TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

DATA-INDEPENDENT ACQUISITION MASS SPECTROMETRY FOR HUMAN GUT MICROBIOTA METAPROTEOME ANALYSIS

Sami Pietilä

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA - SER. D OSA - TOM. 1603 | MEDICA – ODONTOLOGICA | TURKU 2021

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

DATA-INDEPENDENT ACQUISITION MASS SPECTROMETRY FOR HUMAN GUT MICROBIOTA METAPROTEOME ANALYSIS

Sami Pietilä

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. D OSA – TOM. 1603 | MEDICA – ODONTOLOGICA | TURKU 2021

University of Turku

Faculty of Medicine
Institute of Biomedicine
Medical Microbiology and Immunology
Turku Doctoral Programme of Molecular Medicine

Supervised by

Professor Laura L. Elo
Turku Bioscience Centre,
University of Turku and Åbo Akademi University,
Turku, Finland
Institute of Biomedicine,
University of Turku,
Turku, Finland

Adjunct Professor Arno Hänninen
Institute of Biomedicine,
University of Turku,
Turku, Finland
Turku University Hospital,
Turku, Finland

Reviewed by

Adjunct Professor Reetta Satokari
Faculty of Medicine,
University of Helsinki,
Helsinki, Finland

Assistant Professor Jarkko Salojärvi
Nanyang Technological University,
Singapore

Opponent

Professor Lennart Martens
Faculty of Medicine and Health Sciences,
Ghent University,
Ghent, Belgium

ISBN 978-951-29-8740-5 (PDF)
ISSN 0355-9483 (Print)
ISSN 2343-3213 (Online)
PunaMusta Oy, Turku, Finland 2021

To all the people who have been working on this project.

UNIVERSITY OF TURKU
Faculty of Medicine
Institute of Biomedicine
Medical Microbiology and Immunology
SAMI PIETILÄ: Data-independent acquisition mass spectrometry for human gut microbiota metaproteome analysis
Doctoral Dissertation, 65 pp.
Turku Doctoral Programme of Molecular Medicine
December 2021

ABSTRACT

Human digestive tract microbiota is a diverse community of microorganisms having complex interactions between microbes and the human host. Observing the functions carried out by microbes is essential for gaining understanding on the role of gut microbiota in human health and associations to diseases. New methods and tools are needed for acquirement of functional information from complex microbial samples. Metagenomic approaches focus on taxonomy or gene based function potential but lack power in the discovery of the actual functions carried out by the microbes. Metaproteomic methods are required to uncover the functions. The current high-throughput metaproteomics methods are based on mass spectrometry which is capable of identifying and quantifying ionized protein fragments, called peptides.
Proteins can be inferred from the peptides and the functions associated with protein expression can be determined by using protein databases. Currently the most widely used data-dependent acquisition (DDA) method records only the most intensive ions in a semi-stochastic manner, which reduces reproducibility and produces incomplete records impairing quantification. Alternative data-independent acquisition (DIA) systematically records all ions and has been proposed as a replacement for DDA. However, recording all ions produces highly convoluted spectra from multiple peptides and, for this reason, it has not been known if and how DIA can be applied to metaproteomics where the number of different peptides is high. This thesis work introduced the DIA method for metaproteomic data analysis. The method was shown to achieve high reproducibility enabling the usage of only a single analysis per sample while DDA requires multiple. An easy to use open source software package, DIAtools, was developed for the analysis. Finally, the DIA analysis method was applied to study human gut microbiota and carbohydrate-active enzymes expressed in members of gut microbiota.

KEYWORDS: DIA, enzyme, gut, metaproteomics, microbiota, peptide

TURUN YLIOPISTO
Lääketieteellinen tiedekunta
Biolääketieteen laitos
Lääketieteellinen mikrobiologia ja immunologia
SAMI PIETILÄ: Ihmisen suolistomikrobiston analyysi DIA-massaspektrometriamenetelmällä
Väitöskirja, 65 s.
Molekyylilääketieteen tohtoriohjelma
Joulukuu 2021

TIIVISTELMÄ

Ihmisen suoliston mikrobisto on monien mikro-organismien yhteisö, joka on vuorovaikutuksessa ihmisen kehon kanssa. Suoliston mikrobien toiminnan ymmärtäminen on keskeistä niiden roolista ihmisen terveyteen ja sairauksiin. Uusia tutkimusmenetelmiä tarvitaan mikrobien toiminnallisuuden määrittämiseen monimutkaisista, useita mikrobeja sisältävistä, näytteistä. Yleisesti käytetyt metagenomiikan menetelmät keskittyvät taksonomiaan tai geenien perusteella ennustettuihin funktioihin, mutta metaproteomiikkaa tarvitaan mikrobien toiminnan selvittämiseen. Metaproteomiikka-analyysiin voidaan käyttää massaspektrometriaa, jolla pystytään tunnistamaan ja määrittämään ionisoitujen proteiinien osasten, peptidien, määrä. Proteiinit voidaan päätellä peptideistä ja näin pystytään määrittämään proteiineihin liittyviä toimintoja hyödyntäen proteiinitietokantoja. Nykyisin käytetty DDA-menetelmä tunnistaa vain runsaimmin esiintyvät ionit, mikä rajoittaa sen hyödyntämistä. Siinä mitattavien ionien valinta on jossain määrin satunnainen, mikä vähentää tulosten toistettavuutta. Vaihtoehtoinen DIA-menetelmä analysoi järjestelmällisesti kaikki ionit ja kyseistä menetelmää on ehdotettu DDA:n tilalle. DIA-menetelmä tuottaa päällekkäisiä peptidispektrejä ja siksi aiemmin ei ole ollut tiedossa, onko se soveltuva menetelmä tai miten sitä olisi mahdollista soveltaa metaproteomiikkaan, jossa on suuri määrä erilaisia peptidejä. Tämä tutkimus esittelee soveltuvia tapoja DIA-menetelmän käyttöön metaproteomiikkadatan analysoinnissa. Työssä osoitetaan, että DIA-metaproteomiikka tuottaa luotettavasti toistettavia tuloksia. DIA-menetelmää käyttäessä riittää, että näyte analysoidaan vain yhden kerran, kun vastaavasti DDA-menetelmän käyttö vaatii useamman analysointikerran. Tutkimuksessa kehitettiin avoimen lähdekoodin ohjelmisto DIAtools, joka toteuttaa kehitetyt DIA-datojen analysointimenetelmät. Lopuksi DIA-analyysiä sovellettiin ruoansulatuskanavan mikrobien ja niiden tuottamien CAZy-entsyymien tutkimiseksi.

AVAINSANAT: DIA, entsyymi, metaproteomiikka, mikrobiomi, peptidi, suolisto

Table of Contents

Abbreviations **8**

List of Original Publications **9**

1 Introduction **10**

2 Review of the Literature **12**
2.1 Gut microbiota and human health 12
2.2 Modern methods for human gut microbiota research 14
2.3 Tandem mass spectrometry based metaproteomics 16
2.3.1 Data-dependent acquisition 17
2.3.2 Data-independent acquisition 18
2.3.3 The challenges of peptide identification 18
2.3.4 Sequence database 19
2.3.5 Spectral library 20
2.3.6 OpenSWATH method 21
2.3.7 DIA spectra deconvolution 21
2.3.8 Annotation 22
2.4 Metaproteomics studies on human gut microbiota 22

3 Aims **24**

4 Materials and Methods **25**
4.1 Datasets 25
4.1.1 Mixture of 12 bacterial strains 25
4.1.2 Human fecal dataset (6 samples) 26
4.1.3 Human fecal dataset from a clinical study 26
4.2 Laboratory methods 26
4.2.1 Preprocessing 26
4.2.2 Protein isolation 27
4.2.3 Liquid chromatography and mass spectrometry 27
4.3 Data-analysis methods and workflow 29
4.3.1 Sequence database and annotations 29
4.3.2 Spectral / pseudospectral library 30
4.3.2.1 Obtaining spectra from DDA data 31
4.3.2.2 Obtaining spectra from DIA data 31
4.3.2.3 Building the library 31
4.3.3 Peptide identification and quantification 32
4.3.4 Peptide annotation 32
4.3.5 Statistical analyses 32

5 Results 34
5.1 DIAtools software package 34
5.1.1 Software environment and packaging 34
5.1.2 Deployment 35
5.1.3 Graphical interface and usage 35
5.1.4 Command line interface 37
5.2 Technical assessment of the DIA-method 38
5.2.1 Peptide identification 38
5.2.2 Taxonomic and functional annotations 39
5.2.3 Reproducibility 40
5.2.4 Quantification consistency between approaches 43
5.2.5 Peptide prevalence 43
5.3 Metaproteomic analysis of human gut microbiota 43
5.3.1 Peptide identifications and taxonomy 44
5.3.2 Carbohydrate-active enzyme profiles 45

6 Discussion 47
6.1 Background, novelty and importance 47
6.2 Results and implications 48
6.3 Bioinformatics software and DIAtools 50
6.4 Limitations and future research 51

7 Summary/Conclusions 53

Acknowledgements 55

References 56

Abbreviations

CAZyme	Carbohydrate-active enzyme
CID	Collision-induced dissociation
CLR	Centred Log-Ratio
CPU	Central Processing Unit
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DNA	Deoxyribonucleic acid
ESI	Electrospray ionization
FDR	False discovery rate
GI	Gastrointestinal
GPU	Graphics Processing Unit
HPC	High performance computing
HPLC	High-performance liquid chromatography
IGC	Integrated reference catalog of the human gut microbiome
iRT	Indexed retention time
KO	KEGG orthology
LC	Liquid chromatography
LCA	Lowest common ancestor
m/z	Mass to charge ratio
MAG	Metagenome-assembled genome
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
NCBI	National Center for Biotechnology Information
PCA	Principal component analysis
PSM	Peptide-spectrum match
RNA	Ribonucleic acid
SWATH-MS	Sequential window acquisition of all theoretical mass spectra
TRIC	TRansfer of Identification Confidence

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Juhani Aakko*, Sami Pietilä*, Tomi Suomi, Mehrad Mahmoudian, Raine Toivonen, Petri Kouvonen, Anne Rokka, Arno Hänninen, Laura L. Elo. Data-independent acquisition mass spectrometry in metaproteomics of gut microbiota – implementation and computational analysis. Journal of Proteome Research 1:432-436, 2020.

