human peptides in a sample in parallel. Comparing the 212 individual sample specific databases to each other (Figure 2B) shows that the individual databases are very diverse. Only around 100,000 peptides are shared between at least 50% of the samples. This corresponds to just 12% of the median size for each sample specific database. This illustrates that even with the sensitivity and comprehensiveness of genome sequencing, a large proportion of the proteins in a sample are unique or contain amino acid variations, demonstrating very considerable diversity of the gut microbiota between and across individuals. This, in turn, severely limits the extent to which quantitative comparisons can be made on the level of individual proteins between patients, their colonization status, medication or other metadata.
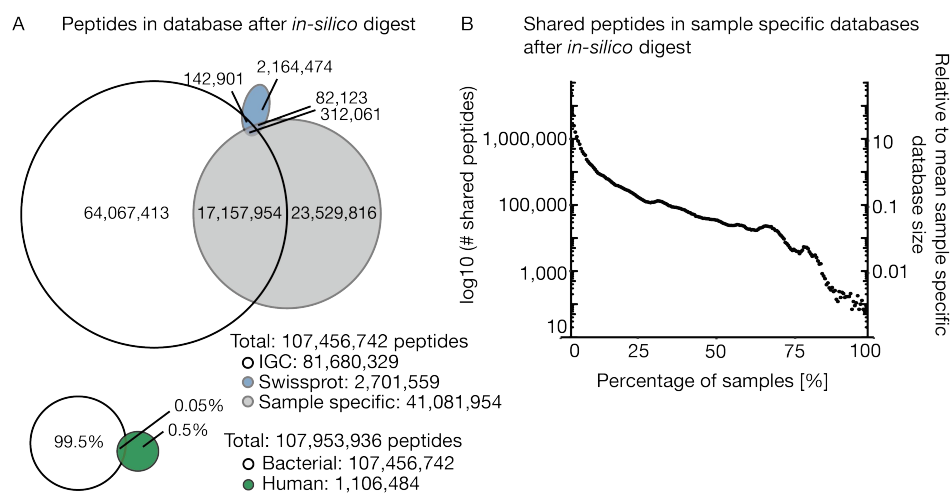

Figure 2. *In silico* comparison of four different databases. Four different databases (Integrated Genome Reference Catalog (IGC), SWISS-PROT bacteria, SWISS-PROT human and sample specific metagenome-based databases) were digested *in silico,* and the possible search space was compared. (**A**) Venn diagram of the resulting peptides after *in silico* digestion comparing the three bacterial databases and all bacterial databases combined versus the peptides from the *in silico* digested human database. (**B**) Number of shared peptides in the 212 sample specific databases against the percentage of samples. The right axis indicates to which the percentage of the average sample specific database the number of shared peptides corresponds.

3.3. The Challenge of Extensive Demand for Computational Power and Storage Capacity

In moving from theoretical consideration to the analysis of the experimental data, we performed MaxQuant searches for all samples and all four databases separately. As the false discovery rate (FDR) estimation procedure of MaxQuant shows strong limitations when very large sequence databases are used (Supplementary Figure S2), MaxQuant outputs were post-processed using Percolator to recover identifications. For the combination of database searches ('Combined'), all MaxQuant results of the individual searches against IGC, SWISS-PROT bacteria and the particular Sample specific database were combined but only retaining the highest scoring peptide sequence per spectrum. The combined MaxQuant results were then post-processed in Percolator. Searching the IGC sequence database (~10 million entries) required considerable computational power and time. For one single sample, the search against IGC produced an output of 85 GB on average. In contrast, searching the much smaller sample-specific databases generated an output of 4.5 GB on average per sample. One theoretical advantage of searching IGC using MaxQuant is that all samples of a project can be combined into a single run, which improves the grouping of protein sequences into a minimal set. However, such a search was not practical given the 2×212 LC-MS/MS files produced in this study. Searching all samples against IGC (the pair of pellet and supernatant for each sample were searched together), produced an output of 17.8 TB of files and required ~1834 h of run time (10 cores, Intel® Xeon® Gold 6150 CPU @ 2.70GHz (4 processors)). In contrast, searching all 212 sample specific databases produced an output of 'just' 830 GB in ~310 h of run time (2 cores, Intel® Xeon® Gold 6150 CPU @ 2.70GHz (4 processors)). This illustrates that searching IGC is not practical in most instances, particularly given the often limited computational infrastructure available in hospitals. And even searching the sample-specific databases will pose challenges when analyzing large sample cohorts.

The total number of identified peptides across all databases showed considerable variation ranging from just a few hundred to up to 10,000. To check for biases in identifications, we compared peptide identification between supernatant and pellet fraction of each sample, Shannon-diversity [41], 16S rRNA count and Bradford protein concentration with the number of identified peptides of each sample (Supplementary Figure S3, Supplementary Table S2). As expected, good correlation for peptide identification between pellet and supernatant was observed, whereas the other parameters showed low correlation. The Venn diagram of shared identified peptide sequences for all bacterial databases shows that most peptides are covered by the sample-specific databases (Figure 3A, top right inset). High positive correlations can be observed when comparing the number of identified peptides between the bacterial databases. Only for SWISS-PROT, the correlations are slightly lower, which probably arises from the overall lower number of identifications. Negative correlations were obtained for comparison to the human database (Figure 3A), i.e., the more bacterial peptides were identified, the fewer human peptides could be found. This is readily explained by the fact that the mass spectrometer cannot exhaustively measure all peptides in a sample and the detection of (largely constant) human background proteome is increasingly suppressed the more bacteria are in a sample. As one would expect, more peptides were identified in pellet samples than in the corresponding supernatants (Figure 3B). In addition to expectation, peptides of human origin were more frequently identified in the supernatant. The numerical range of identified peptides from the human database search was smaller compared to the bacterial peptides, again indicating that there is a largely constant background of human proteins in the samples.

