Computational and Structural Biotechnology Journal 20 (2022) 274–286

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Short review

Inferring early-life host and microbiome functions by mass spectrometry-based metaproteomics and metabolomics

Veronika Kuchařová Pettersen [a,b,c], Luis Caetano Martha Antunes [d,e], Antoine Dufour [f], Marie-Claire Arrieta [f,g,h,*]

[a] Research Group for Host-Microbe Interactions, Department of Medical Biology, UiT The Arctic University of Norway, Tromsø, Norway
[b] Pediatric Research Group, Department of Clinical Medicine, UiT The Arctic University of Norway, Tromsø, Norway
[c] Centre for New Antibacterial Strategies, UiT The Arctic University of Norway, Tromsø, Norway
[d] Oswaldo Cruz Institute, Oswaldo Cruz Foundation, Rio de Janeiro, RJ, Brazil
[e] National Institute of Science and Technology of Innovation on Diseases of Neglected Populations, Center for Technological Development in Health, Oswaldo Cruz Foundation, Rio de Janeiro, RJ, Brazil
[f] Department of Physiology & Pharmacology, University of Calgary, Calgary, Canada
[g] Department of Pediatrics, University of Calgary, Calgary, AB, Canada
[h] International Microbiome Centre, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

ARTICLE INFO

Article history:
Received 14 August 2021
Received in revised form 8 December 2021
Accepted 8 December 2021
Available online 20 December 2021

Keywords:
Metaproteomics
Metabolomics
Early life human microbiome
Microbial colonisation
Metagenomics

ABSTRACT

Humans have a long-standing coexistence with microorganisms. In particular, the microbial community that populates the human gastrointestinal tract has emerged as a critical player in governing human health and disease. DNA and RNA sequencing techniques that map taxonomical composition and genomic potential of the gut community have become invaluable for microbiome research. However, deriving a biochemical understanding of how activities of the gut microbiome shape host development and physiology requires an expanded experimental design that goes beyond these approaches.

In this review, we explore advances in high-throughput techniques based on liquid chromatography–mass spectrometry. These omics methods for the identification of proteins and metabolites have enabled direct characterisation of gut microbiome functions and the crosstalk with the host. We discuss current metaproteomics and metabolomics workflows for producing functional profiles, the existing methodological challenges and limitations, and recent studies utilising these techniques with a special focus on early life gut microbiome.

1. Introduction

The mammalian gastrointestinal tract accommodates one of the densest microbial populations known, the gut microbiome. Each mammalian species, including humans, has a unique microbial community that has coevolved with its host and is finely adapted to the species lifestyle [1]. The trillions of microbial cells, including bacteria, fungi, protozoa, archaea, as well as viruses, all take advantage of the nutrient-rich gut environment, but it is mainly bacteria for which there is evidence of benefits being provided to host physiology. Commensal bacteria augment host functions by breaking down indigestible food components, synthesising essential vitamins, stimulating the immune system, and protecting against invading pathogens [2–4]. Still, the nature of the relationship mammalian hosts share with their gut microbiomes is convoluted, and research has so far elucidated only initial clues of the functions involved in microbiome-host crosstalk.

The gut microbiome has been linked to the development and progression of both infectious [5] and chronic non-communicable diseases [6,7], including cancer [8], autoimmune [9], and neurological disorders [10]. Practical knowledge about the gut microbiome is highly relevant for medicine because characteristics of the gut microbiome can be used as a complementary tool to clinical diagnosis and be a target for therapeutic interventions by itself. By being a diagnostic adjunct, microbially derived biomarkers could inform on treatment response [11], serve as a window into the side-effects of antibiotics [12] and other drugs [13], or be a baseline measurement before therapy initiation [14]. Most importantly,

* Corresponding author at: University of Calgary, Health Research Innovation Centre, 3330 Hospital Drive N.W., Calgary T2N 4N1, Alberta, Canada.
E-mail address: marie.arrieta@ucalgary.ca (M.-C. Arrieta).

https://doi.org/10.1016/j.csbj.2021.12.012

because of its inherent connection to human physiology, the gut microbiome will serve as a signature of diseases, based on which targeted therapeutic interventions such as guided nutritional plans can be recommended [15].

In this minireview, we outline advances in gut microbiome characterisation using high-resolution liquid chromatography–mass spectrometry (LC–MS) for large-scale profiling of proteins and metabolites. Initially, we discuss steps in a typical workflow used in LC–MS-based metaproteomics and metabolomics (Fig. 1). Although there are differences in microbial colonisation and dissimilar protein and metabolite profiles along the gastrointestinal tract [16], we concentrate on stool-based approaches because of their application for biomarker discovery and non-invasive nature. Furthermore, feces are a heterogeneous material rich in various macromolecules and small metabolites, introducing challenges for analysis using instrumental methods and subsequent computational workflows. We conclude the review with recent studies using LC–MS omics for gut microbiome characterisation in the pediatric population (Table 1), which have enabled deeper biological insights on microbe-microbe and host-microbe interactions during early life.

2. Need for functional description of the gut microbiome

The potential for exploiting the gut microbiome in biomedical applications is immense; however, host-microbiome molecular interactions are still largely uncharacterised. This is in part because of the microbiome multi-layered complexity. The gut microbiome configuration depends on a metabolically active microbial community (microbiota), which dynamically responds to fluctuating physio-chemical properties of the gut [17], the host control of its composition [18], and other potential factors influencing the host-microbiota interactions such as microbial pathogens and medications. Additionally, only a small portion of human gut microorganisms have been cultured in specialized laboratories, and most of the microbiota remains uncharacterised using cultivation techniques. Because of the latter, culture-independent approaches such as profiling taxonomical marker genes [16S ribosomal RNA gene for bacteria and the internal transcribed spacer (ITS) region for fungi] have gained a major foothold among methods for microbiota characterisation. Although amplicon sequencing based on 16S and ITS is limited to describing taxonomic composition, researchers can use bioinformatics methods such as PICRUSt (https://picrust.github.io/picrust) to predict the microbial community functional profiles based on the taxa found [19].

Amplicon sequencing and predictive functional profile tools are restricted by low power for taxonomy resolution. Therefore more robust whole genome shotgun sequencing is used to answer specific biological questions about less abundant taxa [20], interindividual strain transfer [21], or prevalence of gene families such as those involved in antimicrobial resistance [22]. Besides delivering more refined information on the microbiome taxonomic composition, metagenomics gives insights into the functional capabilities of the microbiome by profiling the relative abundances of genes within the microbial community. Still, similar to other omics strategies, many challenges remain, including efforts to answer questions about lower abundant taxa. Aspects such as sequencing depth, human DNA content removal, and targeted enrichment methods for less abundant microbial taxa need to be therefore considered in the design of metagenomics experiments [23].

DNA sequencing techniques will continue to be indispensable in microbiome studies. Still, conclusions about microbiome function derived from metagenomics predictions must be treated as hypotheses requiring functional validation [24]. Despite an earlier belief that the gut microbiome functional profile is more stable and generally conserved, based on the bioinformatic annotation of

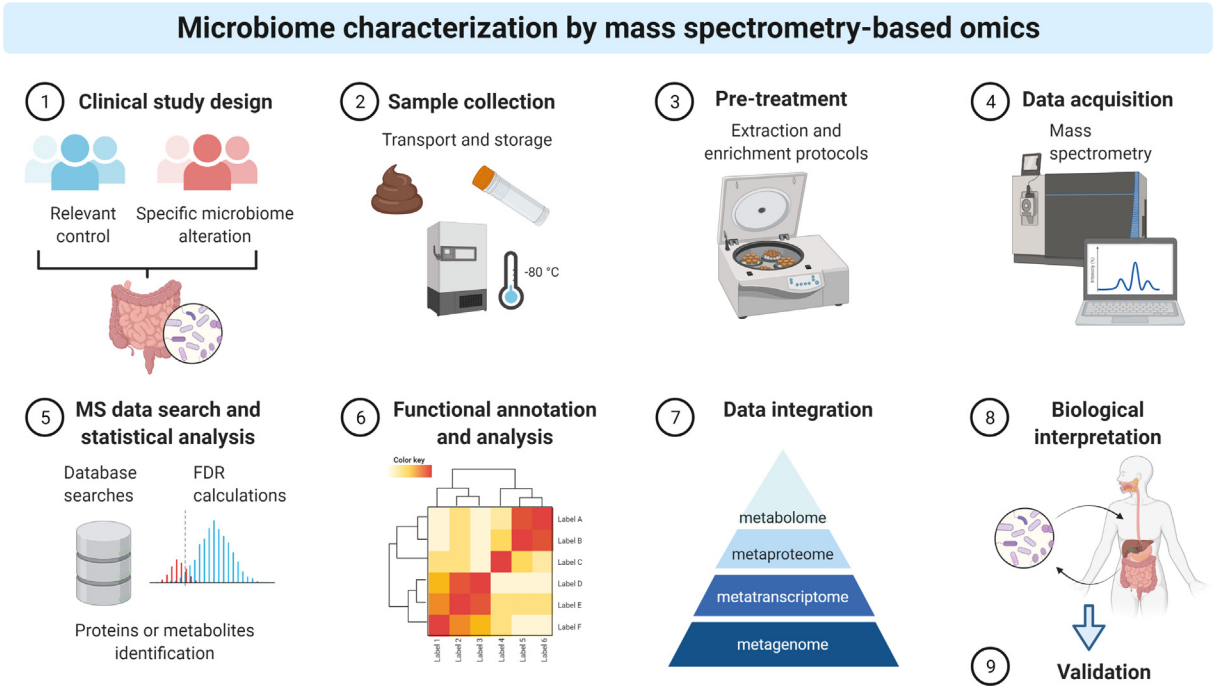

Fig. 1. Key steps during functional investigations of the human microbiome by techniques based on liquid chromatography-mass spectrometry. The workflow starts with a robust design of a clinical study and experimental controls. Sample transport chain, storage, and pretreatment methods need to be carefully evaluated as any of these steps might influence the composition of microbial cells and different biomolecules. Mass spectra acquisition is followed by searching the data against a sample-specific database and statistical filtering of false-positive matches. Metaproteome and metabolome datasets can be analysed by different bioinformatics and statistical approaches to extract biological information (see text for details). Finally, further experimental design is needed to validate identified proteins and metabolites significantly associated with a specific phenotype. Figure was created with Biorender.com.

Table 1
Metaproteomic and metabolomic studies describing early life gut microbiome functions.