II Sami Pietilä, Tomi Suomi, Juhani Aakko, Laura L. Elo. A Data Analysis Protocol for Quantitative Data-Independent Acquisition Proteomics. Methods in Molecular Biology 1871:455-465, 2019.

III Sami Pietilä*, Tomi Suomi*, Laura L. Elo. Metaproteomics boosted up by untargeted data-independent acquisition data analysis framework. Manuscript.

IV Juhani Aakko, Sami Pietilä, Raine Toivonen, Anne Rokka, Kati Mokkala, Kirsi Laitinen, Laura L. Elo, Arno Hänninen. A carbohydrate-active enzyme (CAZy) profile links successful metabolic adaptation of *Prevotella* to its abundance in gut microbiota. Scientific Reports 10(1):12411, 2020.

* shared first author

1 Introduction

Cells, the unit of all living organisms, can be thought to be tiny biochemical machines having programming and executive machinery to govern the cell behavior. This machinery enables cells to react to various environmental conditions such as nutrition availability. The programming code is stored as a nucleotide base pair sequence of DNA or directly as RNA (Alberts, 2017). The code includes instructions for building various protein products, such as enzymes that break down available food molecules (Alberts, 2017). The execution of the programming is controlled by various signalling systems which regulate the transcription of the DNA into RNA with subsequent translation of the RNA into proteins. This is the simplified interpretation of the information flow described by the central dogma of the molecular biology (Crick, 1958a, 1958b). The proteins finally perform various tasks. The protein production is a vital component in many cell mechanisms, which facilitate complex interactions between the cell and the environment and also enable co-operative communities of single cell microorganisms that can even have social structures (Cordero *et al.*, 2012).

The human gastrointestinal tract is a habitat of a complex co-operative, but also concurring, community of microorganisms performing important functions in the human body and having a central role in human health. The microorganisms live by breaking down undigested food macromolecules. The host mucus and shed epithelial cells provides a food source as well as by-products excreted by other microorganisms. However, some by-products can be toxic and the overgrowth of microorganisms not adapted to peaceful symbiosis with the host can lead to disruption to the microbiota homeostasis known as dysbiosis caused by an imbalance of microbiota taxonomic and functional composition. (Tamboli *et al.*, 2004; Moos *et al.*, 2016) Dysbiosis of the human microbiota has been linked to multiple diseases, yet the functional aberrations are unknown (Turnbaugh *et al.*, 2006; Marteau, 2009; Castellarin *et al.*, 2012). Evaluating the functions carried out by microbes is essential for gaining understanding on the role of gut microbiota in human health and its associations to diseases. The functions carried out by the microbiota can be studied by analyzing the expressed proteins and their abundances.

Metaproteomics can be defined as the study of all proteins recovered from a particular environment such as human digestive tract. Similarly, metaproteome is defined in this work as the entire set of proteins expressed by species occupying the environment. In this particular environment, proteins such as carbohydrate-active enzymes are of special interest as they indicate how molecules are digested by the bacterial community. However, new methods and tools are needed to obtain metaproteomic information from complex microbial samples.

The current high-throughput method for producing metaproteomic data is mass spectrometry (MS), combined with liquid chromatography. This method enables the detection of protein fragments called peptides. The instrument is typically run in so-called data-dependent acquisition (DDA) mode in metaproteomics analysis. However, DDA discards other than the most intensive ions to achieve clean non-convoluted spectra. A key limitation of this method is that the ion selection procedure is semi-stochastic introducing inconsistency, which causes less than ideal reproducibility. This inconsistency can be mitigated by running multiple repeated DDA analysis, but this consumes sample material and increases costs. Furthermore, the ion selection procedure of the DDA produces incomplete records and hinders the applicability of the data for ion intensity quantification. Data-independent acquisition (DIA) has been proposed as an alternative to DDA as DIA systematically fragments all ions and produces complete records along the chromatographic profile.

The aim of this study was to discover an analysis approach for metaproteomic DIA data and demonstrate its applicability, and by doing so answer the research question if DIA is a feasible method for metaproteomics. This thesis presents the analysis of complex metaproteome DIA data for the first time and introduces data-analysis methods and DIAtools software package for data-analysis. In **Publication I**, a metaproteomic DIA data analysis was presented where peptides were identified by using a spectral library built from DDA data. The detailed analysis protocol was presented in **Publication II**. The method was later improved, in **Publication III,** to the sole use of DIA data. The method was validated with a laboratory assembled bacterial mixture of 12 bacteria strains and with more complex human fecal samples. Finally, in **Publication IV**, DIA analysis was applied to study carbohydrate-active enzymes of the human digestive tract microbiota to demonstrate the potential of the DIA-analysis as a research method.

2 Review of the Literature

2.1 Gut microbiota and human health

The human gastrointestinal (GI) tract contains a large 250 - 400 m^2 surface area colonised by a complex microbial community. The community, called gut microbiota (Thursby and Juge, 2017), consists of eukarya, archaea and bacteria. In addition, viruses present in the gut (gut virome) are considered part of the gut microbiota. The GI tract of an adult contains 100 trillion (10^{14}) microorganisms that exceed the number of human cells in the body (Sender, Fuchs and Milo, 2016). The microorganisms of the GI tract encode 3 million genes (Rinninella *et al.*, 2019) while, by comparison, the human genome contains approximately only 20 thousand protein encoding genes.

Harbouring such a large microbial community in the GI tract is made possible by having an intestinal barrier limiting microorganism exposure on the human immunity system. The barrier consists of mucus and epithelial layers having antimicrobial proteins, secretory immunoglobulin A (sIgA) molecules and the inner lamina propria where immune cells, such as B cells and T cells, macrophages and dendritic cells reside (Vancamelbeke and Vermeire, 2017). The microorganisms harvest energy by breaking down food molecules and byproducts from other microorganisms. Additionally, the mucus produced by the host provides a source of food molecules for gut microorganisms. The major mucus protein is MUC2 encoded by *MUC2* gene. Over 80% of MUC2 glycoprotein mass consists of oligosaccharide side chains providing energy source for microorganisms which are able to degrade oligosaccharides (Schroeder, 2019).

There is a symbiotic relationship between the host and gut microbiota regulated by complex interactions of microbial-synthesized metabolites and host endocrine and immune systems (Kho and Lal, 2018). The microbiota performs vital functions in the human body such as producing substances needed by the human body including vitamins (LeBlanc *et al.*, 2013). Bacteria augment human digestion by producing enzymes to break complex carbohydrates such as cellulose, hemicellulose and pectin into simple sugars, which are then further fermented to create short chain fatty acids available for human cells as nutrition (Inman, 2011). Microorganisms

take part in the regulation of host immunity and assist to protect against colonization by pathogens (Bäumler and Sperandio, 2016). Microbiota utilizes various means to protect against pathogens, such as competing for nutrients and attachment sites and producing directly or stimulating antimicrobial substance production by the host (Ubeda, Djukovic and Isaac, 2017).

The composition of the gut microbiota varies greatly between individuals even in the same geographical regions, which can be presumed to be possible by the fact that there is functional redundancy between microorganisms allowing multiple different configurations (Moya and Ferrer, 2016; Sonnenburg *et al.*, 2016). Two major factors for shaping the gut microbiota are diet and gut colonization during early life. In the infancy, the primary source of carbohydrates is human milk. While infants are capable to degrade lactose from human milk, the oligosaccharides require an extensive set of glycoside hydrolases (Marcobal and Sonnenburg, 2012). *Bacteroides* and *Bifidobacterium* are able of degrading oligosaccharides, which promotes their growth. During the maturation process, the gut microbiota matures typically to become dominated by Bacteroidetes and Firmicutes. The introduction of solid food changes the ratios of fat, carbohydrate and fiber intake and may initiate the shift toward adult-like microbiota (Homann *et al.*, 2021). The shift increases *Bacteroidetes* and *Firmicutes* related species (Bergström *et al.*, 2014). In particular, the intake of complex plant polysaccharides promotes the growth of *Prevotella*, which has plant molecule digestion enzymes such as xylanase (Linares-Pastén *et al.*, 2021). The colonization events in early life also shape the immune system (Gensollen *et al.*, 2016). During adult life the composition remains fairly stable and after a disruption, such as antibiotic medication, the microbiota typically shifts back to previous composition (Palleja *et al.*, 2018). However, antibiotic treatment may cause a permanent shift to an altered stable gut microbiota composition (Dethlefsen and Relman, 2011). Exposing factors for such a shift are broad-spectrum antibiotics and frequent antibiotic treatments. Altered microbiota composition and function may lead to dysbiosis and degradation of immune responses thus promoting disease outcomes (Francino, 2015).