Comparing the total number of peptide identifications showed that sample specific databases yielded the highest number of identification, followed by the IGC (Figure 3C). This underscores that sample specific databases are more representative of metaproteomic content than IGC as already expected from the theoretical considerations discussed above. In addition, we observed that combining all bacterial databases ('Combined', including IGC, Sample specifics, and SWISS-PROT Bacteria) for post-processing into one leads to a substantial decrease in peptide identifications. As noted above, this is a direct consequence of the limitations of current database search engines in controlling peptide identification FDR when using large sequence collections [47]. Unfortunately, no fully satisfactory

practical solution has been found to this issue yet. When combining all unique peptide sequence identifications of the three bacterial databases ('all DBs additive') identifications would be increased by ~20%. However, this is not a very practical approach because of the demands on computational power also discussed above and limited FDR estimation accuracy.

Any comparative analysis requires that the same peptides/proteins are found between samples. Figure 3B highlights the large range of peptide identifications per sample. When we calculated the number of identified peptides that are shared across samples, an extremely small overlap was observed (Figure 3D). The extent of this non-overlap is stunning as there were only very few peptides that were identified in all samples and most of these were consistently identified from the human host. This has profound implications that set metaproteomics apart from any single species proteomic investigation in that some type of heuristics must be applied to aggregate results into a level which allows comparison. This may be as high as the species level which would preclude any functional interpretation of the molecular level beyond what, for example, gene ontology (GO) analysis may have to offer.

From the above analysis, we conclude that sample-specific sequence databases are the preferred option for metaproteomic projects today. Therefore, all the analysis presented below is based on results from sample specific database searches.
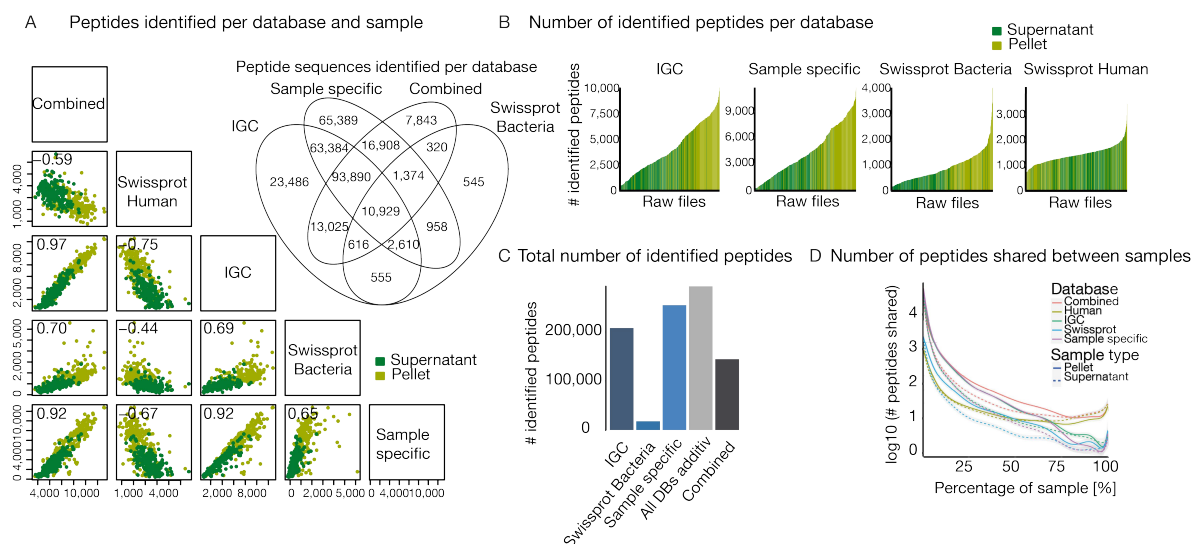

Figure 3. Comparing the influence of database selection on peptide identification. (**A**) Multi-scatter plot of identified peptides at 1% PSM and peptide FDR for the four different databases and all databases combined. Identification for pellet and supernatant fraction of each sample is shown separately. Pearson correlation is shown in top left of each box (highest p value is 8.2×10^{-21}). The Venn diagram shows the overlap of identified peptides over all samples for the three bacterial and the combination of all four databases. (**B**) Histogram of the number of identified peptides of supernatant or pellet for each sample. Raw files are sorted according to the number of identified peptides. (**C**) Bar plot of the number of total identified peptides over all samples per database. 'All DBs additive' shows the theoretical identification by summing up all unique peptides of the three bacterial database types. (**D**) Polynomial curve fit for the number of shared peptides across all samples for the different databases. Separated for supernatant and pellet fraction of the samples.

3.4. The Challenge of Functional and Taxonomic Annotation

Although metagenome-derived sequence databases are currently the best available representation of the possible species and protein content of a sample, subsequent annotation by mapping peptide or proteins to taxonomic and functional information is required. For this purpose, we used the Unipept metaproteome analyzer tool for the annotation of tryptic peptides. In addition to metagenome and metaproteome data, we also generated 16S rRNA gene sequencing data for all samples (Supplementary Table S3). To compare class-level 16S rRNA information to the proteomic

data, peptides were mapped to their lowest common ancestor. For proteomics, it was possible to map 60% of the identified peptides to a taxonomy level. The remaining peptides could not be found in the Unipept database (NCBI) or were not unique for taxonomic identification. In that way, we identified 72 classes, 383 genera, and 595 species in the patient samples (Supplementary Table S4). In comparison, 16 S rRNA data identified 32 classes and 334 genera across all samples. Gratifyingly, the overall taxonomic distribution per sample derived from the proteomic and 16S rRNA data showed a mean Pearson correlation of 0.95 at the class level (0.94 for order, 0.78 for family, and 0.87 for genus), meaning that it was possible to derive taxonomic distributions from metaproteomic analysis, and there was no evidence that poor correlations generally resulted from poor proteome coverage (Figure 4A, Supplementary Figure S4). In line with prior results [48,49], the most abundant classes over all samples were Bacteroidia, Clostridia, and Verrucomicrobiae (Figure 4B).