Reference	Main objectives(s) with highlighted LC–MS techniques	Study Population[2]	Samples Collected (Age)	Sample Storage and Pre-processing	LC–MS/MS analysis; instrument, software and database used	Detected peptide, proteins, or metabolites	Key findings
Henderickx et al. 2021 [34]	To characterise GIT functionality and maturation of preterm infants by GIT enzyme activity assays and metaproteomics.	Preterm infants n = 40 (GA 24–33), term infants n = 3 (GA 37–42)	Gastric aspirates (PW 1–2), feces (PW 1–6)	Samples were frozen at −20 °C after collection, and stored at −80 °C.	nano-LC–LTQ-Orbitrap-MS; MaxQuant; in-house database based on 16S rRNA sequencing and the Human Microbiome Project reference genomes	89,294 unique proteins, 2317 protein groups (886-human or bovine, 1431-bacterial)	The fecal proteome of preterm infants was deprived of GIT barrier-related proteins compared to term infants. In preterm infants, bacterial oxidative stress proteins were increased compared to term infants and higher birth weight correlated with higher relative abundance of bifidobacterial proteins.
Lay et al. 2021 [160]	To elucidate characteristics (metabolome, 16S rRNA profile, metagenome, metatransciptome) of a compromised microbiome and study the role of a synbiotic in microbiome restoration.	127 infants born by elective C-section 26 vaginally born infants	Feces (PW 1–22)	Individual stool samples were lyophilized and equal amount of dry weight was combined to prepare a pool sample for each treatment group.	UPLC–MS(QExactive) KEGG and HMDB for metabolomics	Not given	Gut microbiome acquired during elective C-section birth was adapted to a more oxidative environment characterised by reactive oxygen species metabolism, biosynthesis of lipopolysaccharides and the absence of detection of genes, transcripts involved in the metabolism of milk carbohydrates.
Petersen et al. 2021 [119]	Investigation of the meconium metabolome to identify components of the neonatal gut niche that contribute to allergic sensitization.	100 infants of the CHILD study	Meconium -the first stool passed after birth	Not reported besides storage at −80 °C before metabolomic analysis	UPLC–MS/MS Proprietary analysis done at Metabolon, Inc.	714 metabolites	Newborns who develop immunoglobulin E-mediated allergic sensitization by 1 year of age had a less-diverse gut metabolome at birth, and specific metabolic clusters were associated with both protection against atopy and the abundance of key taxa driving microbiota maturation.
Cortes et al. 2019 [172]	To develop metaproteomics approach for assessment of biological phenotype and metabolic status, as a functional complement to DNA sequence analysis.	8 infants	Feces, one timepoint (2–5 months of age)	4 °C for 1 h, homogenised stool aliquots kept at −80 °C Differential centrifugation to enrich for bacterial cells	Fractionation of the peptide mixes by strong cation exchange chromatography; nanoAcquity UPLC–MS (Q Exactive); Mascot software; Custom database based 16S rRNA sequencing	15,250 unique peptides, 2154 protein groups	Metaproteomics data yielded more refined information on microbial composition than 16S rRNA gene sequencing of the same samples.
Levan et al. 2019 [36]	To test whether elevated faecal concentrations of 12,13-diHOME identified in infants by targeted metabolomics promote allergic inflammation in experimental models.	91 infants	Feces (first month of life)	Initial condition of storage not given, later stored at −80 °C	LC–MS (LTQ-Orbitrap-XL)	Faecal oxylipin (9,10-diHOME and 12,13-diHOME)	An increase in the copy number of bacterial epoxide hydrolase genes linked to 12,13-diHOME production, or the concentration of 12,13-diHOME in the faeces of neonates was found to be associated with an increased probability of developing atopy, eczema or asthma during childhood.
Brown et al. 2018 [35]	To study the premature infant gut colonization process by metagenomics and metaproteomics.	35 preterm infant (GA 24–32)	Feces (first 3 months of life)	Direct freezing at −80 °C	Microbial cells enrichment by filtration; LC–MS/MS (LTQ-Orbitrap Elite MS); MyriMatch v2.1; Matched metagenome-based database:	8691 protein families	Infants were found to be colonized by similar microbes, but each underwent a distinct colonization trajectory. Related microbes colonizing different infants were found to have distinct proteomes, indicating that microbiome function is not only driven by which organisms are present, but also largely depends on microbial responses to the unique set of physiological conditions in the infant gut.
Zwittink et al. 2017 [71]	To study microbiota development during the first six weeks in preterm infants by 16S-rRNA gene sequencing and metaproteomics, and to identify the factors associated with this development.	10 preterm infants (GA 25–30)	Feces (PW 1–6)	Direct freezing, temporal storage at −20 °C until transfer to −80 °C	nano-LC–LTQ-Orbitrap-MS, MaxQuant Custom database based on the bacterial part of the Human Microbiome Project (Uniprot)-87 bacterial species, 438,537 sequences	953 bacterial proteins	GA-dependent microbial signature differentiated between extremely preterm (25–27 GA) and very preterm (30 GA) infants. In very preterm infants, the intestinal microbiota developed toward a Bifidobacterium-dominated community and associated with high abundance of proteins involved in carbohydrate and energy metabolism. Extremely preterm infants remained predominantly colonized by facultative anaerobes and associated with proteins involved in membrane transport and translation.
Young et al. 2015 [170]	To determine time-dependent functional signatures of microbial and human proteins during early colonization of the gut.	One preterm infants (GA 28)	Feces (PW 1–3)	Immediately stored at −80 °C until analysis	nano-2D-LC–MS/MS (LTQ Orbitrap Velos); SEQUEST & DTASelect, Database derived from metagenome data	16,605 peptides, and 4031 proteins (per run)	Detected human proteins included those responsible for epithelial barrier function and antimicrobial activity. Neutrophil-derived proteins increased in abundance, suggesting activation of the innate immune system.

Abbreviations used: GA - gestational age; GIT - gastrointestinal tract; HMDB - Human Metabolome Database, PW - Postnatal week.

putative protein-coding genes [25], studies measuring mRNA or proteins have demonstrated that the metatranscriptome and metaproteome display greater variability and sensitivity to perturbation when compared to the information content of the metagenome [26–28]. This is partly because of an imperfect coupling of the gut microbiome composition and function [29], which stems from complex regulatory networks along the gene-transcript-protein expression path. Although metatranscriptomics gives greater insights into the functional potential of the microbial community than metagenomics [26], not all transcripts are translated to proteins in the same manner. For example, timing of expression (transcriptional regulation) and various mechanisms of post-transcriptional regulation, such as differences in mRNA stability, will affect transcript levels [30]. Similarly, protein abundance is a combined result of protein synthesis and degradation, the latter being ignored in metatranscriptomics. Accordingly, a popular strategy to gain insights into the microbiome function has been integration of DNA- or RNA-based information with high-throughput measurements of microbial metabolic products and proteins, *i.e.*, metabolomics and metaproteomics.

3. Proteins and metabolites as microbiome functional descriptors

Each microbial cell responds to the unique physicochemical conditions of the host by adjusting its protein synthesis, metabolism, and secretion of biomolecules that facilitate its adaptation to the environment. Proteins carry out most functions in the cell (e.g., catalysis of biochemical reactions, transport, maintenance of cell structure), and protein amounts reflect the cell's most recent activities. Metaproteomics, the characterisation of the entire set of proteins accumulated by all community members at a given point in time [31], has emerged as the most relevant approach to characterise gut microbiome function. In addition, metaproteomics can simultaneously detect host and microbial proteins and aid in the characterisation of host-microbiome interactions [32]. Besides proteins, the collection of small molecules found in feces, the fecal metabolome, can be seen as a recording of the recent chemical communication between the microbial community and its host. Metaproteomics and metabolomics thus provide insight into the metabolic and physiological state of both the host and microbiome, and give a direct description of their phenotypes (Fig. 2).

Functional characterisation of stool is an attractive option to assess human health and disease due to the non-invasive sampling nature and broad coverage of biomolecules reflecting different physiological processes. Both metaproteomics and metabolomics have been used in clinical research to discover biomarkers that might facilitate early detection and diagnostics of various diseases. For example, several studies demonstrated the potential of proteins and peptides present in stool as biomarkers for colorectal cancer and other bowel-related diseases in the adult population [33]. In the pediatric population, a few metaproteomics studies reported findings on promising protein biomarkers for gastrointestinal tract maturation [34,35]. Further, detection of metabolites identified as key mediators of the interactions between the gut microbiome and the host during early life is critical for disease prevention. A potential biomarker for early prediction of disease risk is 12,13-diHOME, a linoleic acid metabolite produced by certain gut bacteria that was elevated in neonates who developed asthma during childhood [36]. On the other hand, indole-3-lactic acid has been associated with beneficial microbiota in infants, decreased inflammation in intestinal epithelial cells [37], and beneficial immunoregulation [38]. However, the above-mentioned metabolites were identified in small cohorts, and future studies must address their validation on a larger number of clinical samples. Overall, although there are still limited numbers of metaproteomics and metabolomics studies of human diseases, the methodologies and available analytical tools have been recently greatly refined and encourage further in-depth characterization of the gut microbiome.

4. Mass spectrometry-based metaproteomics and metabolomics

General principles. The combination of liquid chromatography (LC) and mass spectrometry (MS) is a powerful analytical method for large-scale identification and quantification of biomolecules. LC–MS can be used in a global discovery mode to identify thousands of compounds or in a targeted manner for detecting specific analytes at levels of a few parts per billion [39]. In a prototypical LC–MS experiment, a solution containing analytes of interest is first separated on an LC column to reduce sample complexity. Then, the LC effluent is directed to the mass spectrometer, where it is nebulised, desolvated, and ionised by an ionisation source, allowing small biomolecules to enter the gas phase as charged particles. By applying electromagnetic fields, the charged particles migrate under a high vacuum through a series of mass analysers, where they are sorted according to their mass-to-charge ratio (m/z). The resulting peak patterns define a fingerprint of the original sample. Tandem mass spectrometry (MS/MS), an analytical setup where two or more MS acquisitions are arranged sequentially, is especially useful for analysing complex biological mixtures and when greater certainty of analyte identification is desired. In the first MS, precursor ions of selected m/z are isolated from the rest of the ions and fragmented by collision with an inert gas into product ions, which are mass analysed in the second MS. This transition from precursor to product ions is specific for each compound and distinguishes even minor changes in molecular structure [40]. Although this method provides a high degree of selectivity, some highly similar isomers still cannot be distinguished, and additional information (discussed in chapter 6 – Metabolite identification) or alternative methods (nuclear magnetic resonance) are required for structural elucidation.

Data acquisition. Data-dependent acquisition (DDA) and data-independent acquisition (DIA) are the two standard modes used in the untargeted identification of biomolecules based on high-resolution mass spectrometry [41,42]. When using DDA, MS/MS data acquisition occurs sequentially, and the resulting data are used to search an existing database. The main DDA advantages include 1) a simpler setup, 2) a need for less computational resources, and 3) a more sensitive quantification than DIA. However, the main issue with DDA has been lower precision and reproducibility and undersampling of low-abundance analytes compared to DIA. In the DIA mode, MS/MS data acquisition occurs in parallel across analytes, and the resulting MS spectra are highly multiplexed. In contrast to DDA, all analytes are analyzed during the second stage of tandem MS, resulting in no need for an *a priori* knowledge of the sample composition. Moreover, DIA can quantify analytes in complex mixtures over a large dynamic range, thereby overcoming the challenge of undersampling when using DDA. One of the current challenges of DIA is an unmet need for tools and software that can be used to deconvolute the complex spectra produced. In this review, we will primarily discuss studies using the DDA approach.