Fiber content of diet containing microbiota-accessible carbohydrates has an impact on microbiota and it is suggested that a low amount of fiber-derived carbohydrates causes reduced microbiota diversity (Sonnenburg *et al.*, 2016). Variations of microbiota composition associate to dietary preferences and body mass index (Tremaroli and Bäckhed, 2012),(Turnbaugh *et al.*, 2009). A low *Bacteroidetes*-to-*Firmicutes* ratio and high abundance of *Faecalibacterium* associate with high dietary energy intake and overweight, while high abundance of *Bacteroides* associates to diets with high in fat and animal-derived protein content

(Johnson *et al.*, 2017). Plant-derived foods often favour *Prevotella* abundance in microbiota (Ley, 2016).

Different bacterial species are able to take in and excrete various carbohydrates and other metabolites (Sung *et al.*, 2017). For example, *Bacteroides thetaiotaomicron* imports 34 and exports 29 metabolites (Sung *et al.*, 2017). On average, species import and export only three metabolites (Sung *et al.*, 2017). The capability to process a certain carbohydrate is determined by the ability of bacteria to produce suitable Carbohydrate-Activated Enzymes (CAZyme) that are involved in synthesizing or break-down of complex carbohydrates. The enzymes are divided into six modules of families according to their specificity: Glycoside Hydrolases (GH), Glycosyltransferases (GT), Polysaccharide Lyases (PL), Carbohydrate Esterases (CE), Carbohydrate-Binding Modules (CBM) and Auxiliary Activity (AA) enzymes (Lombard *et al.*, 2014). Differences in the produced enzymes can lead to species-specific associations between diets and species abundance, such as high abundance of *Prevotella* being associated with plant-fiber rich diets suggesting adaptation to digest plant based molecules (Precup and Vodnar, 2019).

It has been reported that distinct groups of CAZyme profiles in gut microbiota have been found from individuals and it has been suggested that individuals with different CAZyme profiles are likely to have different carbohydrate metabolic capacities (Bhattacharya, Ghosh and Mande, 2015). This might, in its part, explain the different responses, between individuals, to dietary interventions (Garcia-Perez *et al.*, 2020). Particularly, a group of CAZymes have been found to have positive correlation with body mass index (Bhattacharya, Ghosh and Mande, 2015).

2.2 Modern methods for human gut microbiota research

Metagenomics refers to the study of the whole genetic material from multiple organisms present in a sample such as human gut contents or fecal sample. The method analyses sequences derived from the whole genomes of whole microbial population – the metagenome – broken up randomly into appropriate sequencing length (shotgun sequencing) (Pérez-Cobas, Gomez-Valero and Buchrieser, 2020). The DNA and RNA sequences of the genetic material can be obtained by using Next Generation Sequencing platforms, which are able to produce staggering amount of sequence data in a single run (Clooney *et al.*, 2016). Instruments, such as Illumina NovaSeq 6000, can produce continuous base pair reads up to 2x250 nucleotides (paired-end). Large online databases have been built covering genetic sequences from a wide range of organisms with ongoing effort to assign both taxonomic and functional annotations (Cole *et al.*, 2014; NCBI Resource Coordinators, 2018). The

database sequences are often used as references in studies. Furthermore, there have been efforts to build databases specialized for human microbiota such as Human Microbiome Project (Integrative HMP (iHMP) Research Network Consortium, 2019), which has sequenced microbiomes from major human body sites. Traditionally, whole genome sequencing has offered superior taxonomic identification resolution, and recently introduced metagenome-assembled genome (MAG) analysis can provide representation of actual individual genomes in the sample and allow high-resolution species-level analysis directly from microbial populations (Asnicar *et al.*, 2020). However, MAG analysis requires deep metagenomic sequencing, which can be still prohibitively expensive for large scale studies. Fortunately, the cost of metagenomic sequencing has been coming down as there has been advances in the sequencing techniques and instruments. Metagenomics can be considered as the most influential method for studies where microbe population taxonomic characterization is crucial, and it is not overstated to say that metagenomics has revolutionized biomedical studies (De, 2017).

Another popular approach to microbial community analysis is to sequence variable regions of the bacterial 16S rRNA gene (Pérez-Cobas, Gomez-Valero and Buchrieser, 2020). The method typically targets one or two variable regions of the 16S rRNA gene, which are first amplified by specific primer sequences in a polymerase chain reaction (amplicon sequencing). 16S rRNA gene sequencing has been an appealing choice for microbial community studies because of its affordable price. However, the usage of the subregion has reduced the taxonomic identification resolution achieving generally genus level identifications. With upcoming third-generation technologies, the sequencing of the full 16S rRNA gene is becoming available, which improves taxonomic identification resolution (Johnson *et al.*, 2019).

In addition to the characterization of microbial composition, another concerning biological question is what the microbes are doing. There have been attempts to answer this question by using 16S rRNA gene sequencing data and indirect methods, such as implemented by PICRUSt software (Langille *et al.*, 2013), to predict functions from the 16S rRNA gene abundances. The more sophisticated metagenomic sequencing methods aim at identification of all genes and thus, of all potentially expressed proteins from the metagenomes in order to characterize the whole genetic material present in the microbiota (Franzosa *et al.*, 2018; Silva *et al.*, 2021). However, while proteins and functions can be predicted from the DNA sequence data, the predictions cannot take into account accurately the environmental conditions that might have considerable impact on the activity carried out by the bacteria at the sampling time or point. Metagenomic sequencing can reveal members of a bacterial community and their functional potential, but not if they are active or dormant.

Metaproteomics (Xiong *et al.*, 2015) methods can observe the actual functions carried out by the microbes by identifying and quantifying the peptides (Kleiner, 2019), (Wang *et al.*, 2020) with mass spectrometry. The peptides are protein fragments that are small pieces of amino acid sequences originating from proteins. Mass spectrometry methods are able to quantify and determine the amino acid sequences of the peptides. Later on, by using protein databases the peptides can be taxonomically and functionally annotated. However, mass spectrometry data analysis is challenging, and peptide identification typically depends heavily on protein sequence databases, which limits the identifiable peptides to those found in the databases. This also implies that the quality and completeness of databases are highly influential in metaproteomics analyses.

In between metagenomics and metaproteomics there are metatranscriptomics methods that study the transcripts (mRNA molecules) and thus, the transcriptional activity of the microbes (Bashiardes, Zilberman-Schapira and Elinav, 2016). The RNA is converted into cDNA and is read with DNA sequencing technologies. Transcriptomics reveals which proteins are in the transcription process and about to be produced at the given moment. However, mRNA has a short half-life and can be degraded instead of being translated into proteins. The presence of mRNA does not indicate the presence of proteins and vice versa. Instead, the correlation of the abundances between the two is reported to be poor (Maier, Güell and Serrano, 2009; Vogel and Marcotte, 2012), which suggests that proteomics gives better indicative for protein production and degradation.

Metabolomics is the study of chemical products, metabolites, left behind by cell processes. It is also used to study the functions along with the metaproteomics. Metabolomics enables sample classification and discrimination based on metabolite profiles and detection of metabolic fingerprints of specific cellular processes (Krastanov, 2010; Johnson, Ivanisevic and Siuzdak, 2016). The two main instruments used in metabolomics research are nuclear magnetic resonance spectroscopy and mass spectrometry (Segers *et al.*, 2019).

Although metagenomics, metatranscriptomics, metaproteomics and metabolomics have similarities, they provide complementary information and can all be applied to study a single dataset to acquire more comprehensive information than is obtainable alone with a single technology.

2.3 Tandem mass spectrometry based metaproteomics

The current high-throughput method for protein identification is mass spectrometry (Han, Aslanian and Yates, 2008). The proteins are extracted and

digested into peptides with a proteolytic enzyme, such as trypsin, that cuts the amino acid sequences after lysine or arginine except when followed by proline. Using a technique called liquid chromatography (LC) (Pitt, 2009), the peptides are ordered by their properties such as isoelectric point, hydrophobicity or size of the peptide. The physical properties of the molecule determine the time the peptide takes to travel through an LC column. The peptides emerging from the LC column at discrete time-points are fed into a mass-spectrometer where they are ionized and sent into a detector recording the mass-to-charge (m/z) [1] ratios of the ions where z is the charge [2] number. Each detected ion is visible as a peak at a specific location in the m/z range. A popular mass spectrometer type is an Orbitrap analyzer which captures ions into certain trajectory (Hu *et al.*, 2005). The trajectory oscillation determines the mass-to-charge ratio.

All or a subset of peptide ions are selected and further fragmented with a technique like *collision-induced dissociation* (CID) where ions are collided with inert collision gas such as argon. The collisions break each peptide molecule into two fragments along the peptide backbone. There are multiple possible potential breaking points at the backbone and thus, a single type of peptide produces multiple different fragments. Mass-to-charge ratio is recorded for each fragment that has retained a charge thus producing fragment spectra.