For further functional annotation, 80% of the peptides mapped to at least one GO term, and 42% could be associated to an E.C. number. This indicates the detection of proteins related to 1,205 molecular functions, 738 biological processes, and 145 cellular components for bacterial peptides and 298 molecular functions, 337 biological processes, and 139 cellular components for human peptides. The ten most abundant GO in terms of the microbiota over all samples mainly represented cell growth and metabolic homeostasis (Figure 4C). In contrast, abundant human peptides were mainly associated with immune responses and metabolic/catabolic activity and belong to proteins that are located in extracellular regions (Figure 4D), which largely recapitulates observations reported by earlier studies of host-microbiome interactions [6,50]. One important learning from this section is that metaproteomics can represent the taxonomic diversity in a sample as well as 16S rRNA gene sequencing but simultaneously providing more details on functional molecular information.
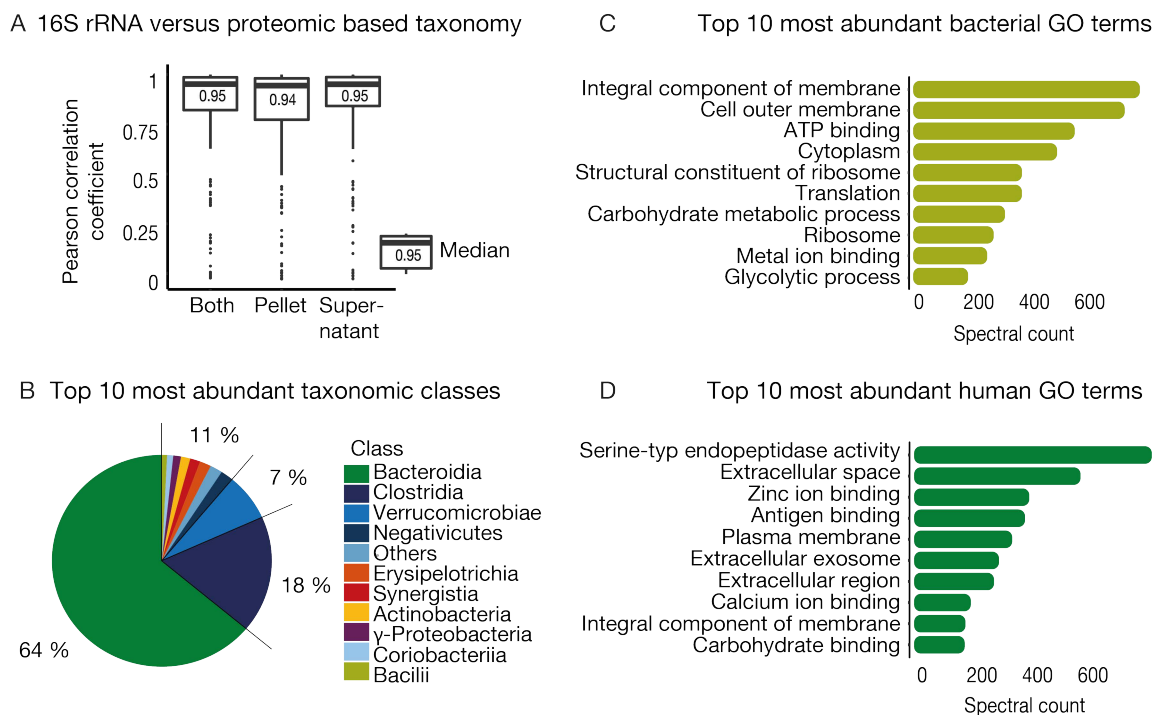

Figure 4. Description of the taxonomic and functional composition. (**A**) Box plot of Pearson correlation of taxonomic composition detected at the class level with 16S rRNA sequencing and proteomic analysis for each sample. Both: supernatant and pellet for each sample combined, Supernatant: only the supernatant fraction of each sample, Pellet: only the pellet fraction of each sample. (**B**) Pie chart shows the most abundant identified taxonomic classes over all samples. (**C**) Bar plot of average spectral counts for the 10 most abundant bacterial gene ontology (GO) term over all samples. (**D**) Bar plot of average spectral counts for the 10 most abundant human GO term over all samples.