Data analyses. An LC–MS run produces raw data, which need to be denoised, peak-picked, feature-detected, deisotoped, and deconvoluted before analyte identification [43–45]. These preprocessing steps are crucial as any errors produced during the initial stage will propagate throughout the analysis. Data preprocessing methods are constantly being improved [46,47] and are an integral part of proteomics and metabolomics software packages, which

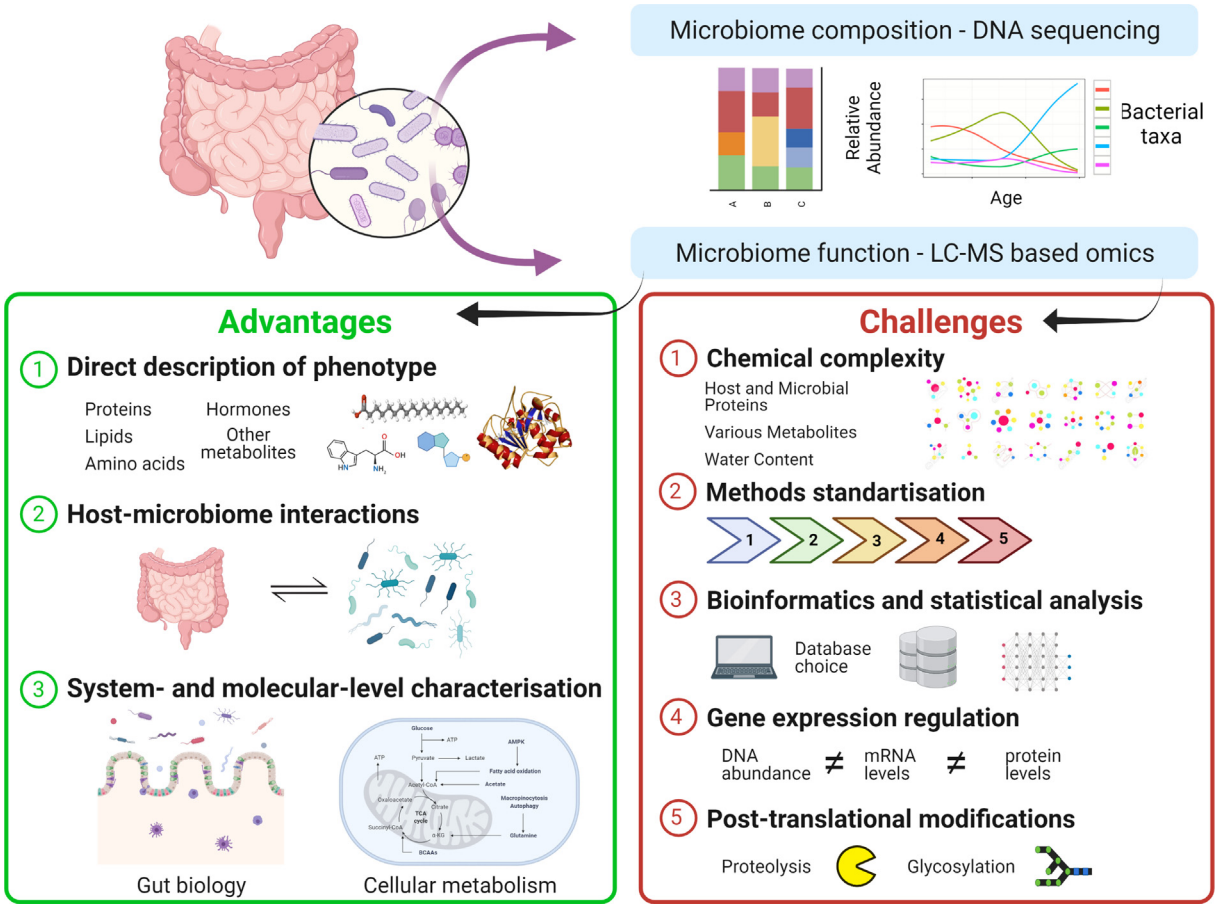

Fig. 2. Advantages and challenges of liquid chromatography-mass spectrometry (LC–MS) omics. Metaproteomics and metabolomics complement other meta-omic approaches such as metagenomics that assess the diversity and functional potential of microorganisms but cannot observe their actual phenotypes. Further, metaproteomics and metabolomics can identify proteins and metabolites originating from either the host or microbiome and give indications of their interactions. However, a wide range of metabolites is common to the human host and gut microbes and thus not possible to discriminate by metabolomics. A significant advantage of LC–MS omics is their ability to characterise cellular metabolism at the molecular level for different microbial species and provide system-level information for the host. Besides these advantages, five challenges of LC–MS omics are listed. These include the chemical complexity of fecal samples, lack of standardisation, and especially bioinformatics and statistical challenges associated with large datasets. Moreover, even if different omics analyses are done on the same sample, complex gene expression regulation processes will hinder direct comparison between DNA abundance and the levels of transcripts and proteins, and consequently the omics data interpretation. Also, complex sample preparation protocols and incomplete databases on proteins cleavage sites and other post-translational modifications are hindering the use of metaproteomics in the discovery of novel regulatory mechanisms. In metaproteomics a formidable issue is the assignment of shared peptides to proteins that originate from different microbial species. Finally, both metabolomics and metaproteomics face the challenge of low abundant molecules detection in complex mixtures. For the sake of clarity, the last two points are not illustrated in the figure. Figure was created with Biorender.com.

search the mass spectra against a database of candidate biomolecules and compare the experimental observations to theoretical patterns (discussed in detail in chapters 5 and 6). However, due to noisy data containing high background signals and incomplete databases, analyte identification is prone to false positives and mismatches. Statistical analyses that assign quality metrics are therefore needed to ascertain the significance of analytes identifications. Several reviews have summarised recent metaproteomics [48,49] and metabolomics [50,51] software and addressed in detail issues associated with database size and completeness, demand for computational power, and identification of false-positive matches.

Finally, all areas of mass spectrometry applications are currently being challenged when it comes to the standardisation of analytical workflows [52] as well as data and method transparency. Nevertheless, developments in the metagenomics community [53,54] predict that the potential of mass spectrometry applications can be fully realised and LC–MS techniques more widely adopted as long as the community guidelines [55] and the FAIR, *i.e.*, findable, accessible, interoperable, and reusable, principles are followed for reporting of methods, data, and software [56].

5. Current challenges of stool-based metaproteomics

Sample storage and processing. Metaproteomics and metabolomics approaches based on LC–MS technology share similar sample collection and processing workflows, yet unique methodological and computational challenges exist for both techniques (Fig. 2). Among the biggest challenges of stool-based metaproteomics is sample processing. In addition to carrying a complex microbial community, the stool matrix comprises undigested food particles and various host components (see chapter 6 for the macromolecular composition of feces). Therefore, an appropriate sample processing protocol needs to be evaluated in the context of each study's aims and should consider unbiased methods for storage of collected samples, microbial protein enrichment, and protein extraction efficiency [57,58]. Sample storage is a crucial step in any omic study because different storage temperatures introduce alterations to microbial profiles [59,60]. Frozen intact stool material is more stable than frozen extracted proteins [61] and thus recommended for long-term storage. Several studies tested preservatives that maintain sample integrity at room temperature when immediate freezing is not possible, and the results indicated

RNA*later* as suitable for metaproteome preservation [62,63]. However, these studies only examined environmental samples, and the effects of preservatives have not yet been characterised for stool-derived metaproteomes.

Further, different enrichment methods, such as strategies based on double filtering [64] and differential centrifugation [65], have been applied to concentrate microbial cells from stool samples. The differential centrifugation step was later shown to cause non-specific removal of microbial cells and proteins [66]. Stool without pretreatment thus likely provides the best representation of the microbial proteins. Finally, differences in cell membranes between microbes require combining chemical and physicomechanical methods to ensure proper disruption of different cell types and consequently optimal metaproteome coverage [67].

Protein databases. Similar to proteomics, metaproteomics aims to identify and quantify all proteins in a sample, but in addition, each protein has to be correctly assigned to a microbial species [55]. Proteins extracted from stool samples are first digested into peptides whose smaller size is better suited for LC–MS analysis. The most common approach for peptide identification is matching the experimental MS/MS spectra against theoretical fragmentation patterns of peptides derived from *in silico* digestion of a protein sequence database [48,68,69]. Currently, shared peptides originating from homologous proteins remain a challenge when searching for protein IDs from a specific species and this complexity is greatly enhanced when profiling the microbiome.

The success of peptide identifications depends on the provided database, making the protein database selection crucial in any proteomic workflow [69]. Estimations for fecal samples suggest the presence of 200,000 [55] to 1,000,000 [70] proteins, leading to enormous sequence databases that bring associated bioinformatics challenges. The larger the protein database is, the lower the sensitivity of identifications, the higher the computational requirements and the chance of false-positive matches. Hence, more tailored databases give better results, and ideally, spectral searches should be performed using matched metagenome or metatranscriptome databases derived from the same sample. Although metagenome-based databases have several drawbacks, such as being prone to sequencing and assembly errors, often lacking useful sequence annotation, and introduction of additional costs, the benefit of increased protein identification rates outweighs these potential pitfalls [69]. Alternatively, the use of 16S-guided metaproteome databases is a practical solution. For example, a custom-made library based on representative bacterial genera identified by 16S rRNA sequencing [71] was compiled from reference proteomes (http://www.uniprot.org/proteomes/) of species within these genera and merged into one database together with the human proteome.

Coverage. One of the obstacles hindering a wider use of metaproteomics is low coverage of the expected metaproteome. Currently, up to 60,000 protein groups have been identified in individual metaproteomics studies [72,73], which might correspond only to a fraction (∼15-25%) of the expected proteins in the adult gut microbiome [55,70]. The gut microbiome of an adult might contain ∼1000 bacterial species and ∼10 million genes [74]. Using protein abundances from metaproteomics analysis of a patient cohort of pediatric inflammatory bowel disease [72], Zhang and Figeys estimated that over 90% of gut microbiome-derived biomass comes from less than 100 most abundant species [55], while the rest of the species is identified only with one or two peptides. They further emphasized the need for techniques that increase protein identification for low abundance microbial taxa. Among these methods is a combination of stable isotope labeling with activity-based probe enrichment that allows for quantification of low-abundance proteins with specific functionalities, and which was recently used in an animal study [75]. The gut microbiome of children, and infants in particular, displays a lower species richness and overall microbial diversity than adults [76], alluding that more complete metaproteome coverage can be achieved even with present-day metaproteomics workflows.

A case study recently demonstrated how low abundance microbial taxa, fungal species in this example, affect microbiome dynamics in a preterm infant [77]. Using a strategy of two bioinformatics pipelines for deriving eukaryotic and prokaryotic metagenomes, and creating a custom-built database composed of the concatenated metagenome-derived predicted proteomes, the authors described unique interactions between the fungus *Candida parapsilosis* and the bacterium *Enterococcus faecalis* within the infant gut microbiome. Similarly, our recent findings from a gnotobiotic study of germ-free mice colonised with defined consortia of bacterial and fungal species showed that metaproteomics could describe interkingdom interactions in the gut microbiome with high resolution [78]. The results from this animal study further highlighted that genome-matched databases are critical for the correct assignment of proteins to individual species. For 12 bacterial species with sequenced genomes, which were used as templates for the database construction, MS searches yielded relatively high coverage of the bacterial proteomes. However, for the fungal strains that did not have sequenced genomes and only general, species-specific databases were available, decreased specificity of MS data searches and lower coverage of the fungal proteomes were achieved.

Bioinformatics. Dedicated bioinformatics software tools have been developed and used to deal with the computing demands of large database searches, including MetaLab [79], MetaProteomeAnalyzer [80], PEAKS [81], Galaxy-P [82], and ComPIL [83]. Two approaches have proven particularly useful for improving the identification rate: combining multiple search engines that match the theoretical spectra to the measured ones [52] and iterative search strategies that significantly speed up the database search process [84]. Another search strategy based on multi-staged filtering of peptide-spectrum matches has been implemented in the ProteoStorm tool [85]. A percentage of the identified peptide-to-spectrum matches will eventually be false positives, which need to be distinguished from the correct matches. The proportion of false-positive identifications is usually controlled by searching a decoy database containing reversed or scrambled protein sequences and calculating the false discovery rate threshold. However, the target-decoy approach is less sensitive in metaproteomics because of the large search space and high sequence similarity between many proteins, especially proteins from different taxa with the same function [69]. Alternatives include the use of machine learning approaches for modelling incorrect peptide-to-spectrum matches [86,87]. These approaches distinguish correct and incorrect peptide-to-spectrum matches using a classifier based on learning algorithms from real data. Still, despite recent advances in big data analyses and newly available software tools, bioinformatics assessments of metaproteomics data remains a formidable challenge.