This procedure is called tandem mass spectrometry, also known as MS/MS, because it requires two mass analyzers and produces an output that contains peptide ions (aka. precursors) and their fragment ion m/z spectra. LC combined with MS/MS produces multiple precursor and fragment spectra over time where each time point associates to a certain type of peptide. A time point is called *retention time*, as it describes the delay caused by LC.

2.3.1 Data-dependent acquisition

In data-dependent acquisition (DDA) mode, the instrument chooses a small set of most abundant precursors (peptides) and fragments only those one by one. Thus, in this mode, the instrument records the m/z values of a precursor ion and its fragments. The output is an untangled precursor and fragment spectra. Only the most abundant peptides are detected, while peptides with small abundance are undetected. This is especially an issue with complex samples where several peptides may elute

[1] Note, m/z unit is u/e, where u is the unified atomic mass unit and e is the charge of proton when positive or electron when negative.
[2] A charge of an electron or proton corresponds to one unit.

at a given point in time. Additionally, the selection is a semi-stochastic process meaning that a repeated analysis of a sample gives a partially different set of detected peptides resulting in a poor run-to-run reproducibility (Hu, Noble and Wolf-Yadlin, 2016). It takes several repeated analysis runs to get a good coverage of peptides that can be detected with DDA. Furthermore, the selection procedure prevents ions from being recorded along the full chromatographic peak profile and for this reason DDA produces incomplete records for the purpose of peptide quantification. DIA has been proposed to overcome DDA limitations (Barkovits *et al.*, 2020).

2.3.2 Data-independent acquisition

In data-independent acquisition (DIA) mode, the instrument systematically fragments all peptide ions with *sequential window acquisition of all theoretical mass spectra* (SWATH-MS (Ludwig *et al.*, 2018)) strategy in the following way. At each retention time point, the full m/z range of detectable peptides is recorded and divided into subranges. The instrument chooses all peptides belonging to the given subrange to be fragmented and recorded. The instrument processes through all subranges consecutively. For example, if the full peptide m/z range is from 400 to 1000, the instrument first chooses peptides from 400 - 415 to be fragmented and then proceeds to the next consecutive subrange 415 - 430. This process is repeated until the last range 985 - 1000 is reached. This procedure produces a single spectrum from the precursor ions from the full range and related fragment spectra from each window range. If the window range at the specific retention time contains multiple peptides, the window readout contains entangled fragment spectra.

2.3.3 The challenges of peptide identification

Detecting a peptide is challenging even from a simple mass spectrum. Ions have a mass probability distribution instead of a single mass as they might contain varying isotopes with associated probabilities. For this reason, using short peptides are preferred over proteins as the variability increases with longer sequence. Also, the amino acid masses can be close to each other and be even indistinguishable like leucine and isoleucine. The particles might become charged multiple times during ionization, which influences the detected mass-to-charge ratio. In practice, the fragmentation step in MS/MS leaves some fragments without a charge, thus the MS/MS does not produce ions from all theoretical fragments making the determination of the amino acid residues and especially their order very difficult. Generally, the recorded mass spectra also contain noise from electrical and chemical deviations and machine originated artifacts. For these reasons, the de-novo (Muth

and Renard, 2018) approaches that infer the peptide sequences directly from the experimental MS/MS spectra have not been particularly successful.

2.3.4 Sequence database

The identification of peptides by their mass per charge ratios is difficult because multiple different peptides can have the same m/z ratios down to fragment level. The de facto standard approach currently to mitigate this difficulty is to use protein sequences to provide a limited search space for all identifiable sequences. There are multiple online databases, which provide protein sequences and related annotations such as taxonomic information (originating species) and functional information. Two popular databases are Uniprot (UniProt Consortium, 2021) and National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov). Furthermore, there are specific databases such as *Integrated reference catalog of the human gut microbiome* (IGC), which contains proteins from human gut microbes (J. Li *et al.*, 2014).

During analysis, the protein sequences from databases are typically appended with decoy sequences, which do not exist in nature but are similar to the true proteins, to facilitate controlling the *false discovery rate* (FDR). The added sequences are manufactured typically by reversing or shuffling the true sequences. The modified sequences are labeled as decoys, so that they can be recognized as decoys at later stages (Elias and Gygi, 2007, 2010).

The protein sequences are digested in-silico into peptides with knowledge of the digestion enzyme. The theoretical masses of the peptides and their fragments are calculated and compared against the observed spectra from the instrument (Eng, McCormack and Yates, 1994; Perkins *et al.*, 1999; Craig and Beavis, 2004). Each match is scored and the highest scoring peptides are reported as the search result. The theoretical fragment peaks have no intensities assigned to them, which limits the identification accuracy. Only recently, deep learning approaches have been emerging and applied to assign the intensities (Gessulat *et al.*, 2019).

The number of the sequences in the database should provide a good coverage for the peptides expected to be identified, but excessive numbers of sequences should be avoided as they hinder the ability for finding the correct match by providing a high number of alternatives to choose from (Kumar, Yadav and Dash, 2017). Ultimately, the match-based technique defines a correct match in terms of match scores and false discovery rate. A common technique of confining the search space is to perform a proteogenomic analysis for the samples and utilizing the predicted proteins for construction of the database (Qin *et al.*, 2010).

2.3.5 Spectral library

In comparison to DDA, it is much harder to identify a peptide from the DIA data where spectra are convoluted by multiple peptides co-eluting at the same time. Furthermore, the spectra produced by mass spectrometers are imperfect and only partially match to the theoretical peptide spectra. For these reasons, the DIA method typically depends on a spectral library (Yates *et al.*, 1998; Lam *et al.*, 2007). A spectral library is a collection of peptide-spectra, where each peptide-spectrum consists of amino acid sequence, precursor and related fragment ion spectrum. Typically, the spectra are collected from DDA data.

A DDA-based spectral library is optimally generated with the same instrument and from the same samples as the DIA data. It is important that the instrument settings, especially the collision energy, closely match those used with DIA data so that the properties such as the fragmentation pattern are similar. These are the most important characteristics for generation of the library, which should capture precursor ions and their fragment ions (i.e fragmentation pattern) with intensities and retention times. The peak intensities are strongly characteristic for a peptide sequence and are taken into account in library-based search, but ignored in theoretical spectra search.

A popular approach to generate a spectral library is first to analyze the sample data by pooling multiple samples together and then analyzing each pooled sample with DDA method using multiple injections. The DDA data from the pooled samples are analyzed with search engines such as Comet and X!Tandem (Eng, Jahan and Hoopmann, 2013), (Craig and Beavis, 2004). The results from multiple search engines are aggregated to increase the number of the peptide identifications because different search engines tend to identify somewhat different sets of peptides (Jones *et al.*, 2009),(Shteynberg *et al.*, 2013). This generates a library of peptide sequences with corresponding consensus spectra, where peaks are averaged among the replicates, for each peptide ion.

Library-based search methods are highly dependent on the quality of the spectra in the library. Low quality spectra can impair library-based searching. For this reason, it is important to include only high confidence peptides in the library. This is done by filtering peptides with a strict FDR threshold (Aggarwal and Yadav, 2016). The library also retains information from which proteins the peptides originate. Building a library might be unfeasible in cases where there is a high diversity between the samples as the library should contain all the peptides expected to be found.

2.3.6 OpenSWATH method

With the presence of a library, the peptides can be identified and quantified from DIA data with OpenSWATH method (Röst *et al.*, 2014). First, the retention times are aligned with iRT peptides, which are a known set of peptides acting as markers for specific retention time points. The aligned retention time information from the library is used to extract the corresponding MS/MS data along the chromatogram profile while the m/z value of the precursor is used to select the correct SWATH window. Precursor and fragment ion traces, indicated by the library, are extracted from the DIA data. Co-eluting traces, called peak groups, are scored by their elution profile and by the correspondence of the fragment ion-intensity and retention time to those indicated by the library. The peak groups are quantified by summing the integrated peak area of each transition ion in a peak group along the chromatographic profile. TRansfer of Identification Confidence (TRIC) alignment algorithm can be applied to cross-align fragment ion trace groups between the samples to facilitate a consistent dataset-wide ion identification and quantification (Röst *et al.*, 2016) .

2.3.7 DIA spectra deconvolution

Recently, methods have been developed to deconvolute the DIA spectra. The deconvolution, in this context, refers to the separation of spectra associated to a peptide from the spectra from the other peptides. Tools, such as DIA-Umpire (Tsou *et al.*, 2015) and Group DIA (Li *et al.*, 2015), perform deconvolution relying on the coexistence of precursor and fragment spectra. The deconvoluted spectra can then be processed with conventional DDA methods where spectra are searched with assistance from protein sequence databases.