3.5. The Challenge of High Microbiome Variability in Clinical Samples

The data shown in Figure 3 already indicated that there is a very high degree of variability in the metaproteomes of clinical samples, and the quality control standards within the proteomic workflow show that this is not the result of technical variations (Supplementary Figure S1). Therefore, simple comparisons between samples based on peptide and protein intensities, as are common for single species studies, were not possible. Instead, we resorted to using Jaccard similarity (Figure 5) to compare protein expression profiles of samples. Here, similarity is described by the presence and absence of peptides between pairs of samples. Clustering of dendrograms was based on Pearson correlation of Jaccard distances (see methods for details). Clustering of all samples (Supplementary Figure S5) showed a clear separation of supernatant and pellet samples. In addition, it was apparent that supernatant samples showed higher overall similarities than pellet samples because human proteins were more consistently identified in supernatants. We next selected six patients for whom the most multiple samples were available. The heat map shown in Figure 5A shows that the microbiomes within patients were more similar than between patients (Figure 5A). This indicates that peptide variation among samples is mainly driven by individual microbiome differences rather than by technical variation of the proteomic analysis. For peptides derived from human proteins, the overall similarity was expectedly much higher than for the microbiomes (Figure 5B). Patients did not, however, cluster based on the human peptide expression, indicating that the human expression profiles were much less variable.
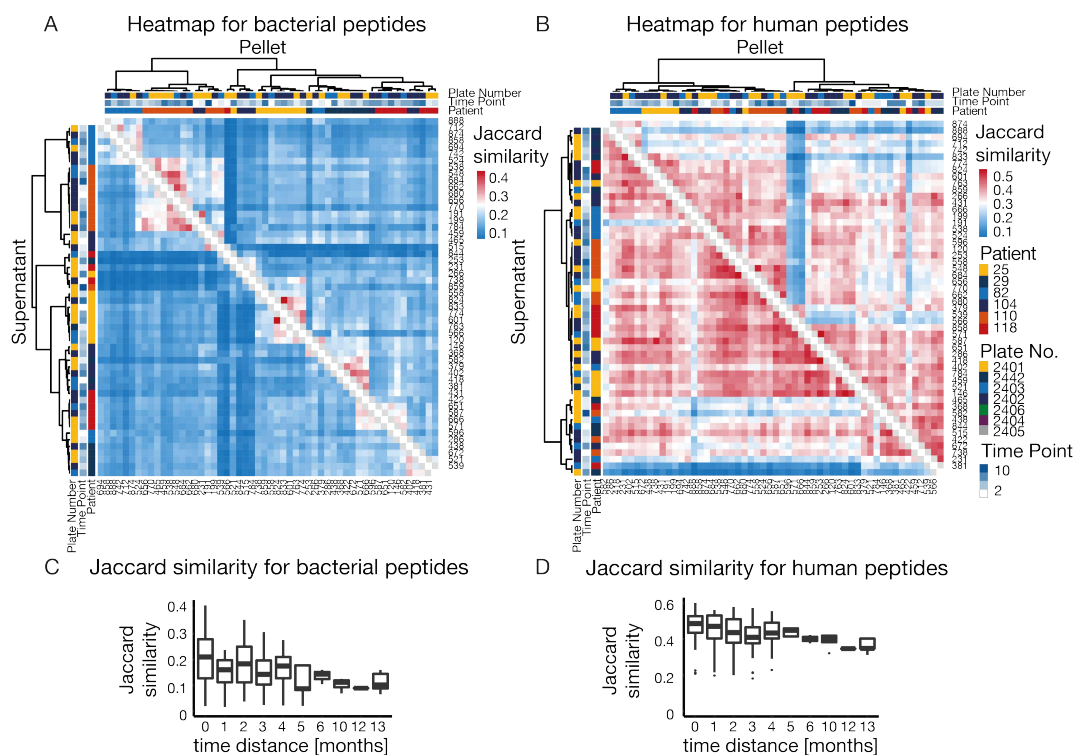

Figure 5. Sample variability (**A**) Heatmap of Jaccard similarities based on the presence/absence of bacterial peptides for the top six patients with the most sampling time points. Dendrogram clustering is based on Pearson correlation of Jaccard distances. Bottom triangle for the supernatant fraction of the sample. Top triangle for pellet fraction of the sample. (**B**) Heatmap of Jaccard similarities based on the presence/absence of human peptides for the top six patients with the most sampling time points. Dendrogram clustering is based on Pearson correlation of Jaccard distances. Bottom triangle for the supernatant fraction of the sample. Top triangle for pellet fraction of the sample. (**C**) Boxplot of Jaccard similarities for bacterial peptides of paired samples with different time distances between sampling points. (**D**) Boxplot of Jaccard similarities for human peptides of paired samples with different time distances between sampling points.

The longitudinal data collected for several patients also allowed us to assess how their metaproteome changed over time. Figure 5C,D depict the grouped Jaccard pair-wise similarities for bacterial and human peptides based on the time between sampling. It is apparent that with increasing time, microbiome similarity decreased from 0.35 to below 0.2, whereas the human profile showed a more stable similarity of 0.5 to 0.4. This observation has important consequences for the future design of clinical metaproteomics projects as it seems important to collect a significant number of longitudinal samples, which we expect to improve the interpretability of individual metaproteomes. Nevertheless, it should be mentioned that patients in this study received a variety of antibiotics altering microbiome composition and, therefore, higher variability compared to a healthy population is expected [51]. To overcome the problem of missing values, analyses are often performed using functional or taxonomic annotation levels, where several peptides/proteins are combined and, therefore, variability is decreased (Supplementary Figure S6) [52]. In this way, the taxonomic and functional composition of patient samples can be compared quantitatively (Figure 6). Taxonomic and GO term distribution of samples for every patient can be found in the Supplementary Material 1.
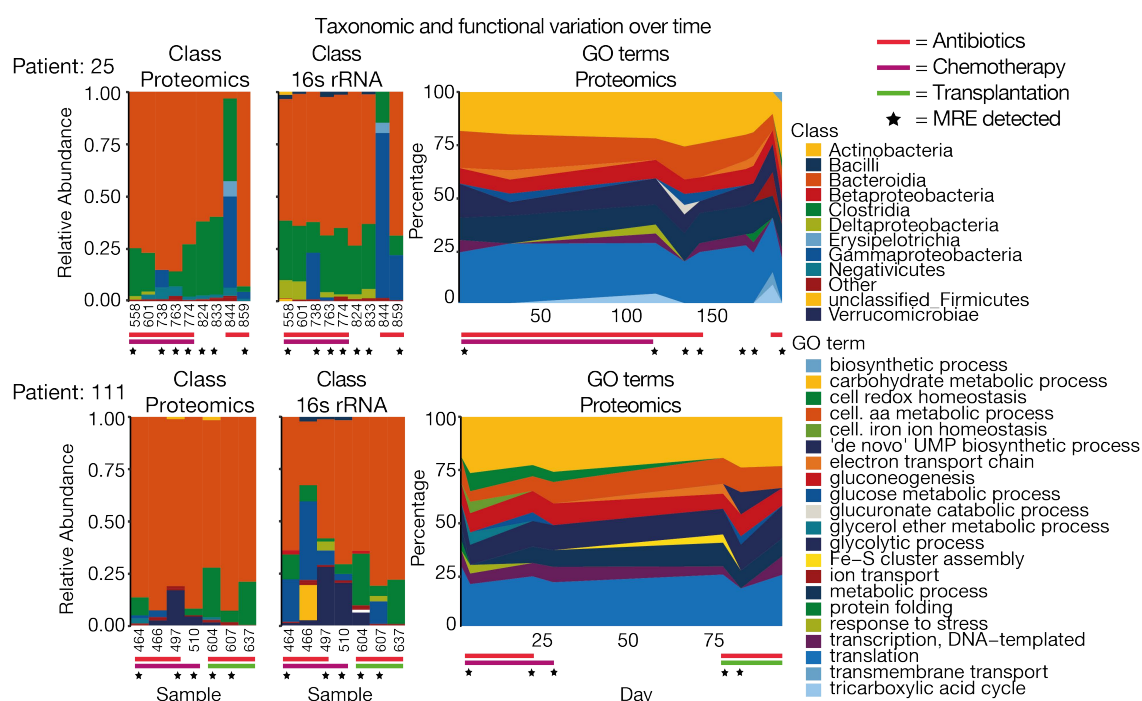
Figure 6. Comparing taxonomic and functional data for longitudinal samples. Taxonomic class abundances retrieved from proteomic and 16S rRNA data as well as GO term abundances were compared for samples for two patients over time. In addition, antibiotic treatment at sampling time point and type of hospital admission (i.e., chemotherapy or transplantation) for the sampling time is indicated.