Protein and species inference. A nontrivial task in metaproteomics, which follows peptide identification and validation, is peptide-to-protein-to-microbial species inference. Due to many similar proteins resulting from closely related species and horizontal gene transfer events within the microbiome, a peptide identification can potentially be matched to several proteins from different taxa. This is a major issue if metaproteomics data are used for species quantification; for example, using peptides shared by two microbial taxa will result in an overestimation of the taxon's abundance [88]. Therefore, the use of highly specific protein inference criteria is recommended if the aim is to accurately quantify microbial taxa abundance.

Furthermore, longer peptides are more likely to be unique to a single protein, while short peptides often match multiple proteins. It is therefore advantageous to optimise the mass spectrometer acquisition settings for preferential analysis of longer peptides [89]. The protein inference problem can be further mitigated by grouping together proteins inferred from the same set of identified peptides. The proteins of the same group usually exhibit the same function but have different taxonomic origins; therefore, the taxonomic origin of the entire protein group can be described by the lowest common ancestor within a phylogenetic tree [48]. Other methods for taxonomical annotation of metaproteomic data include the use of taxon-specific peptides, such as UniPep and ProteoClade [90,91].

Post-translational modifications. Identification of post-translational modifications (PTMs) is an aspect of metaproteomics that can inform on regulatory mechanisms within different microbial taxa or host cells. The study of PTMs using proteomics typically requires an enrichment or depletion step, and a limited number of studies have profiled PTMs in the human gut environment from intestinal biopsies or stool samples. For example, a pioneering study has used a peptide immuno-affinity enrichment strategy to profile an abundant PTM in prokaryotes, lysine acetylation, in the gut microbiome [92]. The study identified lysine-acetylated sites on both host and microbial proteins that were differentially abundant in patients with Crohn's disease and healthy controls. Another form of PTMs is proteolytic processing of proteins by proteases, which act in concerted networks to amplify regulatory signals and are hypothesized to be molecular effectors involved in all aspects of biology, including microbiota homeostasis [93]Dysregulated proteolysis is often implicated in the initiation of inflammation but also persist in chronic inflammatory diseases [94,95]. Using an N-terminomics approach that enriches N-termini to determine protein cut by proteases, TAILS (terminal amine isotopic labelling of substrates) was used to profile human colonic mucosal biopsies where over 1642 human N-termini were identified [96]. Using the bioinformatics software TopFIND [97], cleavage peptide positions was compared to known proteolytic processing preferences of human, bacterial, fungal and viral protease using the MEROPS database (https://www.ebi.ac.uk/merops/index.shtml). Interestingly, based on the reported site of cleavage preferences, the predicted proteolytic activity was identified to be potentially from human proteases (63%), followed by bacterial (27%), fungal (7%), and viral sources (3%) [96]. It is important to mention that proteases cleavage sites are largely uncharacterized; therefore, such analysis is likely to change as more information is added to the MEROPS database. Furthermore, little is known about the key PTMs involved in microbiome homeostasis, their provenance (human vs bacterial, fungal, or viral) and their roles in promoting human pathologies.

Protein Annotation. Another important aspect of metaproteomics data analysis is protein functional annotation. With a well-annotated metaproteomics dataset, one can access multiple functional levels, from exploring broad classes of the metaproteome that give hints to overall functional changes to focused pathway-level analysis within specific taxa (Fig. 2). Nevertheless, proteins can often be assigned to multiple functional groups, which further augments the existing challenge of assigning the identified peptides to proteins sharing similar sequences but originating from different species. A variety of metaproteomics software tools for functional microbiome analysis is available and have been recently reviewed [58] and compared [98]. The performance of these computational tools differed to a large extent when tested on a single dataset, indicating potential difficulties for cross-study comparisons of data acquired by different labs, with different sample processing protocols and MS settings. Finally, bioinformatics tools for taxonomic and functional analysis face a large number of unannotated sequences. This is partly because the quality of annotations of sequence databases originating from metagenomic projects might be low, and for most proteins there is missing biochemical evidence of their function. Resources for proteins functional annotations [e.g, Gene Ontology [99], eggNOG [100], UniProt [101]], biochemical pathways [MetaCyc [102], KEGG [103], neXtProt [104]] and interactions [STRING [105]], are essential for the use of metaproteomics to address biological questions.

In summary, the complexity and heterogeneity of stool samples brings considerable wet lab challenges to the metaproteomics field, but tailored protein databases, combined search algorithms, and iterative workflows, improve protein identification. This was demonstrated in a recent multi-lab comparison of metaproteomics workflows, where the same samples were given to 7 different labs. Different wet lab processing protocols introduced a variability at the peptide level, which, however, largely disappeared at the protein level in downstream bioinformatic analysis [52]. Nonetheless, there are still substantial bioinformatics limitations in metaproteomics related to the identification of false positives and functional annotation of the data. Metaproteomics will benefit from standardised bioinformatics pipelines that reliably process metaproteome data within a short time frame and link protein sequence to the taxonomic and biochemical information available from community resources. Without a doubt, new efficient bioinformatics tools adapted to the complexity of microbiomes are the key for more routine application of metaproteomics.

6. Current challenges of LC–MS-based metabolomics of stool samples

Chemical complexity of feces. In addition to proteins, fecal matter contains other biomolecules that reflect the process of nutrition to which the gut microbiome significantly contributes. Feces contain typically between 60 and 85% of water, depending on the fiber intake, and the dry matter consists of microbial biomass (25–54%), shredded epithelial cells and mucus, undigested food residues, macromolecules (fiber, protein, DNA, mucopolysaccharides) and small molecules or metabolites [106]. The fecal metabolome refers to the collection of these small molecules, *i.e.*, sugars, organic acids, amino acids, nucleotides, phenols, indoles, lipids, and hormones, all of which might have roles as signaling molecules, metabolic intermediates, or secondary metabolites [107]. Thus, the metabolome can be interpreted as a molecular signature of the host under certain physiological conditions and a record of the interactions between the host and the gut microbiome. A recent estimate suggested that gut bacterial products account for up to 90% of the fecal metabolome [107], reflecting the gut microbiota composition and explaining on average ~68% of its variance [108]. Thus, the fecal metabolome is considered a functional readout of the microbiome [108]; however, some of the metabolites will be common for the gut microbiota and the host as feces contain a combined metabolic output of both.

Volatile metabolites. Currently, there are over 115,000 characterised metabolites in the Human Metabolome Database [109], of which 5.9% (6810) originate from feces and are accessible in the Human Fecal Metabolome Database [107]. The annotation of metabolites in the Human metabolome database is based on CFM-ID, a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra [110]. In addition, many of the assignments were performed using a combination of manual annotation and data mining software tools such as PolySearch2 [111].

The composition of the fecal metabolome depends on the diet [112], and although the majority of the metabolites is non-volatile, some of the most abundant metabolites in human feces

are short-chain fatty acids (SCFA) such as acetic, propionic, and butyric acid [107]. SCFA are the best-known representatives of almost 400 volatile organic compounds that have been identified in human fecal samples [113]. SCFA are the end products of bacterial fermentation in the gut and function as energy sources for epithelial cells and as bioactive metabolites regulating the immune system, intestinal barrier function [114], and microbial behavior [115,116]. Despite SCFA's importance, it is difficult to handle volatile organic compounds because of their gaseous form, resulting in their loss when using common sample preparation methods. Still, several targeted metabolomics methods for SCFA detection and quantification have been developed and successfully applied, using chemical derivatization and a highly sensitive MS method of multiple reaction monitoring [117,118].

Methods standardisation. Recent functional investigations of the gut microbiome document growing interest in stool metabolomics [119,120]. However, standardised methods for collecting, processing, and analysing fecal samples are still lacking, and their paucity greatly limits study-to-study comparisons. There is inherent variability in fecal samples even within one individual that depends on feeding status and bowel activity, reflected by dynamic changes in metabolite composition over time [121]. Consequently, multiple-day sampling and pooling have been proposed to minimise day to day variation in metabolite profiles [122]. In addition, the fact that feces contain metabolically active microbial cells makes their analysis sensitive to differences in collection methods, as exposure to aerobic conditions and different temperatures can change the metabolite composition of samples. Ethanol preservation is an alternative when immediate freezing of samples is not possible, as shown for samples stored in 95% ethanol up to 4 days that exhibited a metabolic profile similar to fresh samples [83].

The topographical position from which the fecal sample is taken can also affect the metabolic profile, and therefore it is crucial to homogenise the fresh sample before aliquoting [123]. Alternatively, small molecule extraction can be performed using the entire fecal sample to avoid missing metabolites present in unsampled areas. One of the most critical steps in sample preparation is normalisation to account for feces water content. Fecal samples can have up to 30% variation in water content, which is significant enough to affect downstream statistical analysis and skew especially modest metabolite differences between samples [106]. Finally, the chemical diversity of the metabolome makes metabolite extraction a formidable problem. A particular solvent can extract metabolites of the same chemical class, and no single extraction method is optimal for all metabolites [124]. If metabolite coverage is of utmost importance, multiple extractions should be performed using solvents of various polarity indices. Although this will increase coverage, it can also significantly increase financial and logistical burdens. The points above, plus several other guidelines regarding sample collection and preparation [107,125], must be considered in the experimental design of metabolomics studies.

Quality control. In addition to inherent biological variation, analytical variation of LC–MS instruments can also cause issues if not appropriately addressed. In general, LC–MS data are collected over long periods of time and, although not a recommended practice, are sometimes analysed in multiple batches. Consequently, LC–MS and MS/MS data exhibit significant variability depending on the instrument condition and operating environment. Shifts in m/z values and retention times of molecular features between runs might result in different spectral patterns, negatively impacting metabolite identification or quantification. Therefore, quality control (QC) samples should be applied and used to model and correct systematic measurement bias and between-batch errors [126]. In targeted metabolomics, the QC sample often consists of a mixture of the authentic chemical standards representing the target analytes. The selection of an appropriate QC sample for untargeted assays is more complex. It is generally recommended that the QC sample reflects the aggregate metabolite composition of the biological samples in a given study, and a homogenous pooled QC sample prepared from all biological samples under study should be analyzed before injection of individual samples, after a fixed number of samples have been injected, and also after injection of the last sample [126]. Several software tools based on mathematical models for signal correction [127–129] and simulation of QC sample data [130] have the potential to correct batch-to-batch variations and instrumental drift. Of note, the use of pooled QC samples is also valid for metaproteomics studies.

Metabolite identification. From the popular instrumental platforms used for metabolomics, *i.e.*, nuclear magnetic resonance spectroscopy, LC–MS, and gas chromatography coupled to MS, LC–MS approaches offer higher sensitivity and relatively broad metabolite coverage. However, this sensitivity often results in more laborious identification of analytes [107]. A standard approach for metabolite identification in untargeted, discovery-based analysis, analogous to the one used in (meta)proteomics, is querying metabolomic databases for the molecular mass values of the identified peaks using a tolerance window. Because metabolomics databases lack genetic templates as those used in metaproteomics, the databases will likely be incomplete with missing candidate matches for more rarely occurring compounds. Moreover, compared to peptides, small metabolites often lack common building blocks and are built from both very frequently occurring elements (C, H, O, N, S, and P) and trace elements (e.g., Na, K, Mg, Zn, Fe, Ca, Mo, Cu, Co, and Mn). Even though MS analysis can accurately determine the mass of a compound, this information alone is not sufficient to differentiate isomers, and additional information, including the fragmentation spectrum and retention time, is critical for structural elucidation of a mass measurement [131].