DIA-umpire deconvolution process, utilized in this study, starts by feature detection to locate precursor-ion signals in MS1 and MS2 data. The process begins with peak curve, which is a mass trace continuous in time, detection from the LC-elution profile. The peak curves are smoothed by B-spline interpolation and split by using continuous wavelet transformation using a Mexican-hat wavelet to separate multiple maxima peak curves into unimodal peak curves. The intensities of the unimodal peak curves are determined as the maximum intensity (at the apex). The peak curves are grouped together on the basis of RT apex distances and m/z spacing to form an isotopic cluster, as the presence of isotopes helps to distinguish precursor signals from noise. Higher number of peaks in a group increases the probability for the group presenting a true peptide and acts as a quality measure. The presumed peptide ions are divided into quality tiers where Q1 indicates ions having three or more isotopes while Q2 means only two isotopic peaks. Ions which have an isotopic envelope when counting also unfragmented ions at MS2 are assigned to quality tier

Q3. Ions with no detected isotopic envelope are discarded. Fragment ion detection is performed in a similar manner, but the isotopic envelope checking is applied only after precursor-fragment grouping. The precursor-fragment ion groups are formed by pairing highest correlating co-eluting profiles, i.e. retention time and apex peak, of precursors and possible fragments restricted by proper m/z range. The convolution process is described in detail in the original publication (Tsou *et al.*, 2015).

2.3.8 Annotation

In protein sequence databases, the functional and taxonomic annotation is typically assigned to the sequences. Commonly, the set of proteins present in a sample is inferred from the identified peptides (Nesvizhskii and Aebersold, 2005). However, this can be tricky as multiple proteins can share one or more peptides. Among other strategies (Huang *et al.*, 2012), a popular strategy for protein inference is to identify a minimum set of proteins that explain the observed peptides. This is called Occam's razor principle, which dictates that the simplest explanation is the most probable one (Nesvizhskii *et al.*, 2003; Kumar, Filipski and Greenbaum, 2004). However, in some cases, a set of proteins can even share exactly the same set of peptides. These proteins are typically grouped together without means to make a difference between them.

When applying spectral library based peptide identification techniques, the library may contain a list of possible originating proteins, which can be used to assign annotation for the identified peptide. This approach was followed in this work. There are caveats such as multiple originating proteins, in which case a peptide can have multiple annotations. However, the peptide annotation is ambiguous only if originating proteins have conflicting annotations.

2.4 Metaproteomics studies on human gut microbiota

The metaproteomic research is gradually gaining momentum to uncover the functions carried out of human gut microbiota and fill in information that has been unobtainable with other methods. In the following, a small selection of studies is briefly discussed for an insight into metaproteomic studies on human gut microbiota and what those studies have revealed.

The study conducted by Verberkmoes et al. compared the gut metaproteome, obtained from fecal samples, of two healthy females (twins). It was discovered that the metaproteome functions differ from those predicted by metagenomics (Verberkmoes *et al.*, 2009). Proteins for translation, energy production and carbohydrate metabolism were found more than predicted. Also, human

antimicrobial peptides were found indicating ongoing host response to the microbiota.

Kolmeder *et al*. applied metaproteomics to study if temporal stability of intestinal tract microbiota is reflected at the functional level (Kolmeder *et al.*, 2012). Metaproteome was obtained from fecal samples collected from three healthy individuals over a period of six to 12 months. The results of this study indicated that fecal metaproteome is subject-specific and stable during the studied period of 12 months. The study found a stable common core of approximately 1,000 proteins, mainly involving carbohydrate transport and degradation, in each of the subjects. The study reports that Clusters of Orthologous Groups of proteins (COGs) (Tatusov *et al.*, 2000) could be assigned to over 70% of identified proteins. Still, this leaves a significant proportion of proteins unannotated, which implies there is room for improvement in methodology and databases.

In a later work, Kolmeder *et al*. studied fecal samples from 16 healthy adults in a probiotic intervention trial (Kolmeder *et al.*, 2016). Half of the subjects consumed placebo and half consumed *Lactobacillus rhamnosus GG* for three weeks. A common core of shared microbial function was identified from all subjects, but no significant changes in the metaproteome were found to be attributable to the probiotic intervention.

In a recent study, Long *et al*. studied the differential expression of microbial proteins between healthy individuals and individuals with colorectal cancer. A set of differently expressed microbial proteins were found related to iron intake/transport; oxidative stress; DNA replication, recombination, and repair, which resulted from high local concentration of iron and high oxidative stress in the large intestine of the patients with colorectal cancer (Long *et al.*, 2020).

In an exploratory study (preprint available) Armengaud *et al.* studied the altered micobiota molecular functions associated with high levels of *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) RNA in fecal samples. The identified molecular functions that were altered in the microbiota highlighted mechanisms that may contribute to vicious disease cycles (Armengaud *et al.*, 2021).

These and other metaproteomic studies have given valuable information, typically collected from fecal samples about human gut microbiota. On the methods, including the coverages of protein databases, there is still room for improvements. Kolmeder *et al.* 2012 reported 70% of identified proteins could be assigned with Clusters of Orthologous Groups (COGs) (Tatusov *et al.*, 2000). The reproducibility is potentially also a concern which is, interestingly, mostly not discussed in these studies (Barkovits *et al.*, 2020). Finally, metaproteomic studies would likely benefit from quantification, which is currently underutilized (Kolmeder and de Vos, 2014).

3 Aims

The aim of this thesis was to study the data analysis approaches for metaproteomic DIA data and to develop novel methods for human gut microbiota analysis.

The following objectives facilitate the achievement of this aim:

1. Study the feasibility of DIA for metaproteomics
2. Develop an approach for DIA metaproteomic data-analysis.
3. Implement a data-analysis tool for DIA metaproteomics.
4. Investigate the association of proteins expressed by microbial taxa and taxonomic abundances, with a particular focus on carbohydrate-active enzyme analysis.

4 Materials and Methods

4.1 Datasets

Two metaproteomic datasets were prepared for technical validation of the data-analysis methods in **Publications I and III**. The first dataset was a laboratory assembled mixture of 12 bacterial strains and the second dataset was a human fecal dataset from six healthy individuals. The **Publication IV** presents a large human fecal dataset from 63 healthy donors.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD008738 for the technical validation sets and PXD017059 for the large fecal dataset.

4.1.1 Mixture of 12 bacterial strains

The laboratory assembled mixture, referred as 12mix, contains the following 12 bacterial strains belonging to the following species: *Bacteroides vulgatus, Parabacteroides distasonis, Enterorhabdus sp., Bifidobacterium pseudocatenulatum, Escherichia coli, Streptococcus agalactiae, Bacteroides fragilis, Alistipes onderdonkii, Collinsella aerofaciens, Clostridium sordellii, Eubacterium tenue, and Bifidobacterium bifidum*. The strains were isolated from fecal samples of three human donors and were grown on fastidious anaerobe agar (LAB 090; LAB M, UK) and tentatively identified by sequencing their 16S-rDNA. Prior to mixing, the bacterial cell counts were equalized to 10×10^8 cells/mL using flow cytometry (bacteria counting kit for FLO, Fisher Scientific) and 1×10^8 cells of each isolate were added to the final mixture.

The proteins were isolated from the mix four times and analyzed in DDA mode with a single injection. In addition, all the four peptide isolations were pooled together and analyzed with four injections in DDA mode. From the resulting spectrum files, three DDA and DIA files were selected for data-analysis. The raw spectrum files used in the data-analysis are listed in the supplement of **Publication III**.

4.1.2 Human fecal dataset (6 samples)

Human fecal samples were collected from six anonymous healthy individuals, whose ages ranged from 20 – 60 (three men and three women), under the permission of Southwest Finland Hospital District. Each biological sample was analyzed with a single DIA injection. In addition, all six samples were pooled together and the pooled sample was analyzed in DDA mode with six injections. The raw spectrum files used in the analysis are listed in the supplement of **Publication III**.

4.1.3 Human fecal dataset from a clinical study

The human fecal samples (63 samples) were originally collected for a mother infant dietary intervention trial (ClinicalTrials.gov, NCT01922791) from healthy overweight and obese pregnant women at the trial baseline. Three-day-food diaries recorded by women in the week before sample donation. Each fecal sample was analyzed in DIA mode with a single injection. In addition, seven pooled samples were created and each pooled sample was analyzed with six injections using DDA mode.

4.2 Laboratory methods

4.2.1 Preprocessing

The microbial mixture (12mix) contained twelve different strains isolated from fecal samples of three human donors grown on fastidious anaerobe agar (LAB 090; LAB M, UK) and were tentatively identified by sequencing their 16S rDNA. Bacterial cell counts were determined by in situ labelling of a 16S-RNA consensus sequence with a fluorochrome to allow detection by flow cytometry. Prior to mixing, the bacterial cell counts were equalized to 10×10^8 cells / ml and 1×10^8 cells of each isolate were added to the final mixture. The kit (LIVE/DEAD™ BacLight™ Bacterial Viability and Counting Kit, Fisher Scientific) also allowed to ensure the viability of cells subjected to proteomic analysis.

The fecal samples were put at + 4 °C immediately after their collection and an aliquot of the sample was stored at − 80 °C within hours. Thawed fecal material was dissolved in phosphate buffered saline (PBS) at + 4 °C including protease inhibitor (aprotinin) and allowed to dissolve with gentle agitation. Bulk material was removed by spinning the samples at low G force, and supernatant containing bacteria was collected in consecutive steps after repeatedly dissolving the remaining particulate matter to PBS. Finally, all aliquots of the supernatant were spun at high G force to

bring bacteria down. Pelleted bacteria were resuspended in a smaller volume to allow cell counting. Bacteria counts were equalized similarly to the bacterial mixture described above. Following flow cytometry, an aliquot of supernatant containing 10^8 bacteria was used to prepare each sample. Bacteria for each sample were pelleted down and stored as pellets at − 80 °C until protein isolation.