4. Conclusions

Metaproteomics is a young and developing field of research. PubMed currently (20th October 2018) lists a total of ~500 publications, whereas ~7500 scientific reports are published per year in proteomics as a whole (Supplementary Figure S7). Although metaproteomic analysis has seen substantial progress, major challenges remain to be overcome. When designing future clinical metaproteomic studies, our data suggest that it is advisable to include longitudinal sampling systematically for each patient and to keep sampling intervals short and consistent within the cohort. Clinical studies often suffer from small sample sizes and, therefore, poor statistical power [53]. We emphasize this point for future clinical study designs as generating the actual metaproteomic data is no longer a bottleneck. As in all clinical studies, it is important to record as much (and correct)

meta information as possible about the patients and their treatments to be able to account for potential confounding factors and to distinguish interesting effects from uncontrolled factors.

For the time being, the generation of sample-specific databases is highly recommended to support comprehensive peptide and protein identification. While this requires metagenome sequencing of each sample, it mitigates the inevitable loss in confident peptide identifications when using community-based resources, such as IGC. In addition, we propose that transcriptomics data from RNA-Seq could be another way to generate databases and would likely assess even better the contribution of individual species and protein to the overall protein expression. These approaches do, however, come at the price of having to generate metagenomes for each sample and to process each of these into a list of protein sequences. That may not always be feasible in terms of cost and time, in which case, IGC is the next best alternative. However, large database sizes come with several issues. First, the ability to distinguish correct from incorrect matches is strongly impaired. The concept of FDR estimation as defined by Elias and Gygi [54] comes with the assumption that the database is a comprehensive representation of the real search space. This assumption is, most of the time, not justified in metaproteomic samples. One option to overcome such artificial loss of identifications is to include semi-supervised machine learning algorithms like Percolator [28] or Nokoi [55] for the PSM scoring. Another approach to circumvent the issue of large search spaces is the clustering of peptides [56,57] or the use of 2-step searches like proposed by Jagtap et al. [58]. However, some controversy exists in the field of metaproteomics as to what degree the latter method leads to an under-estimation of the true FDR. To solve this problem for metaproteomics, a major rethinking of peptide match scoring is necessary. We anticipate that substantial progress will be made when using synthetic peptides as a ground truth for training predictors of tandem mass spectra. Large collections of synthetic peptides are becoming available by initiatives, such as the ProteomeTools project [59]. The use of sample-specific sequence databases for peptide identification also controls, at least to some degree, demand for large computational power and storage capacity.

Another obvious challenge of metaproteomics is sample variability and the high proportion of missing values which impairs the use of many statistical methods that require complete data matrices. Due to their high species complexity, metaproteomic samples show generally high variability. Sample variability could be enhanced in the present patient cohort due to the administration of chemotherapy, increasing mutational load in bacteria [60], and the administration of antibiotics, altering the intestinal microbiome composition [51]. This variability may be attenuated by increasing the dynamic range of the analytical workflow, e. g., using deep fractionation of peptides prior to LC-MS/MS analysis or depletion of non-bacterial contaminants. This would, however, imply increased time requirements and cost, and the production of an even higher 'data mountain'. In addition, feasibility may not always be ensured especially for large clinical studies and low sample availability [61]. In contrast to mainstream proteomics, which makes strong use of intensity-based abundance estimation, metaproteomics is still largely confined to spectral counting methods because only few peptides are detected in many samples, which limits the accuracy with which changes can be measured [62]. In addition, most quantification methods require robust normalization of the data. In microbiology, samples are regularly normalized on the sample input weight (here feces). This may or may not be a fair representation of actual bacterial/protein amount/variation in a sample. It has become standard procedure in proteomics to normalize input material for LC-MS/MS measurements on the basis of total peptide or protein content to ensure equal depth of analysis and reproducible identification [63]. This may not be possible in metaproteomics: Comparability and normalization may be compromised because the feces may contain proteineous material other than from bacteria and host contributions. Many researcher are turning attention to data-independent-acquisition (DIA) strategies because it promises to improve reproducibility and quantification and could decrease the level of missing values. Yet, spectrum annotation and availability of suitable spectral libraries for DIA is still challenging for single proteomes and, in our view, the concepts and tools need to be much improved before DIA is applicable for complex metaproteomic samples.