The use of standards for comparing analyte retention times and mass fragmentation patterns assists in accurate biomolecules identification and especially quantification. Once the metabolite has been confidently identified, an additional challenge posed is the determination of its concentration in the sample studied. Given that potential biomarkers of health and disease states will most often be found in both conditions, though at different levels, accurate quantification of promising targets is a desirable feature. Although a regular MS run will provide semi-quantitative information on a metabolite, such as peak area and signal intensity, due to the variation between runs commonly seen in MS experiments, more careful analyses are required to directly compare metabolite concentrations in different samples. This is usually achieved using metabolite standards containing deuterium, a heavy hydrogen isotope. By spiking samples with known concentrations of the deuterated standard and comparing the peaks of the standard and target compounds, one can accurately determine the absolute concentration of the metabolite studied. Still, because of an impracticality for untargeted analyses in which standards are not available for most compounds, general approaches based on prediction models are gaining importance [132].

In conclusion, high mass accuracy of state-of-the-art MS instruments and complementary analysis of molecular patterns are increasingly able to assign putative structures to the detected features despite the inherent challenges that metabolomics faces. But instrument advances can do little if the databases are not in constant improvement. As mentioned above, only 6810 metabolites in the Human Metabolome Database are annotated as being found in feces. This is possibly orders of magnitude below the real chemical diversity of the human gastrointestinal tract. Equally important is the development and improvement of metabolomics software tools, which still need to address many challenges associated with metabolite identification, diverse data types, and large volumes of data [133–135].

7. Gut microbiome establishment: Insights from metaproteomics and metabolomics studies

Stool metaproteomics and metabolomics have been used to study various diseases; yet, here we focus on their use for functional characterisation of the early life gut microbiome (Table 1). Understanding the establishment of the human gut microbiome during infancy is paramount for modern medicine because of its implications for long-term health [136,137]. Numerous reports have demonstrated that mammalian systems are adapted to receive specific microbial signals necessary for optimal physiological development [138,139]. Specifically for humans, an infant gut microbiome characterised by early bacterial colonisers from the genera *Bifidobacterium* and *Bacteroides*, adapted to utilise human milk oligosaccharides, appears to be a cornerstone of healthy development. Perturbations of the microbiome at the earliest time in life during maximal immune, metabolic, and neuroendocrine development predispose infants to non-communicable diseases caused by underlying defects in physiology [9,140] as well as more frequent infections [141–143]. The biochemical processes that govern the microbial dynamics during gut colonisation remain a poorly understood yet exciting research frontier.

Initial colonisation. Integrative analysis of metagenomic data from 34 longitudinal studies worldwide showed that gut microbiome maturation happens in an orchestrated manner, suggesting that the timing of microbial succession is biologically determined [144]. The gut microbiome of infants born at term and vaginally is seeded with vertically transmitted microbes from the mother and is initially dominated by facultative anaerobic bacteria (*i.e.*, *Streptococcus* spp., Enterobacterales, *Staphylococcus* spp.), which are soon replaced by a community dominated by *Bacteroides* and especially *Bifidobacterium* during the lactation period [145,146]. A common belief has been that the initial facultative anaerobes consume oxygen and facilitate the subsequent engraftment of obligate anaerobes. This view was recently questioned by a multi-omic study that provided evidence of anaerobic fermentation of amino acids as a mechanism for the initial growth of *E. coli*, the most common early colonizer [120]. A gnotobiotic animal study showed a similar finding: establishment of the dominant intestinal anaerobe *Bacteroides thetaiotaomicron* was dependent on the *Bacteroides* inoculum size and preestablishment by bacteria capable or not of consuming oxygen [147]. Another example of the versatile metabolic capacities of facultative anaerobes from the order Enterobacterales is their ability to degrade fatty acids and lipids [148], which constitute ~50% of the infant's first stool [119]. Additional translational research using functional omics needs to clarify whether the presence of the very first microbial colonisers is driven by their better survival in the environment [147] and increased capacity to degrade host-derived components such as proteins [149] and lipids [150], or a combination of these factors.

Foundation species. The strongest documented disruptors to the gut microbiome development are birth by C-section, lack of breastfeeding, and antibiotic use in infancy [141,151,152]. Alterations of the microbiota composition by these adverse external factors have also been documented at the metabolome level. For example, early antibiotic exposure in preterm infants functionally altered the gut metabolic output, including pathways related to vitamin biosynthesis, bile acids, amino acid metabolism, and neurotransmitters [153]. Similarly, different early life feeding methods, *i.e.*, breastfeeding, formula-feeding, or their combination, induced distinct fecal metabolite profiles in infants [154]. The above reports illustrate how the metabolic output of the gut microbiome directly depends on its composition [155], and adverse external factors may affect the levels of most metabolites currently detectable, as predicted in an animal study [156].

A signature of C-section born infants is a lack of *Bacteroides* spp., delayed *Bifidobacterium* development, and an expansion of facultative anaerobes adapted to a more oxidative environment and without the genomic capability to metabolize milk carbohydrates [152]. Although several metagenomic studies compared fecal microbiota composition of infants delivered vaginally and by C-section, characterisation of the microbiota functions is largely missing. Nevertheless, there is evidence for the functional benefits of *Bifidobacterium* species, strict anaerobes from the phylum Actinobacteria that are the founder species of the gut microbiome associated with a protective immune system modulation [157]. *Bifidobacterium* persists at high levels during lactation because of their unique genomic capacity to utilise human milk oligosaccharides. In breastfed infants, high bifidobacterial levels lead to a high SCFA concentration and a decrease in the gut pH [158–160], limiting the growth of other bacteria, such as *Enterobacterales* [161]. Supplementation with *Bifidobacterium* strains has also been associated with altered fecal metabolome [162], lower levels of potential pathogens such as *Enterococcus*, *Enterobacter*, and *Klebsiella*, and reduced carriage of antimicrobial genes [163,164].

In addition to host management of the first gut microbiota through breastfeeding, IgA – the most abundant immunoglobulin isotype secreted into the gut, and received by the infant via breastmilk – appears to play a key role in gut microbiota maturation [165]. This was demonstrated in a metaproteomic investigation that assessed gut microbiota maturation in newborn mice [166], using an IgA-deficient ($Rag2^{-/-}$) mouse genetic background. The results confirmed the role of breastfeeding in modulating the mouse gut microbiota in the first days of life, but at the same time suggested other concurrent factors, related to the mother's gut microbiota, immune response, or regulation by the mucosal immune system itself.

Preterm infants. The above-described gut colonisation process is very different in infants born prematurely. Reports describing the gut microbiome composition showed that premature infants display reduced alpha diversity, delayed colonization with obligate anaerobic bacteria, and increased abundance of opportunistic pathogens compared to term infants [167,168]. A recent metaproteomics study followed the fecal microbiome of preterm infants during the first six weeks of life and brought additional information on the physiology of the premature gut [34]. Compared to term infants, gastrointestinal barrier related proteins were less abundant in preterm infants' feces, while bacterial oxidative stress proteins of facultative anaerobes were increased. The authors hypothesised that these findings might suggest the introduction of oxygen into the gut lumen by respiratory support commonly used in neonatal intensive care units. Previously, respiratory support was associated with delayed colonisation by strict anaerobes [138], a hallmark of the preterm gut colonisation process. Subsequently, the aerobic environment might decrease the abundance of strict anaerobes such as *Bifidobacterium* spp., the primary producers of SCFA involved in the production of anti-inflammatory cytokines and stimulation of the intestinal barrier function [37,38].

A study that combined metagenomics and metaproteomics has given an ecological perspective on the premature infant gut colonization process. This genome-resolved metaproteomics study demonstrated that the contributions of individual organisms to microbiome development depend on microbial community context [35]. Furthermore, the microbial metaproteome was more variable over time than the community composition, and genetically similar microbes colonizing different infants were found to have distinct proteomes. These results indicated that microbiome function is not only driven by the type of organisms present but largely depends on microbial responses to the unique set of physiological conditions in the infant's gut. Similarly, the stool

metabolome of preterm infants appears to be distinct between individuals, without any apparent associations to health outcomes, such as necrotizing enterocolitis and sepsis [169].

Studies describing preterm infants gut colonisation have dominated the early life microbiome functional description because of more straightforward sample collection logistics and availability of detailed clinical data. Although these reports are specific for the premature gut microbiome, there might be certain parallels with the general microbial colonisation process in humans, regardless of gestational age. For example, a case study of one preterm infant documented how bacterial activity transits toward more complex metabolic functions over the first month of life [170]. Based on the identified proteins functional classification, the authors predicted that the gut microbial community first focused its resources on biomass growth, protein production, and lipid metabolism. After this initial microbiome establishment during the first two weeks, it switched to more complex metabolic functions, such as carbohydrate metabolism. Several reports on the preterm infant fecal metaproteome also documented low bacterial load in the first weeks after birth, showing a time-dependent increase in the relative abundance of microbial proteins while the abundance of host- and dietary-derived proteins gradually decreased [71,170,171]. A similar trend has also been observed for term infants, although the observation was based on metaproteomes of only three infants [34]. Overall, these studies highlight the strong interdependency between the human host and the gut microbiome for both to reach maturity. Proteins that directly regulate gut colonisation and maturation will serve as valuable markers for intestinal barrier development and immune system education.

Time course metaproteomics had also been applied to specific actions of microbial eukaryotes within the gut microbiome [77]. This case study of low abundance microbial taxa characterised fecal samples from a premature infant with a documented *Candida* blood infection with the aim to describe the behaviour of the fungi in the human gut. Metagenomic sequencing confirmed the presence of *C. parapsilosis* in the infant's fecal sample, with indications of robust establishment and active function within the gut microbiome. Further, protein-derived metabolic activities of bacteria, fungi, and their shared activity showed distinct partitioning of function and cooperation between eukaryotes and prokaryotes within the community during early life. This study highlights the importance of characterising interkingdom interactions within the human microbiome, as these are essential components of the relationship between the microbiome and its host.

8. Conclusions and outlook

Omics based on LC–MS are gradually gaining momentum to identify with high precision functionalities of the gut microbiome. LC–MS analyses of stool assist in unravelling interactions between different microorganisms residing in the gut as well as those with the host, offering insights beyond taxonomic composition and genomic information. MS-based omics provide data that DNA sequences cannot; that is which proteins and metabolites are present and their quantitative information. In addition, identification of post-translational modifications is only possible by metaproteomics.

The combination of LC–MS techniques with DNA sequencing applied on longitudinal human studies has already led to the description of nuanced signatures of healthy and disease states. Still, several methodological and bioinformatics challenges persist, with stool sample chemical complexity, lack of standardised method, and incomplete databases being the main issues contributing to low metaproteome and metabolome coverages (Fig. 2). However, once the current challenges are overcome, it will be possible to fully define the intertwined metabolic networks of individual gut microbes and the human host.

Declaration of Competing Interest

The study funding sources are listed in the Acknowledgements. The authors have no financial/commercial conflicts of interest.