4.2.2 Protein isolation

Proteins, from the bacterial mixture, were isolated by using a Barocycler instrument NEP3229 (Pressure BioSciences Inc., South Easton, Easton, Massachusetts, USA), which uses pressure cycles to lyse the cells. Protein concentrations were determined with Bradford method. Fifty µg of protein was used for trypsin digestion. The proteins were reduced with dithiothreitol (DTT) and alkylated with iodoacetamide. The trypsin digestion was performed in two steps: first trypsin was added in a 1:50 ratio and digested for 4h and then with a 1:30 ratio overnight at 37 °C. After digestion, the peptides were desalted using a SepPak C18 96-well plate (Waters Corporation, Milford, Massachusetts, USA).

Proteins, from the fecal samples, were extracted from pelleted bacteria using NoviPure Microbial Protein kit (MO BIO Laboratories Inc.) according to manufacturer's instructions. Protease inhibitors (Pierce Protease Inhibitor Tablets, Thermo Scientific) were added to the lysis buffer. Mechanical cell lysis was performed by bead-beating using TissueLyser-device (Qiagen) and two 5 min cycles at 50 Hz. Between cycles samples were placed on ice for 5 min. Protein concentrations were determined by DC Lowry (BioRad) method. Fifty microgram proteins were digested by trypsin using filter aided sample preparation (FASP) method39. Peptides were desalted by SepPac C18 96-well plate (Waters), evaporated to dryness and dissolved in 0.1% formic acid. Peptide concentrations were checked with NanoDrop device (Thermo Fisher Scientific), and iRT peptides (Biognosys AG) required for retention time calibration were added to all samples according to manufacturer's instructions before mass spectrometry analysis.

4.2.3 Liquid chromatography and mass spectrometry

The LC-ESI-MS/MS analyses were performed on a nanoflow HPLC system (Easy-nLC1200, Thermo Fisher Scientific, Waltham, Massachusetts, USA) coupled to a Q Exactive HF mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with a nano-electrospray ionization source. Digested protein samples in the range of 500 ng to 2 µg (varying with the sample sets and acquisition modes) were first loaded on a trapping column and subsequently separated inline on a 15 cm C18

column (75 µm × 15 cm, ReproSil-Pur 5 µm 200 Å C18-AQ, Dr. Maisch HPLC, Ammerbuch-Entringen, Germany) and on a 40 cm C18 column (75 µm × 40 cm, ReproSil-Pur 1.9 µm 120 Å C18-AQ; Dr. Maisch HPLC GmbH, Ammerbuch-Entringen, Germany). The mobile phase consisted of water with 0.1% formic acid (solvent A) or acetonitrile/water (80:20 volume/volume) with 0.1% formic acid (solvent B). Depending on the dataset, peptides were eluted from a gradient (7% to 35%) of solvent B during a 90 min elution period, followed by a wash with undiluted solvent B; or in two-steps, first from a gradient (7% to 25%) of solvent A during a 75 min elution period, followed by a gradient (25 to 35%) of solvent B during a 15 min elution period.

Proteins from the samples were identified and quantitated using data independent acquisition (DIA) based MS method. DIA quantification was performed with a resolution of 30 000. AGC target was set at $5*10^5$ with automatic maximum injection time. The DIA MS method covered a mass range from 400 to 1,000 m/z through 40 consecutive windows with isolation width of 15 m/z. For the DIA analysis, the samples were spiked with indexed retention time peptides (HRM Calibration kit, Biognosys, Schlieren, Switzerland) and each sample was injected once.

The DDA method consisted of an Orbitrap MS survey scan of mass ranges 375-1500 and 380–1,200 m/z were used followed by higher energy collisional dissociation (HCD) fragmentation for 15 and 20 of the most intense peptide ions, depending on the dataset. The mass ranges and the number of the chosen intensive ions varied slightly between the data sets. The survey scan was done with 120 K resolution. AGC target was $3*10^6$ and max injection time 50 ms. Monoisotopic masses were then selected for further fragmentation for ions with 2 to 5 charge within a dynamic exclusion range of 30 s and a minimum intensity threshold of $2*10^4$ ions. Precursor ions were isolated using the quadrupole with an isolation window of 1.4 m/z, NCE 27% was used, the AGC target was set at 10^5 and maximum injection time was 50 ms. For the DDA analysis, the samples of each type were pooled and spiked with indexed retention time peptides (HRM Calibration kit, Biognosys, Schlieren, Switzerland). The pooled 12mix samples were analyzed three times and the pooled human fecal samples six times in DDA mode. Clinical study samples were pooled into seven (7) pools and each pool was analyzed five (5) times with data dependent acquisition method (DDA).

4.3 Data-analysis methods and workflow

The overall metaproteomics DIA analysis workflow is presented in **Figure 1** [3]. The first step of the process is the creation of a spectral/pseudospectral library. The library can be built from DDA data or directly from the DIA samples. In this work, the former approach is referred as *DDA-assisted* and the latter is referred as *DIA-only* approach. With the library, peptides are identified and quantified from DIA data. The resulting peptide intensity matrix contains identified peptides and their intensities, i.e. abundances, in each sample. The analysis workflow is implemented by DIAtools software package, which was developed as part of this work.

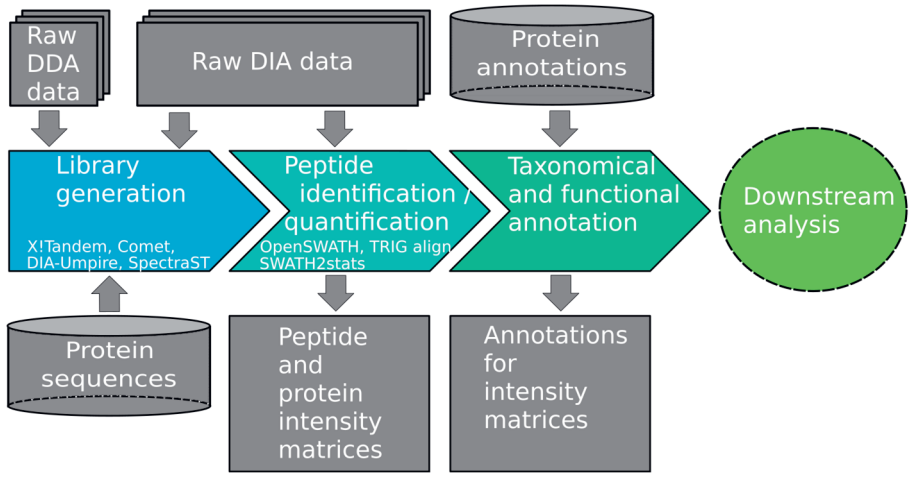

Figure 1. Metaproteomics DIA data analysis workflow with essential proteomic software components.

4.3.1 Sequence database and annotations

The data analysis methods deployed in this study require a database of protein sequences representing the search space for identifiable peptides. This study uses

[3] The first iteration of the workflow and the figure is presented in **Publication I**. The first iteration of the method has only DDA-assisted method while the current one has both DDA-assisted and DIA-only methods.

IGC [4], which is a very large protein sequence database from human gut microbiota containing almost 10 million protein sequences. The database proteins have been obtained by locating *open reading frames* (ORF) from genomic sequence data and translating the sequences to protein sequences. The database contains taxonomic and functional annotations such as phylum, genus and KEGG functional categories for the protein sequences.

A subset of the proteins are carbohydrate-active enzymes, discussed in **Publication IV**. The CAZyme annotations were obtained from the *Carbohydrate-Active enZYmes Database* (Lombard *et al.*, 2014) (www.cazy.org). A Python (https://www.python.org/) script was written to fetch and extract bacterial enzyme identifiers and family modules from HTML pages as the database does not support bulk download. This resulted in a total of 367595 protein identifiers. With the identifiers, GenBank formatted protein records were downloaded from the NCBI. Protein sequences and protein products (enzymes) were extracted from the records and combined to the family module information. The protein sequences were aligned, with STAR aligner (Dobin *et al.*, 2013), to IGC sequences to identify which proteins in the IGC are carbohydrate-active enzymes. The identified IGC proteins were annotated with CAZyme family module and protein product information. From all IGC proteins, 5.8% were identified as carbohydrate-active enzymes and 4.6 % were annotated unambiguously with a single CAZyme family and enzyme.

The sequence search space was extended to contain iRT peptides (Escher *et al.*, 2012), contaminants and decoy sequences. Trypsin, Lysis enzyme and reviewed human proteins from Uniprot sequences were added to identify and filter spectra from the contaminants. The decoy sequences were generated by reversing the sequences. The sequence reversing decoy generation strategy was chosen as it is a widely used approach and retains the properties of the original protein and the behavior characteristics through various processing steps such as the ionization (Elias and Gygi, 2010).

4.3.2 Spectral / pseudospectral library

Two different approaches for building a library of precursor ions and related fragment spectra were introduced to facilitate the SWATH-MS data analysis. In addition to the traditional approach where the library was built from DDA data, the library was built directly from the DIA data. Library built from DDA data is called

[4] The online summary of the database content is available at https://db.cngb.org/microbiome/genecatalog/genecatalog_human/.

a spectral library and the library built from DIA data is called a pseudospectral library.