An entirely different option to circumvent missing values on peptide/protein level is to compare abundances of GO terms and taxonomic distributions. Although, this shows promising results, clear taxonomic and functional annotation is not always feasible in metaproteomics. Because peptides can be shared between different proteins of the same organism or between multiple organisms, the protein inference problem in metaproteomic is even more pronounced than in single organism proteomics [21]. Unequivocal annotation to one species is, therefore, often not possible. To circumvent this problem, peptides and proteins are often mapped to the lowest common ancestor (LCA) as first described by Huson DH et al. [64] However, this clearly results in loss of information and potentially ambiguous annotations, limiting its applicability to higher phylogenetic levels, such as classes or phyla. Still, it is a very practical approach that does provide functional annotation and, thus, helps in the interpretation of metaproteomic data. This was strongly facilitated by the recent extension of the frequently used Unipept metaproteome analyzer to not only map the LCA on peptide level but also annotates peptides with GO terms and E.C. numbers. This functionality offers an alternative to other commonly used protein-based tools, such as Megan (Metagenome Analyzer) [65], eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) [66] and KEGG (Kyoto Encyclopedia of Genes and Genomes) [67]. Of note, since Unipept works at the peptide level, it simplifies data analysis by side-stepping the need of BLAST searches, especially for sample-specific genomic databases but it needs to be mentioned that, to our knowledge, no validation study comparing peptide vs. protein level based annotation has been published. Despite this, both are frequently used in metaproteomics, and there is no consensus opinion on this point in the field of metaproteomics.

This report describes the taxonomic composition and functional process of patients during the MRE gut colonization progress. Further improvements in data analysis strategies and study designs are needed to explore the processes and interactions in the microbiome and the host in more detail. Elucidating the mechanism of microbiome provided colonization resistance against multidrug-resistant pathogens (e.g., bacteriocins), the microbiome influence in disease development following transplantation (e.g., graft versus host disease) or chemotherapy efficacy. We are confident that with new technology and software, most of the challenges will eventually be solved, enabling future studies to move from merely describing taxonomic and functional composition changes to revealing significant protein-centric molecular and functional processes.

References

1. Bäckhed, F.; Ley, R.E.; Sonnenburg, J.L.; Peterson, D.A.; Gordon, J.I. Host-bacterial mutualism in the human intestine. *Science* **2005**, *307*, 1915–1920. [CrossRef] [PubMed]

2. Yoon, S.S.; Kim, E.K.; Lee, W.J. Functional genomic and metagenomic approaches to understanding gut microbiota-animal mutualism. *Curr. Opin. Microbiol.* **2015**, *24*, 38–46. [CrossRef] [PubMed]

3. Graham, R.L.J.; Graham, C.; McMullan, G. Microbial proteomics: A mass spectrometry primer for biologists. *Microb. Cell Fact.* **2007**, *6*, 1–14. [CrossRef] [PubMed]

4. Wilmes, P.; Bond, P.L. Metaproteomics: Studying functional gene expression in microbial ecosystems. *Trends Microbiol.* **2006**, *14*, 92–97. [CrossRef]

5. Rodríguez-Valera, F. Environmental genomics, the big picture? *FEMS Microbiol. Lett.* **2004**, *231*, 153–158. [CrossRef]

6. Lichtman, J.S.; Marcobal, A.; Sonnenburg, J.L.; Elias, J.E. Host-centric proteomics of stool: A novel strategy focused on intestinal responses to the gut microbiota. *Mol. Cell Proteom.* **2013**, *12*, 3310–3318. [CrossRef]

7. Kolmeder, C.A.; de Been, M.; Nikkilä, J.; Ritamo, I.; Mättö, J.; Valmu, L.; Salojärvi, J.; Palva, A.; Salonen, A.; de Vos, W.M. Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. *PLoS ONE* **2012**, *7*, e29913. [CrossRef]

8. Verberkmoes, N.C.; Russell, A.L.; Shah, M.; Godzik, A.; Rosenquist, M.; Halfvarson, J.; Lefsrud, M.G.; Apajalahti, J.; Tysk, C.; Hettich, R.L.; et al. Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* **2009**, *3*, 179–189. [CrossRef]

9. Pérez-Cobas, A.E.; Gosalbes, M.J.; Friedrichs, A.; Knecht, H.; Artacho, A.; Eismann, K.; Otto, W.; Rojo, D.; Bargiell, R.; von Bergen, M.; et al. Gut microbiota disturbance during antibiotic therapy: A multi-omic approach. *Gut* **2012**, *62*, 1591–1601. [CrossRef]

10. Chuong, K.H.; Mack, D.R.; Stintzi, A.; O'Doeherty, K.C. Human Microbiome and Learning Healthcare Systems: Integrating Research and Precision Medicine for Inflammatory Bowel Disease. *OMICS* **2018**, *22*, 119–126. [CrossRef]

11. Juste, C.; Kreil, D.P.; Beauvallet, C.; Guillot, A.; Vaca, S.; Carapito, C.; Mondot, S.; Sykacek, P.; Sokol, H.; Blon, F.; et al. Bacterial protein signals are associated with Crohn's disease. *Gut* **2014**, *63*, 1566–1577. [CrossRef] [PubMed]

12. Ferrer, M.; Riuz, A.; Lanza, F.; Haange, S.B.; Oberbach, A.; Till, H.; Bargiella, R.; Campoy, C.; Segura, M.T.; Richter, M.; et al. Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. *Environ. Microbiol.* **2013**, *15*, 211–226. [CrossRef] [PubMed]

13. Kolmeder, C.A.; Ritari, J.; Verdam, F.J.; Muth, T.; Keskitalo, S.; Varosalo, M.; Fuentes, S.; Greve, J.W.; Buurman, W.A.; Reichl, U.; et al. Colonic metaproteomic signatures of active bacteria and the host in obesity. *Proteomics* **2015**, *15*, 3544–3552. [CrossRef] [PubMed]

14. Gavin, P.G.; Mullaney, J.A.; Loo, D.; Gottlieb, P.A.; Hill, M.M.; Zipris, D.; Hamilton-Williams, E.E. Intestinal Metaproteomics Reveals Host-Microbiota Interactions in Subjects at Risk for Type 1 Diabetes. *Diabetes Care* **2018**, *41*, 2178–2186. [CrossRef] [PubMed]

15. Hammond, S.P.; Baden, L.R. Antibiotic prophylaxis for patients with acute leukemia. *Leuk Lymphoma* **2008**, *49*, 183–193. [CrossRef] [PubMed]