Acknowledgements

This work was funded by The Research Council of Norway Grant No. 274296 (V.K.P). This work was also supported by the Cumming School of Medicine, the Alberta Children Hospital Research Institute, the Snyder Institute of Chronic Diseases, the Canadian Institutes for Health Research, the Sick Kids Foundation, and W. Garfield Weston Foundation (M.C.A). A.D. was supported by an NSERC Discovery Grant (DGECR-2019-00112). L.C.M.A. was supported by the Canadian Institutes for Health Research, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (grants E-26/202.705/2018, E-26/210.209/2018, E-26/010.001280/2016, and E-26/211.554/2019) and Fundação Oswaldo Cruz Inova Fiocruz/VPPCB Program (VPPCB-007-FIO-18-2-51). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. Figures were created with BioRender.com.

Authors contributions

Manuscript draft (VKP), critical text review (VKP, M-CA, AD, LCMA).

References

[1] Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV, Muller MN, et al. Cospeciation of gut microbiota with hominids. Science 2016;353(6297):380–2.
[2] Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med 2018;24(4):392–400.
[3] Manor O, Dai CL, Kornilov SA, Smith B, Price ND, Lovejoy JC, et al. Health and disease markers correlate with gut microbiome composition across thousands of people. Nat Commun 2020;11(1):5206.
[4] Kelsen JR, Wu GD. The gut microbiota, environment and diseases of modern society. Gut microbes 2012;3(4):374–82.
[5] Roubaud-Baudron C, Ruiz VE, Swan AM, Vallance BA, Ozkul C, Pei Z, et al. Long-term effects of early-life antibiotic exposure on resistance to subsequent bacterial infection. mBio 2019;10(6):e02820–e2919.
[6] Korpela K, Salonen A, Virta LJ, Kekkonen RA, Forslund K, Bork P, et al. Intestinal microbiome is related to lifetime antibiotic use in Finnish preschool children. Nat Commun 2016;7(1):10410.
[7] Uzan-Yulzari A, Turta O, Belogolovski A, Ziv O, Kunz C, Perschbacher S, et al. Neonatal antibiotic exposure impairs child growth during the first six years of life by perturbing intestinal microbial colonization. Nat Commun 2021;12(1):443.
[8] Sepich-Poore GD, Zitvogel L, Straussman R, Hasty J, Wargo JA, Knight R. The microbiome and human cancer. Science 2021;371(6536):eabc4552.
[9] Aversa Z, Atkinson EJ, Schafer MJ, Theiler RN, Rocca WA, Blaser MJ, et al. Association of infant antibiotic exposure with childhood health outcomes. Mayo Clin Proc 2020.
[10] Cryan JF, O'Riordan KJ, Sandhu K, Peterson V, Dinan TG. The gut microbiome in neurological disorders. Lancet Neurol 2020;19(2):179–94.
[11] Shaw KA, Bertha M, Hofmekler T, Chopra P, Vatanen T, Srivatsa A, et al. Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. Genome Med 2016;8(1):75.
[12] Langdon A, Crook N, Dantas G. The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. Genome Med 2016;8(1):39-..
[13] Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature 2018;555(7698):623–8.
[14] Bizzarro S, Laine ML, Buijs MJ, Brandt BW, Crielaard W, Loos BG, et al. Microbial profiles at baseline and not the use of antibiotics determine the clinical outcome of the treatment of chronic periodontitis. Sci Rep. 2016;6:20205-.

[15] Grady NG, Petrof EO, Claud EC. Microbial therapeutic interventions. Semin Fetal Neonatal Med 2016;21(6):418–23.
[16] Haange S-B, Oberbach A, Schlichting N, Hugenholtz F, Smidt H, von Bergen M, et al. Metaproteome analysis and molecular genetics of rat intestinal microbiota reveals section and localization resolved species distribution and enzymatic functionalities. J Proteome Res 2012;11(11):5406–17.
[17] Berg G, Rybakova D, Fischer D, Cernava T, Vergès M-C-C, Charles T, et al. Microbiome definition re-visited: old concepts and new challenges. Microbiome 2020;8(1):103.
[18] Byndloss MX, Bäumler AJ. The germ-organ theory of non-communicable diseases. Nat Rev Microbiol 2018;16(2):103–10.
[19] Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol 2020;38 (6):685–8.
[20] Durazzi F, Sala C, Castellani G, Manfreda G, Remondini D, De Cesare A. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. Sci Rep 2021;11(1): 3030.
[21] Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. Cell Host Microbe. 2018;24(1):133-45.e5.
[22] Gasparrini AJ, Wang B, Sun X, Kennedy EA, Hernandez-Leyva A, Ndao IM, et al. Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. Nat Microbiol 2019;4 (12):2285–97.
[23] Pereira-Marques J, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, van Doorn L-J, et al. Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. Front Microbiol 2019;10(1277).
[24] Langille MGI. Exploring linkages between taxonomic and functional profiles of the human microbiome. mSystems 2018;3(2):e00163–e217.
[25] Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature 2012;486(7402):207–14.
[26] Heintz-Buschart A, Wilmes P. Human gut microbiome: function matters. Trends Microbiol 2018;26(7):563–74.
[27] Li L, Ning Z, Zhang X, Mayne J, Cheng K, Stintzi A, et al. RapidAIM: a culture- and metaproteomics-based Rapid Assay of Individual Microbiome responses to drugs. Microbiome 2020;8(1):33.
[28] Blakeley-Ruiz JA, Erickson AR, Cantarel BL, Xiong W, Adams R, Jansson JK, et al. Metaproteomics reveals persistent and phylum-redundant metabolic functional stability in adult human gut microbiomes of Crohn's remission patients despite temporal variations in microbial taxa, genomes, and proteomes. Microbiome 2019;7(1):18.
[29] Inkpen SA, Douglas GM, Brunet TDP, Leuschen K, Doolittle WF, Langille MGI. The coupling of taxonomy and function in microbiomes. Biol Philos 2017;32 (6):1225–43.
[30] Vargas-Blanco DA, Shell SS. Regulation of mRNA stability during bacterial stress responses. Front Microbiol 2020;11(2111).
[31] Wilmes P, Bond PL. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environ Microbiol 2004;6(9):911–20.
[32] Lichtman JS, Marcobal A, Sonnenburg JL, Elias JE. Host-centric proteomics of stool: A novel strategy focused on intestinal responses to the gut microbiota *. Mol Cell Proteomics 2013;12(11):3310–8.
[33] Jin P, Wang K, Huang C, Nice EC. Mining the fecal proteome: from biomarkers to personalised medicine. Expert Rev Proteomics 2017;14(5):445–59.
[34] Henderickx JGE, Zwittink RD, Renes IB, van Lingen RA, van Zoeren-Grobben D, Jebbink LJG, et al. Maturation of the preterm gastrointestinal tract can be defined by host and microbial markers for digestion and barrier defense. Sci Rep 2021;11(1):12808.
[35] Brown Christopher T, Xiong W, Olm Matthew R, Thomas Brian C, Baker R, Firek B, et al. Hospitalized premature infants are colonized by related bacterial strains with distinct proteomic profiles. mBio.9(2):e00441-18.
[36] Levan SR, Stamnes KA, Lin DL, Panzer AR, Fukui E, McCauley K, et al. Elevated faecal 12,13-diHOME concentration in neonates at high risk for asthma is produced by gut bacteria and impedes immune tolerance. Nat Microbiol 2019;4(11):1851–61.
[37] Ehrlich AM, Pacheco AR, Henrick BM, Taft D, Xu G, Huda MN, et al. Indole-3-lactic acid associated with Bifidobacterium-dominated microbiota significantly decreases inflammation in intestinal epithelial cells. BMC Microbiol 2020;20(1):357.
[38] Henrick BM, Rodriguez L, Lakshmikanth T, Pou C, Henckel E, Arzoomand A, et al. Bifidobacteria-mediated immune system imprinting early in life. Cell. 2021;184(15):3884-98.e11.
[39] Griffiths WJ, Wang Y. Mass spectrometry: from proteomics to metabolomics and lipidomics. Chem Soc Rev 2009;38(7):1882–96.
[40] Aretz I, Meierhofer D. Advantages and pitfalls of mass spectrometry based metabolome profiling in systems biology. Int J Mol Sci 2016;17(5):632.
[41] Guo J, Huan T. Comparison of full-scan, data-dependent, and data-independent acquisition modes in liquid chromatography-mass spectrometry based untargeted metabolomics. Anal Chem 2020;92 (12):8072–80.
[42] Defossez E, Bourquin J, von Reuss S, Rasmann S, Glauser G. Eight key rules for successful data-dependent acquisition in mass spectrometry-based metabolomics. Mass Spectrom Rev 2021.
[43] Beata R, Attila K-F, Sandor P, Michael PM. Data preprocessing and filtering in mass spectrometry based proteomics. Curr Bioinform 2012;7(2):212–20.
[44] Smith R, Mathis AD, Ventura D, Prince JT. Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. BMC Bioinformatics. 2014;15 Suppl 7(Suppl 7):S9-S.
[45] Tsai T-H, Wang M, Ressom HW. Preprocessing and analysis of LC-MS-based proteomic data. Methods Mol Biol (Clifton, NJ) 2016;1362:63–76.
[46] Riquelme G, Zabalegui N, Marchi P, Jones CM, Monge ME. A python-based pipeline for preprocessing LC-MS data for untargeted metabolomics workflows. Metabolites 2020;10(10):416.
[47] Deng Y, Ren Z, Pan Q, Qi D, Wen B, Ren Y, et al. pClean: an algorithm to preprocess high-resolution tandem mass spectra for database searching. J Proteome Res 2019;18(9):3235–44.
[48] 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.
[49] Chen C, Hou J, Tanner JJ, Cheng J. Bioinformatics methods for mass spectrometry-based proteomics data analysis. Int J Mol Sci 2020;21(8):2873.
[50] Puckett SP, Samples RM, Schloss PD, Balunas MJ. 7.25 - metabolomics and the microbiome: characterizing molecular diversity in complex microbial communities. In: Liu H–W, Begley TP, editors. Comprehensive natural products III. Oxford: Elsevier; 2020. p. 502–18.
[51] O'Shea K, Misra BB. Software tools, databases and resources in metabolomics: updates from 2018 to 2019. Metabolomics 2020;16(3):36.
[52] Van Den Bossche T, Kunath BJ, Schallert K, Schäpe SS, Abraham PE, Armengaud J, et al. Critical assessment of metaproteome investigation (CAMPI): A multi-lab comparison of established workflows. bioRxiv. 2021:2021.03.05.433915.
[53] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010;7(5):335–6.
[54] Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, et al. Community-led, integrated, reproducible multi-omics with anvi'o. Nat Microbiol 2021;6(1):3–6.
[55] Zhang X, Figeys D. Perspective and guidelines for metaproteomics in microbiome studies. J Proteome Res 2019;18(6):2370–80.
[56] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3(1):160018.
[57] Peters DL, Wang W, Zhang X, Ning Z, Mayne J, Figeys D. Metaproteomic and metabolomic approaches for characterizing the gut microbiome. Proteomics 2019;19(16):1800363.
[58] Salvato F, Hettich RL, Kleiner M. Five key aspects of metaproteomics as a tool to understand functional interactions in host-associated microbiomes. PLoS Pathog. 2021;17(2):e1009245-e.
[59] Choo JM, Leong LEX, Rogers GB. Sample storage conditions significantly influence faecal microbiome profiles. Sci Rep 2015;5(1):16350.
[60] Marotz C, Cavagnero KJ, Song SJ, McDonald D, Wandro S, Humphrey G, et al. Evaluation of the effect of storage methods on fecal, saliva, and skin microbiome composition. mSystems 2021;6(2):e01329–e1420.
[61] Morris LS, Marchesi JR. Assessing the impact of long term frozen storage of faecal samples on protein concentration and protease activity. J Microbiol Methods 2016;123:31–8.
[62] Jensen M, Wippler J, Kleiner M. Evaluation of RNAlater™ as a field-compatible preservation method for metaproteomic analyses of bacteria-animal symbioses. bioRxiv. 2021:2021.06.16.448770.
[63] Saito MA, Bulygin VV, Moran DM, Taylor C, Scholin C. Examination of microbial proteome preservation techniques applicable to autonomous environmental sample collection. Front Microbiol 2011;2:215.
[64] Xiong W, Giannone RJ, Morowitz MJ, Banfield JF, Hettich RL. Development of an enhanced metaproteomic approach for deepening the microbiome characterization of the human infant gut. J Proteome Res 2015;14(1):133–41.
[65] Tanca A, Palomba A, Pisanu S, Deligios M, Fraumene C, Manghina V, et al. A straightforward and efficient analytical pipeline for metaproteome characterization. Microbiome 2014;2(1):49.
[66] Tanca A, Palomba A, Pisanu S, Addis MF, Uzzau S. Enrichment or depletion? The impact of stool pretreatment on metaproteomic characterization of the human gut microbiota. Proteomics 2015;15(20):3474–85.
[67] Zhang X, Li L, Mayne J, Ning Z, Stintzi A, Figeys D. Assessing the impact of protein extraction methods for human gut metaproteomics. J Proteomics 2018;180:120–7.
[68] Rechenberger J, Samaras P, Jarzab A, Behr J, Frejno M, Djukovic A, et al. Challenges in clinical metaproteomics highlighted by the analysis of acute leukemia patients with gut colonization by multidrug-resistant enterobacteriaceae. Proteomes 2019;7(1):2.
[69] Muth T, Renard BY, Martens L. Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. Expert Rev Proteomics 2016;13(8):757–69.
[70] Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 2017;550(7674):61–6.
[71] Zwittink RD, van Zoeren-Grobben D, Martin R, van Lingen RA, Groot Jebbink LJ, Boeren S, et al. Metaproteomics reveals functional differences in intestinal microbiota development of preterm infants. Mol Cell Proteomics 2017;16 (9):1610–20.
[72] Zhang X, Deeke SA, Ning Z, Starr AE, Butcher J, Li J, et al. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle

proteins in pediatric inflammatory bowel disease. Nat Commun 2018;9 (1):2873.
[73] Maier TV, Lucio M, Lee LH, VerBerkmoes NC, Brislawn CJ, Bernhardt J, et al. Impact of dietary resistant starch on the human gut microbiome, metaproteome, and metabolome. mBio 2017;8(5):e01343–e1417.
[74] Tierney BT, Yang Z, Luber JM, Beaudin M, Wibowo MC, Baek C, et al. The landscape of genetic content in the gut and oral human microbiome. Cell Host Microbe. 2019;26(2):283-95.e8.
[75] Mayers MD, Moon C, Stupp GS, Su AI, Wolan DW. Quantitative metaproteomics and activity-based probe enrichment reveals significant alterations in protein expression from a mouse model of inflammatory bowel disease. J Proteome Res 2017;16(2):1014–26.
[76] Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature 2012;486:222.
[77] West PT, Peters SL, Olm MR, Yu FB, Gause H, Lou YC, et al. Genetic and behavioral adaptation of Candida parapsilosis to the microbiome of hospitalized infants revealed by in situ genomics, transcriptomics, and proteomics. Microbiome 2021;9(1):142.
[78] Pettersen VK, Dufour A, Arrieta M-C. Metaproteomic profiling of fungal gut colonization in gnotobiotic mice. bioRxiv. 2020:2020.12.24.424341.
[79] Cheng K, Ning Z, Zhang X, Li L, Liao B, Mayne J, et al. MetaLab: an automated pipeline for metaproteomic data analysis. Microbiome 2017;5(1):157.
[80] Muth T, Behne A, Heyer R, Kohrs F, Benndorf D, Hoffmann M, et al. The MetaProteomeAnalyzer: A powerful open-source software suite for metaproteomics data analysis and interpretation. J Proteome Res 2015;14 (3):1557–65.
[81] Cantarel BL, Erickson AR, VerBerkmoes NC, Erickson BK, Carey PA, Pan C, et al. Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. PLoS One. 2011;6(11):e27173-e.
[82] Jagtap PD, Blakely A, Murray K, Stewart S, Kooren J, Johnson JE, et al. Metaproteomic analysis using the Galaxy framework. Proteomics 2015;15 (20):3553–65.
[83] Park SKR, Jung T, Thuy-Boun PS, Wang AY, Yates JR, Wolan DW. ComPIL 2.0: an updated comprehensive metaproteomics database. J Proteome Res 2019;18(2):616–22.
[84] Bassignani A, Plancade S, Berland M, Blein-Nicolas M, Guillot A, Chevret D, et al. Benefits of iterative searches of large databases to interpret large human gut metaproteomic data sets. J Proteome Res 2021;20(3): 1522–34.
[85] Beyter D, Lin MS, Yu Y, Pieper R, Bafna V. ProteoStorm: an ultrafast metaproteomics database search framework. Cell Systems. 2018;7(4):463-7.e6.
[86] Gonnelli G, Stock M, Verwaeren J, Maddelein D, De Baets B, Martens L, et al. A decoy-free approach to the identification of peptides. J Proteome Res 2015;14 (4):1792–8.
[87] Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 2007;4(11):923–5.
[88] Kleiner M, Thorson E, Sharp CE, Dong X, Liu D, Li C, et al. Assessing species biomass contributions in microbial communities via metaproteomics. Nat Commun 2017;8(1):1558.
[89] Kunath BJ, Minniti G, Skaugen M, Hagen LH, Vaaje-Kolstad G, Eijsink VGH, et al. Metaproteomics: sample preparation and methodological considerations. Adv Exp Med Biol 2019;1073:187–215.
[90] 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(12):5773–80.
[91] Mooradian AD, van der Post S, Naegle KM, Held JM. ProteoClade: A taxonomic toolkit for multi-species and metaproteomic analysis. PLoS Comput Biol 2020;16(3):e1007741.
[92] Zhang X, Ning Z, Mayne J, Yang Y, Deeke SA, Walker K, et al. Widespread protein lysine acetylation in gut microbiome and its alterations in patients with Crohn's disease. Nat Commun 2020;11(1):4120.
[93] Motta J-P, Denadai-Souza A, Sagnat D, Guiraud L, Edir A, Bonnart C, et al. Active thrombin produced by the intestinal epithelium controls mucosal biofilms. Nat Commun 2019;10(1):3224.
[94] Vergnolle N. Protease inhibition as new therapeutic strategy for GI diseases. Gut 2016;65(7):1215–24.
[95] Mainoli B, Hirota S, Edgington-Mitchell LE, Lu C, Dufour A. Proteomics and imaging in Crohn's disease: TAILS of unlikely allies. Trends Pharmacol Sci 2020;41(2):74–84.
[96] Gordon MH, Anowai A, Young D, Das N, Campden RI, Sekhon H, et al. N-terminomics/TAILS profiling of proteases and their substrates in ulcerative colitis. ACS Chem Biol 2019;14(11):2471–83.
[97] Fortelny N, Yang S, Pavlidis P, Lange PF, Overall CM. Proteome TopFIND 3.0 with TopFINDer and PathFINDer: database and analysis tools for the association of protein termini to pre- and post-translational events. Nucleic Acids Res. 2015;43(Database issue):D290-D7.
[98] Sajulga R, Easterly C, Riffle M, Mesuere B, Muth T, Mehta S, et al. Survey of metaproteomics software tools for functional microbiome analysis. PLoS ONE 2020;15(11):e0241503.
[99] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 2018;47(D1):D330–8.
[100] Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 2015;44(D1):D286–93.
[101] Consortium TU. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2020;49(D1):D480–9.
[102] Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 2007;36 (suppl_1):D623–31.
[103] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2015;44(D1):D457–62.
[104] Zahn-Zabal M, Michel P-A, Gateau A, Nikitin F, Schaeffer M, Audot E, et al. The neXtProt knowledgebase in 2020: data, tools and usability improvements. Nucleic Acids Res 2020;48(D1):D328–34.
[105] Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res 2016;45(D1):D362–8.
[106] Rose C, Parker A, Jefferson B, Cartmell E. The characterization of feces and urine: A review of the literature to inform advanced treatment technology. Crit Rev Environ Sci Technol 2015;45(17):1827–79.
[107] Karu N, Deng L, Slae M, Guo AC, Sajed T, Huynh H, et al. A review on human fecal metabolomics: Methods, applications and the human fecal metabolome database. Anal Chim Acta 2018;1030:1–24.
[108] Zierer J, Jackson MA, Kastenmüller G, Mangino M, Long T, Telenti A, et al. The fecal metabolome as a functional readout of the gut microbiome. Nat Genet 2018;50(6):790–5.
[109] Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 2018;46(D1):D608–17.
[110] Allen F, Pon A, Wilson M, Greiner R, Wishart D. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. Nucleic Acids Res. 2014;42(Web Server issue):W94-W9.
[111] Liu Y, Liang Y, Wishart D. PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic Acids Res 2015;43(W1):W535–42.
[112] Tang Z-Z, Chen G, Hong Q, Huang S, Smith HM, Shah RD, et al. Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites. Front Genet 2019;10(454).
[113] de Lacy CB, Amann A, Al-Kateb H, Flynn C, Filipiak W, Khalid T, et al. A review of the volatiles from the healthy human body. J Breath Res 2014;8(1):014001.
[114] Blaak EE, Canfora EE, Theis S, Frost G, Groen AK, Mithieux G, et al. Short chain fatty acids in human gut and metabolic health. Benef Microbes 2020;11 (5):411–55.
[115] Tobe T, Nakanishi N, Sugimoto N. Activation of motility by sensing short-chain fatty acids via two steps in a flagellar gene regulatory cascade in enterohemorrhagic Escherichia coli. Infect Immun 2011;79(3):1016–24.
[116] Hung C-C, Garner CD, Slauch JM, Dwyer ZW, Lawhon SD, Frye JG, et al. The intestinal fatty acid propionate inhibits Salmonella invasion through the post-translational control of HilD. Mol Microbiol 2013;87(5):1045–60.
[117] Han J, Lin K, Sequeira C, Borchers CH. An isotope-labeled chemical derivatization method for the quantitation of short-chain fatty acids in human feces by liquid chromatography–tandem mass spectrometry. Anal Chim Acta 2015;854:86–94.
[118] Song HE, Lee HY, Kim SJ, Back SH, Yoo HJ. A facile profiling method of short chain fatty acids using liquid chromatography-mass spectrometry. Metabolites 2019;9(9):173.
[119] Petersen C, Dai DLY, Boutin RCT, Sbihi H, Sears MR, Moraes TJ, et al. A rich meconium metabolome in human infants is associated with early-life gut microbiota composition and reduced allergic sensitization. Cell Reports Med 2021;2(5).
[120] Bittinger K, Zhao C, Li Y, Ford E, Friedman ES, Ni J, et al. Bacterial colonization reprograms the neonatal gut metabolome. Nat Microbiol 2020;5(6):838–47.
[121] Ramamoorthy S, Levy S, Mohamed M, Abdelghani A, Evans AM, Miller LAD, et al. An ambient-temperature storage and stabilization device performs comparably to flash-frozen collection for stool metabolomics in infants. BMC Microbiol 2021;21(1):59.
[122] Lamichhane S, Sundekilde UK, Blædel T, Dalsgaard TK, Larsen LH, Dragsted LO, et al. Optimizing sampling strategies for NMR-based metabolomics of human feces: pooled vs. unpooled analyses. Anal Methods 2017;9 (30):4476–80.
[123] Gratton J, Phetcharaburanin J, Mullish BH, Williams HRT, Thursz M, Nicholson JK, et al. Optimized sample handling strategy for metabolic profiling of human feces. Anal Chem 2016;88(9):4661–8.
[124] Zhou B, Xiao JF, Tuli L, Ressom HW. LC-MS-based metabolomics. Mol Biosyst 2012;8(2):470–81.
[125] Smith L, Villaret-Cazadamont J, Claus SP, Canlet C, Guillou H, Cabaton NJ, et al. Important considerations for sample collection in metabolomics studies with a special focus on applications to liver functions. Metabolites 2020;10 (3):104.
[126] Broadhurst D, Goodacre R, Reinke SN, Kuligowski J, Wilson ID, Lewis MR, et al. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. Metabolomics 2018;14(6):72.
[127] Kuligowski J, Sánchez-Illana Á, Sanjuán-Herráez D, Vento M, Quintás G. Intra-batch effect correction in liquid chromatography-mass spectrometry using