4.3.2.1 Obtaining spectra from DDA data

For each dataset, the isolated proteins were pooled together to create a single or a small set of samples that represent the peptides present in the whole dataset. The pooled samples were analyzed in DDA mode to create precursor and related fragment spectra for building a spectral library.

4.3.2.2 Obtaining spectra from DIA data

The DIA spectrum was deconvoluted by detecting co-eluting peak groups and the co-existence of a precursor ion and related fragment spectrum and separating those from other precursor and fragment ion spectra. This resulted in a spectrum that resembled the properties of DDA thus making the resulting pseudospectra suitable replacements for DDA spectra. The deconvolution is implemented in DIA-Umpire software (Tsou *et al.*, 2015). The deconvolution process, for producing all quality tiers of DDA-like spectrum, was integrated into the library build process of DIAtools in **Publication III** to achieve completely DDA-free analysis.

4.3.2.3 Building the library

The library is built based on protocol by Schubert et. al. (Schubert *et al.*, 2015). The obtained spectrum or pseudospectrum data were searched against the protein database with Comet (Eng, Jahan and Hoopmann, 2013) and X!Tandem (Craig and Beavis, 2004) search engines, where the search engines assigned each spectrum to the highest scoring peptide match, a.k.a. Peptide-Spectrum Match (PSM). The search results from both search engines were combined to maximize the number of PSMs.

The IGC database was previously appended with a reversed copy of each sequence that were tagged as decoys. The decoys found by the search engines were considered as false positives and their proportion used in the FDR correction. The PSMs from search engines were scored with PeptideProphet (Keller *et al.*, 2002) and subsequently with iProphet (Shteynberg *et al.*, 2011) to calculate more accurate posterior probabilities. The FDR of groups of PSMs mapping to the same sequence database identifiers were calculated with Maya (Reiter *et al.*, 2009) software.

Finally, all precursors and related fragments, the m/z values of b and y type ions, were compiled into a single library by applying the very strict FDR (< 0.01) filtering to include only the high confidence spectra. Also iRT peptides were recorded in the library. The library was compiled with SpectraST software (Lam *et al.*, 2007). For

peptides having multiple spectra, the spectra were combined into a single consensus spectrum. Six of the most intensive fragments were retained in the spectra as recommended in the literature (Reiter *et al.*, 2011). Peptides having less than six fragments were filtered out. Resulting spectral library was stored in a standard TraML format (https://www.psidev.info/traml) (Deutsch *et al.*, 2017). As a part of the library build process, decoy entries were added to the library to enable FDR control in the following peptide identification phase. The library build process is fully automatized in DIAtools, **Publications I, II, III**.

4.3.3 Peptide identification and quantification

Once the spectral/pseudospectral library was built, the peptides from the DIA mass spectra files were identified against the library. The process searches and extracts the chromatographic profiles of library ions from mass spectrum DIA data and scores co-eluting profiles, from which true positive identification are determined under FDR scheme. This results in peak groups. Consistent identification and quantification is achieved by aligning peak groups across samples which boosts identification confidence of peak groups. The integrated area under the peaks is used to determine ion intensities. The ion identification and quantification procedure was carried out with OpenSWATH workflow (Röst *et al.*, 2014) with TRIC alignment (Röst *et al.*, 2016).

4.3.4 Peptide annotation

The peptide annotations, such as taxonomy or functional category, were assigned directly according to the possible originating proteins from the spectral/pseudospectral library. A peptide may originate from multiple proteins and can thus have multiple annotations. An annotation was assigned for a peptide if there was no evidence of conflicting annotations. Specifically, this means that a peptide can have originating proteins with no annotation, but all proteins having an annotation must have the same annotation. Otherwise, the peptide was labeled as having an ambiguous annotation.

4.3.5 Statistical analyses

The DIA-data analysis workflow (**Figure 1**) utilizes FDR correction for peptide identification with cut-off thresholds 0.01 and 0.05 (Reiter *et al.*, 2009; Röst *et al.*, 2016).

The downstream statistical analyses of **Publication IV** were conducted using R software version 4 (https://www.R-project.org/). The peptide intensity matrix was transformed using centred log-ratio transformation (CLR) (Aitchison, 1982). Differently expressed peptides between the sample groups, selected according to the *Bacteroides* 16S rRNA gene abundance, were assessed with ROPECA (Mokkala *et al.*, 2016; Suomi and Elo, 2017) using the modified t-test. Principal component analysis of the peptide intensity data was performed with scikit-learn (https://scikit-learn.org/).

5 Results

To validate the analysis, the DIA-only and DDA-assisted approaches were technically assessed with the laboratory assembled mixture of 12 species and with the set of six human fecal samples in the original **Publications** from **I** to **III**. In **Publication IV,** the method was applied to 63 human fecal samples to study CAZy expression in human gut microbiota.

In the course of the subsequent studies, the DIAtools open source software package was improved with new versions, each adding new capabilities and enhancements, may cause results to vary slightly from those presented in the publications.

5.1 DIAtools software package

The workflow (**Figure 1**) for the analysis of DIA metaproteomic data was implemented as a software package DIAtools in its latest version 3.0, **Publication III**. The previous versions of the workflow, implemented by earlier versions of DIAtools, required the usage of DDA sample for building the spectral library, **Publications I**, **II**. The earlier versions provided solely command line interface, but the latest version provided also a modern web-based graphical interface. DIAtools is implemented with Python and JavaScript programming languages. DIAtools is released as open source software, licensed under the GNU General Public License (gnu.org). The source code and install instructions of the software is available at GitHub (https://github.com/elolab/diatools).

Even though DIAtools is designed for highly complex metaproteomics data, it can be used to analyze simpler proteomics DIA data as well.

5.1.1 Software environment and packaging

The DIA analysis workflow (**Figure 1**) consists of a lengthy series of various operations implemented by multiple tools. The most essential components in the library build phase are: DIA-Umpire, X!Tandem, Comet and SpectraST. In the

subsequent peptide identification and quantification phase, the essential components are: OpenSWATH, TRIG alignment and SWATH2stats (Blattmann, Heusel and Aebersold, 2016). The taxonomic and functional annotation phase is done by DIAtools workflow without calling external programs.

DIAtools provides all the required software as a container. Majority of the required software comes from software collections: OpenMS (Sturm *et al.*, 2008), Trans-Proteomic Pipeline (TPP) (Deutsch *et al.*, 2010) and msproteomicstools (https://github.com/msproteomicstools/msproteomicstools). The formal and complete description of the container image content is provided in the Dockerfile included in the DIAtools source code repository (https://github.com/elolab/diatools). The DIAtools software package is compatible with container technologies such as Docker (https://docker.com) and Podman (https://podman.io).

5.1.2 Deployment

DIAtools is deployed by downloading the container image file and running the container. The container can be run in environments supporting x86/64 based architecture. Detailed usage instructions are available in the software manual (https://github.com/elolab/diatools).

5.1.3 Graphical interface and usage

The graphical user interface is implemented as a web service running in the container and is accessible through web browsers. The web service backend is implemented with Pyramid (https://trypyramid.com/) framework while the frontend is built with Vue (https://vuejs.org/) framework. The service can be made accessible from the network, allowing deployments on servers, or the accessibility can be limited to the local machine for single workstation deployments.

Peptide identification and quantification is initiated from the analysis windows, **Figure 2**. There, DIA mass-spectrum data and protein sequence databases are provided for the analysis. Various analysis and instrument specific settings are available, such as instrument specific precursor and fragment tolerances and FDR thresholds for the analysis. Optionally DDA data can be provided for building a spectral library. Mappings from proteins to annotations, such as Genus and KEGG function, can also be provided, in which case DIAtools writes annotated peptide intensity tables.