16. Heyer, R.; Schallert, K.; Zoun, R.; Becher, B.; Saake, G.; Benndorf, D. Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* **2017**, *261*, 24–36. [CrossRef] [PubMed]

17. Timmins-Schiffman, E.; May, D.H.; Mikan, M.; Riffle, M.; Frazar, C.; Harvey, H.R.; Noble, W.S.; Nunn, B.L. Critical decisions in metaproteomics: Achieving high confidence protein annotations in a sea of unknowns. *ISME J.* **2017**, *11*, 309–314. [CrossRef]

18. Muth, T.; Renard, B.Y.; Martens, L. Metaproteomic data analysis at a glance: Advances in computational microbial community proteomics. *Expert Rev. Proteom.* **2016**, *13*, 757–769. [CrossRef]

19. Lloyd-Price, J.; Mahurkar, A.; Rahnavard, G.; Crabtree, J.; Orvis, J.; Hall, A.B.; Brady, A.; Creasy, H.H.; McCracken, C.; Giglio, M.G.; et al. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **2017**, *550*, 61–66. [CrossRef]

20. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]

21. Kleiner, M.; Thorson, E.; Sharp, C.E.; Dong, X.; Liu, D.; Li, C.; Strous, M. Assessing species biomass contributions in microbial communities via metaproteomics. *Nat. Commun.* **2017**, *8*, 1558. [CrossRef] [PubMed]

22. Haange, S.B.; Jehmlich, N. Proteomic interrogation of the gut microbiota: Potential clinical impact. *Expert Rev. Proteom.* **2016**, *13*, 535–537. [CrossRef] [PubMed]

23. Magiorakos, A.P.; Srinivasan, A.; Carey, R.B.; Carmeli, Y.; Falagas, M.E.; Giske, C.G.; Harbath, S.; Hindler, J.F.; Kahlmeter, G.; Olsson-Liljequist, B.; et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: An international expert proposal for interim standard definitions for acquired resistance. *Clin. Microbiol. Infect.* **2014**, *18*, 268–281. [CrossRef] [PubMed]

24. Shevchenko, A.; Tomas, H.; Halvis, J.; Olsen, B.; Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **2006**, *1*, 2856–2860. [CrossRef] [PubMed]

25. Hahne, H.; Pachl, F.; Ruprecht, B.; Maier, S.K.; Klaeger, S.; Helm, D.; Médard, G.; Wilm, M.; Lemeer, S.; Kuster, B. DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. *Nat. Methods* **2013**, *10*, 989–991. [CrossRef] [PubMed]

26. Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372. [CrossRef] [PubMed]

27. Li, J.; Jia, H.; Cai, X.; Zhong, H.; Feng, Q.; Sunagawa, S.; Arumugam, M.; Kultima, J.R.; Prifti, E.; Nielsen, T.; et al. MetaHIT Consortium An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **2014**, *32*, 834–841. [CrossRef] [PubMed]

28. Käll, L.; Canterbury, J.D.; Weston, J.; Noble, W.S.; MacCoss, M.J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923–925. [CrossRef]

29. Mesuere, B.; Devreese, B.; Debyser, G.; Aerts, M.; Vandamme, P.; Dawyndt, P. Unipept: Tryptic peptide-based biodiversity analysis of metaproteome samples. *J. Proteome Res.* **2012**, *11*, 5773–5780. [CrossRef]

30. Gurdeep, S.R.; Tanca, A.; Palomba, A.; Van der Jeugt, F.; Verschaffelt, P.; Uzzau, S.; Martens, L.; Dawyndt, P.; Mesuere, B. Unipept 4.0: Functional analysis of metaproteome data. *J Proteome Res.* **2018**. [CrossRef]

31. Vizcaíno, J.A.; Côté, R.G.; Csordas, A.; Dianes, J.A.; Fabregat, A.; Foster, J.M.; Griss, J.; Alpi, E.; Birim, M.; Contell, J.; et al. The Proteomics Identifications (PRIDE) database and associated tools: Status in 2013. *Nucleic Acids Res.* **2013**, *41*, 1063–1069. [CrossRef] [PubMed]

32. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **2011**, *17*, 10–12. [CrossRef]

33. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]

34. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef] [PubMed]

35. Hyatt, D.; Chen, G.L.; LoCascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **2010**, *11*, 119. [CrossRef] [PubMed]

36. Rognes, T.; Flouri, T.; Nichols, B.; Quince, C.; Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **2016**, *18*, e2584. [CrossRef] [PubMed]

37. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef] [PubMed]

38. Kang, D.D.; Froula, J.; Egan, R.; Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **2015**, *3*, e1165. [CrossRef]

39. Parks, D.H.; Imelfort, M.; Skennerton, C.T.; Hugenholtz, P.; Tyson, G.W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **2015**, *25*, 1043–1055. [CrossRef]

40. Schloss, P.D.; Westcott, S.L.; Ryabin, T.; Hall, J.R.; Hartmann, M.; Hollister, E.B.; Lesniewski, R.A.; Oakley, B.B.; Parks, D.H.; Robinson, C.J.; et al. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **2009**, *75*, 7537–7541. [CrossRef]

41. Shannon, C.E.; Weaver, W. The mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *34*, 312–313. [CrossRef]

42. Wang, O.; Garrity, G.M.; Tiedje, J.M.; Cole, J.R. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl. Environ. Microbiol.* **2007**, *73*, 5261–5267. [CrossRef] [PubMed]

43. Tanca, A.; Palomba, A.; Deligios, M.; Cubeddu, T.; Fraumene, C.; Biossa, G.; Pagnozzi, D.; Addis, M.F.; Uzzau, S. Evaluating the impact of different sequence databases on metaproteome analysis: Insights from a lab-assembled microbial mixture. *PLoS ONE* **2013**, *8*, e82981. [CrossRef]

44. Muth, T.; Kolmeder, C.A.; Salojärvi, J.; Keskitalo, S.; Varjosalo, S.; Verdam, F.J.; Rensen, S.S.; Reichl, U.; de Vos, W.M.; Rapp, E.; et al. Navigating through metaproteomics data: A logbook of database searching. *Proteomics* **2015**, *15*, 3439–3453. [CrossRef]