quality control samples and support vector regression (QC-SVRC). Analyst 2015;140(22):7810–7.
[128] Wehrens R, Hageman JA, van Eeuwijk F, Kooke R, Flood PJ, Wijnker E, et al. Improved batch correction in untargeted MS-based metabolomics. Metabolomics 2016;12(88).
[129] Rusilowicz M, Dickinson M, Charlton A, O'Keefe S, Wilson J. A batch correction method for liquid chromatography–mass spectrometry data that does not depend on quality control samples. Metabolomics 2016;12(3):56.
[130] Wang S, Yang H. pseudoQC: A regression-based simulation software for correction and normalization of complex metabolomics and proteomics datasets. Proteomics 2019;19(19):1900264.
[131] Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted metabolomics strategies-challenges and emerging directions. J Am Soc Mass Spectrom 2016;27(12):1897–905.
[132] Liigand J, Wang T, Kellogg J, Smedsgaard J, Cech N, Kruve A. Quantification for non-targeted LC/MS screening without standard substances. Sci Rep 2020;10(1):5808.
[133] Allen F, Pon A, Wilson M, Greiner R, Wishart D. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. Nucleic Acids Res 2014;42(W1):W94–9.
[134] Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminform 2016;8(1):3.
[135] Djoumbou-Feunang Y, Fiamoncini J, Gil-de-la-Fuente A, Greiner R, Manach C, Wishart DS. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. J Cheminform 2019;11(1):2.
[136] Laforest-Lapointe I, Arrieta M-C. Patterns of early-life gut microbial colonization during human immune development: an ecological perspective. Front Immunol 2017;8:788.
[137] Sonnenburg JL, Sonnenburg ED. Vulnerability of the industrialized microbiota. Science 2019;366(6464):eaaw9255.
[138] Gensollen T, Iyer SS, Kasper DL, Blumberg RS. How colonization by microbiota in early life shapes the immune system. Science (New York, NY) 2016;352(6285):539–44.
[139] Olin A, Henckel E, Chen Y, Lakshmikanth T, Pou C, Mikes J, et al. Stereotypic immune system development in newborn children. Cell. 2018;174(5):1277-92.e14.
[140] Chavarro JE, Martín-Calvo N, Yuan C, Arvizu M, Rich-Edwards JW, Michels KB, et al. Association of birth by cesarean delivery with obesity and type 2 diabetes among adult women. JAMA Network Open. 2020;3(4):e202605-e.
[141] Shao Y, Forster SC, Tsaliki E, Vervier K, Strang A, Simpson N, et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature 2019.
[142] Miller JE, Goldacre R, Moore HC, Zeltzer J, Knight M, Morris C, et al. Mode of birth and risk of infection-related hospitalisation in childhood: A population cohort study of 7.17 million births from 4 high-income countries. PLoS Med. 2020;17(11):e1003429-e.
[143] Reyman M, van Houten MA, van Baarle D, Bosch AATM, Man WH, Chu MLJN, et al. Impact of delivery mode-associated gut microbiota dynamics on health in the first year of life. Nat Commun 2019;10(1):4997.
[144] Korpela K, de Vos WM. Early life colonization of the human gut: microbes matter everywhere. Curr Opin Microbiol 2018;44:70–8.
[145] Yassour M, Vatanen T, Siljander H, Hamalainen AM, Harkonen T, Ryhanen SJ, et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci Transl Med 2016;8(343):343ra81.
[146] Mitchell CM, Mazzoni C, Hogstrom L, Bryant A, Bergerat A, Cher A, et al. Delivery mode affects stability of early infant gut microbiota. Cell Rep Med 2020;1(9):100156.
[147] Halpern D, Morvan C, Derré-Bobillot A, Meylheuc T, Guillemet M, Rabot S, et al. Do primocolonizing bacteria enable bacteroides thetaiotaomicron intestinal colonization independently of the capacity to consume oxygen? mSphere 2021;6(3):e00232–e319.
[148] Iram SH, Cronan JE. The beta-oxidation systems of *Escherichia coli* and *Salmonella enterica* are not functionally equivalent. J Bacteriol 2006;188(2):599–608.
[149] Diether NE, Willing BP. Microbial fermentation of dietary protein: an important factor in diet-microbe-host interaction. Microorganisms 2019;7(1):19.
[150] Agans R, Gordon A, Kramer DL, Perez-Burillo S, Rufián-Henares JA, Paliy O, et al. Dietary fatty acids sustain the growth of the human gut microbiota. Appl Environ Microbiol 2018;84(21):e01525–e1618.
[151] Korpela K, Salonen A, Saxen H, Nikkonen A, Peltola V, Jaakkola T, et al. Antibiotics in early life associate with specific gut microbiota signatures in a prospective longitudinal infant cohort. Pediatr Res 2020;88(3):438–43.
[152] Shao Y, Forster SC, Tsaliki E, Vervier K, Strang A, Simpson N, et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature 2019;574(7776):117–21.
[153] Patton L, Li N, Garrett TJ, Ruoss JL, Russell JT, de la Cruz D, et al. Antibiotics effects on the fecal metabolome in preterm infants. Metabolites 2020;10(8):331.
[154] Li N, Yan F, Wang N, Song Y, Yue Y, Guan J, et al. Distinct gut microbiota and metabolite profiles induced by different feeding methods in healthy Chinese infants. Front Microbiol 2020;11(714).
[155] Manor O, Levy R, Borenstein E. Mapping the inner workings of the microbiome: genomic- and metagenomic-based study of metabolism and metabolic interactions in the human microbiome. Cell Metab 2014;20(5):742–52.
[156] Antunes LCM, Han J, Ferreira RBR, Lolić P, Borchers CH, Finlay BB. Effect of antibiotic treatment on the intestinal metabolome. Antimicrob Agents Chemother 2011;55(4):1494–503.
[157] Rabe H, Lundell A-C, Sjöberg F, Ljung A, Strömbeck A, Gio-Batta M, et al. Neonatal gut colonization by Bifidobacterium is associated with higher childhood cytokine responses. Gut Microbes 2020;12(1):1–14.
[158] Henrick BM, Hutton AA, Palumbo MC, Casaburi G, Mitchell RD, Underwood MA, et al. Elevated fecal pH indicates a profound change in the breastfed infant gut microbiome due to reduction of *Bifidobacterium* over the past century. mSphere 2018;3(2):e00041–e118.
[159] Sorbara MT, Dubin K, Littmann ER, Moody TU, Fontana E, Seok R, et al. Inhibiting antibiotic-resistant Enterobacteriaceae by microbiota-mediated intracellular acidification. J Exp Med 2019;216(1):84–98.
[160] Lay C, Chu CW, Purbojati RW, Acerbi E, Drautz-Moses DI, de Sessions PF, et al. A synbiotic intervention modulates meta-omics signatures of gut redox potential and acidity in elective caesarean born infants. BMC Microbiol 2021;21(1):191.
[161] Fukuda S, Toh H, Hase K, Oshima K, Nakanishi Y, Yoshimura K, et al. Bifidobacteria can protect from enteropathogenic infection through production of acetate. Nature 2011;469(7331):543–7.
[162] Alcon-Giner C, Dalby MJ, Caim S, Ketskemety J, Shaw A, Sim K, et al. Microbiota supplementation with Bifidobacterium and Lactobacillus modifies the preterm infant gut microbiota and metabolome. Cell Reports Med 2020;1(5):100077.
[163] Casaburi G, Duar RM, Vance DP, Mitchell R, Contreras L, Frese SA, et al. Early-life gut microbiome modulation reduces the abundance of antibiotic-resistant bacteria. Antimicrob Resistance Infect Control 2019;8(1):131.
[164] Esaiassen E, Hjerde E, Cavanagh JP, Pedersen T, Andresen JH, Rettedal SI, et al. Effects of probiotic supplementation on the gut microbiota and antibiotic resistome development in preterm infants. Front Pediatr 2018;6(347).
[165] Guo J, Ren C, Han X, Huang W, You Y, Zhan J. Role of IgA in the early-life establishment of the gut microbiota and immunity: Implications for constructing a healthy start. Gut Microbes 2021;13(1):1908101.
[166] Levi Mortera S, Soggiu A, Vernocchi P, Del Chierico F, Piras C, Carsetti R, et al. Metaproteomic investigation to assess gut microbiota shaping in newborn mice: A combined taxonomic, functional and quantitative approach. J Proteomics 2019;203:103378.
[167] Stewart CJ, Embleton ND, Marrs EC, Smith DP, Nelson A, Abdulkadir B, et al. Temporal bacterial and metabolic development of the preterm gut reveals specific signatures in health and disease. Microbiome 2016;4(1):67.
[168] Korpela K, Blakstad EW, Moltu SJ, Strommen K, Nakstad B, Ronnestad AE, et al. Intestinal microbiota development and gestational age in preterm neonates. Sci Rep 2018;8(1):2453.
[169] Wandro S, Osborne S, Enriquez C, Bixby C, Arrieta A, Whiteson K. The microbiome and metabolome of preterm infant stool are personalized and not driven by health outcomes, including necrotizing enterocolitis and late-onset sepsis. mSphere 2018;3(3):e00104–e118.
[170] Young JC, Pan C, Adams RM, Brooks B, Banfield JF, Morowitz MJ, et al. Metaproteomics reveals functional shifts in microbial and human proteins during a preterm infant gut colonization case. Proteomics 2015;15(20):3463–73.
[171] Xiong W, Brown CT, Morowitz MJ, Banfield JF, Hettich RL. Genome-resolved metaproteomic characterization of preterm infant gut microbiota development reveals species-specific metabolic shifts and variabilities during early life. Microbiome 2017;5(1):72.
[172] Cortes L, Wopereis H, Tartiere A, Piquenot J, Gouw JW, Tims S, et al. Metaproteomic and 16S rRNA gene sequencing analysis of the infant fecal microbiome. Int J Mol Sci 2019;20(6):1430.