45. Hershberg, R. Mutation—The Engine of Evolution: Studying Mutation and Its Role in the Evolution of Bacteria. *Cold Spring Harb. Perspect. Biol.* **2015**, *7*, a018077. [CrossRef]

46. Troung, D.T.; Tett, A.; Pasolli, E.; Huttenhower, C.; Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **2017**, *27*, 626–638. [CrossRef]

47. Savitski, M.M.; Wilhelm, M.; Hahne, H.; Kuster, B.; Bantscheff, M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Mol. Cell. Proteom.* **2015**, *14*, 2394–2404. [CrossRef]

48. Eckburg, P.B.; Bik, E.M.; Bernstein, C.N.; Purdom, E.; Dethlefsen, L.; Sargent, M.; Gill, S.R.; Nelson, K.E.; Relman, D.A. Diversity of the human intestinal microbial flora. *Science* **2005**, *308*, 1635–1638. [CrossRef] [PubMed]

49. Gill, S.R.; Pop, M.; Deboy, R.T.; Eckburg, P.B.; Turnbaugh, P.J.; Samuel, B.S.; Gordon, J.I.; Relman, D.A.; Fraser-Liggett, C.M.; Nelson, K.E. Metagenomic Analysis of the Human Distal Gut Microbiome. *Science* **2006**, *312*, 1355–1359. [CrossRef] [PubMed]

50. Zhang, X.; Deeke, S.A.; Ning, Z.; Starr, A.E.; Butcher, J.; Li, J.; Mayne, J.; Cheng, K.; Liao, B.; Li, L.; et al. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nat. Commun.* **2018**, *9*, 2873. [CrossRef] [PubMed]

51. Cebula, T.A. Genetic and physiological modulation of anthracycline-induced mutagenesis in Salmonella typhimurium. *Environ. Mutagen.* **1986**, *8*, 675–692. [CrossRef] [PubMed]

52. Riffle, M.; May, D.H.; Timmins-Schiffman, E.; Mikan, M.P.; Jaschib, D.; Noble, W.S.; Nunn, B.L. MetaGOmics: A Web-Based Tool for Peptide-Centric Functional and Taxonomic Analysis of Metaproteomics Data. *Proteomes* **2018**, *6*, 2. [CrossRef] [PubMed]

53. Kolmeder, C.A.; de Vos, W.S. Metaproteomics of our microbiome—Developing insight in function and activity in man and model systems. *J. Proteom.* **2014**, *97*, 3–16. [CrossRef]

54. Elias, J.E.; Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *604*, 55–71. [CrossRef] [PubMed]

55. Gonnelli, G.; Stock, M.; Verwaeren, J.; Maddelein, D.; De Baets, B.; Martens, L.; Degroeve, S. A decoy-free approach to the identification of peptides. *J. Proteome Res.* **2015**, *14*, 1792–1798. [CrossRef] [PubMed]

56. Marx, H.; Lemeer, S.; Klaeger, S.; Rattei, T.; Kuster, B. MScDB: A mass spectrometry-centric protein sequence database for proteomics. *J. Proteome Res.* **2013**, *12*, 2386–2398. [CrossRef] [PubMed]

57. May, D.H.; Timmins-Schiffman, E.; Mikan, M.P.; Harvey, H.R.; Borenstein, E.; Nunn, B.L.; Noble, W.S. An alignment-free 'metapeptide' strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. *J. Proteome Res.* **2016**, *15*, 2697–2705. [CrossRef] [PubMed]

58. Jagtap, P.; Goslinga, J.; Kooren, J.A.; McGowan, T.; Wroblewski, M.S.; Seymour, S.L.; Griffin, T.J. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **2013**, *13*, 1352–1357. [CrossRef]

59. Zolg, D.P.; Wilhelm, M.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Bailey, D.J.; Gessulat, S.; Ehrlich, H.C.; Weininger, M.; et al. Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **2017**, *14*, 259–262. [CrossRef]

60. Papanicolas, L.E.; Gordon, D.L.; Wesselingh, S.L.; Rogers, G.B. Not Just Antibiotics: Is Cancer Chemotherapy Driving Antimicrobial Resistance? *Trends Microbiol.* **2018**, *26*, 393–400. [CrossRef]

61. Xiong, W.; Giannone, R.J.; Morowitz, M.J.; Banfield, J.F.; Hettich, R.L. Development of an enhanced metaproteomic approach for deepening the microbiome characterization of the human infant gut. *J. Proteome Res.* **2014**, *14*, 133–141. [CrossRef] [PubMed]

62. Lee, P.Y.; Chin, S.F.; Neoh, H.M.; Jamal, R. Metaproteomic analysis of human gut microbiota: Where are we heading? *J. Biomed. Sci.* **2017**, *24*, 36. [CrossRef] [PubMed]

63. Griffin, N.M.; Yu, J.; Long, F.; Oh, P.; Shore, S.; Li, Y.; Koziol, J.A.; Schnitzer, J.E. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol.* **2010**, *28*, 83–89. [CrossRef] [PubMed]

64. Huson, D.H.; Mitra, S.; Ruscheweyh, H.J.; Weber, N.; Schuster, S.C. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* **2011**, *21*, 1552–1560. [CrossRef] [PubMed]

65. Huson, D.H.; Weber, N. Microbial community analysis using MEGAN. *Methods Enzymol.* **2013**, *531*, 465–485. [CrossRef] [PubMed]

66. Huerta-Cepas, J.; Szklarczyk, D.; Forslund, K.; Cook, H.; Heller, D.; Walter, M.C.; Rattei, T.; Mende, D.R.; Sunagawa, S.; Kuhn, M.; et al. EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **2016**, *44*, 286–293. [CrossRef] [PubMed]

67. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **2016**, *44*, 457–462. [CrossRef] [PubMed]