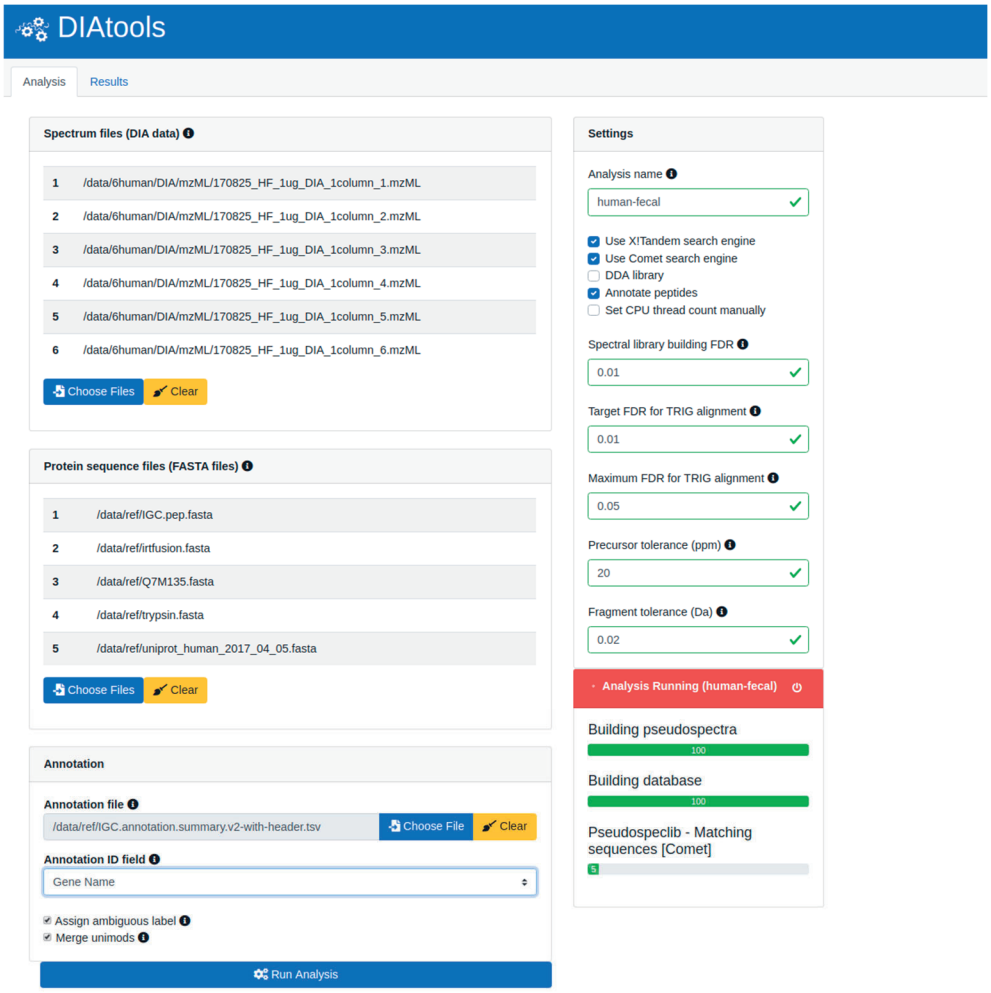

Figure 2. Analysis view of DIAtools

The analysis results (**Figure 3**) can be accessed from the result view tab. The view lists analyses and shows an overview of the results of each analysis. Full peptide intensity tables and related annotations are available for download. A log file for each run is available as well, which can be used to identify the cause of the problem if the analysis run has failed.

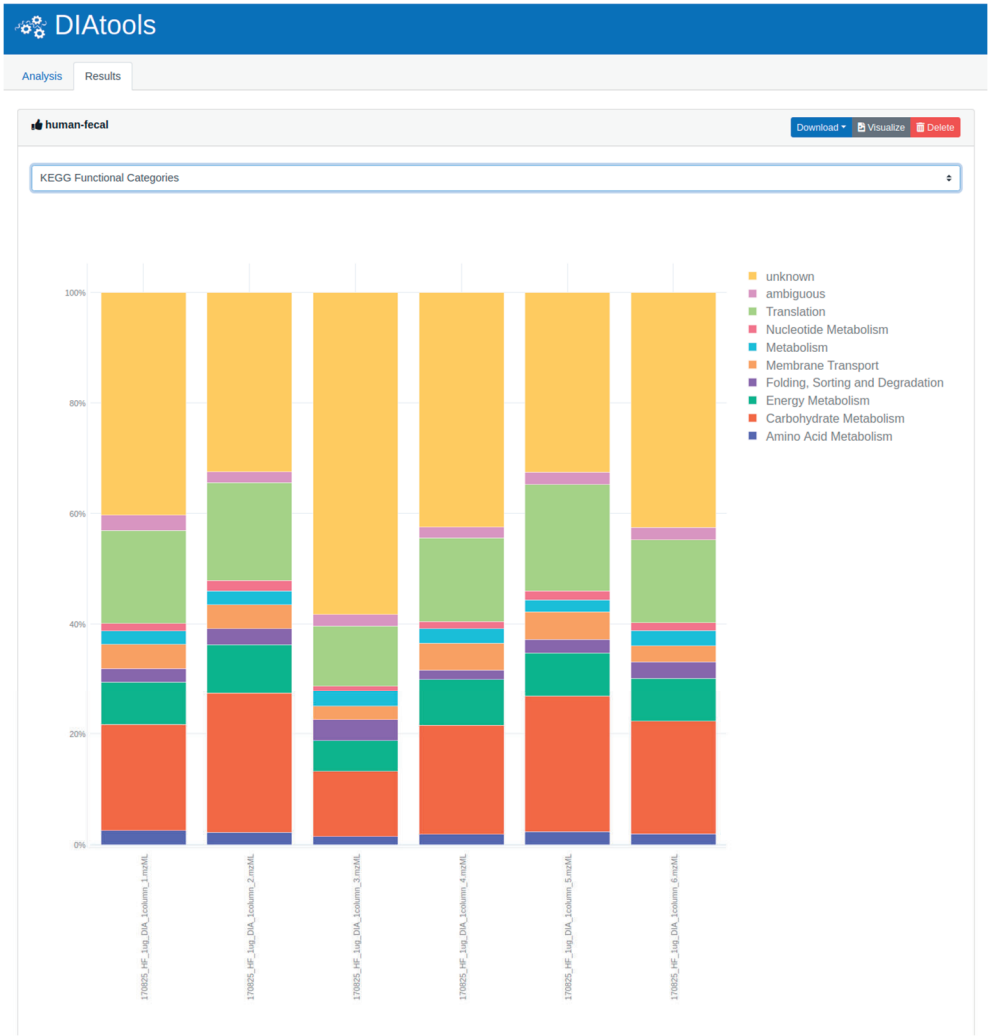

Figure 3. Result view of DIAtools

5.1.4 Command line interface

In addition to the graphical user interface, DIAtools provides a command line tool for executing the workflow with input files and settings taken in as parameters according to which the data is processed. The command line tool is especially useful in scripted works, which are typically used in high performance computer clusters where analyses are typically submitted as tasks controlled by work managers such as SLURM (https://slurm.schedmd.com/documentation.html)

5.2 Technical assessment of the DIA-method

5.2.1 Peptide identification

In total, 7967 and 15742 peptides [5] were identified from the mixture of 12 bacterial species (12mix) and 14691 and 11122 from the six fecal samples with DIA-only and DDA-assisted approaches, respectively (**Figure 4A**). With 12mix, the DIA-only identified 51% from the amount of peptides identified by the DDA-assisted while using three spectrum files in comparison to six files used by DDA-assisted. With human fecal samples, DIA-only gave 32% more identifications by using 6 spectrum files in comparison to 12 files used by DDA-assisted approach. Overall, the number of identified peptides from DIA data was comparable to those reported by DDA studies using similar laboratory protocols (Zhang *et al.*, 2018). Of all the peptides identified by either of the DIA approaches, 37% (12mix) and 30% (human fecal) were identified by both of the methods (**Figure 4B, C**).

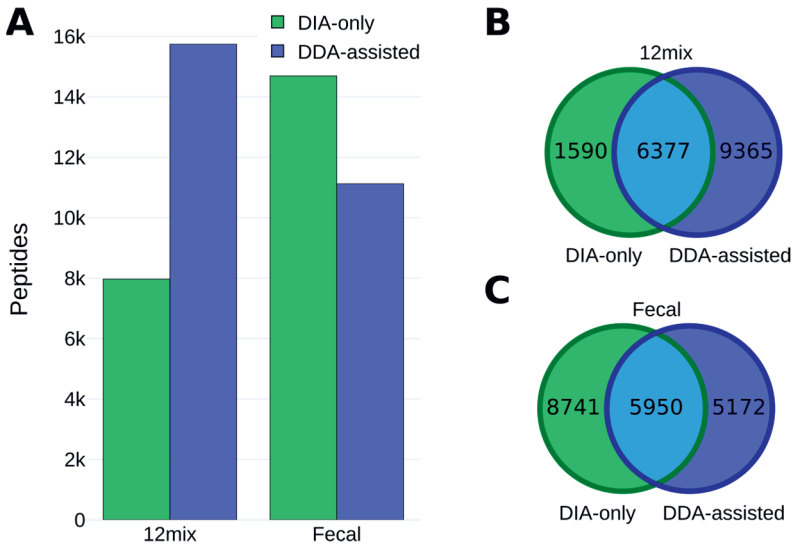

Figure 4. (**A**) Peptide identifications with two DIA-approaches (DIA-only and DDA-assisted DIA) in 12mix and fecal datasets. Common peptides identified by the methods in **(B)** 12mix and **(C)** fecal datasets.

[5] Here, in the context of mass spectrometry based identifications, peptides are used to refer to peptide ions with possible modifications.

5.2.2 Taxonomic and functional annotations

Over 56% of 12mix and over 41% of fecal data peptides were unambiguously assigned into a single genus and over 89% and 87%, respectively, of the peptides were assigned to a single KEGG functional category (**Table 1**).

Table 1. The taxonomically and functionally annotated peptides out of all identified peptides in 12mix and fecal datasets with DIA-only and DDA-assisted methods.

	Genus (%)		KEGG functional categories (%)	
	12mix	**Fecal**	**12mix**	**Fecal**
DIA-only	56	42	89	87
DDA-assisted	61	41	89	89

The number of identified peptides by taxonomy and function are shown in **Figure 5A-D**. In the 12mix, only 2% of peptides were incorrectly annotated to genera not included in the mixture. The taxonomic profile of human fecal samples was similar to those reported in the existing literature (Tanca *et al.*, 2017). However, the definition of similarity is loose and difficult to quantify. Here, it was observed that 80% of the reported most abundant identifications (Tanca et al., 2017, Figure 2B right side panel) were among those genera reported in this thesis study (**Figure 5B**). It should be noted that the differences may originate from biological and technical (analysis method) differences.

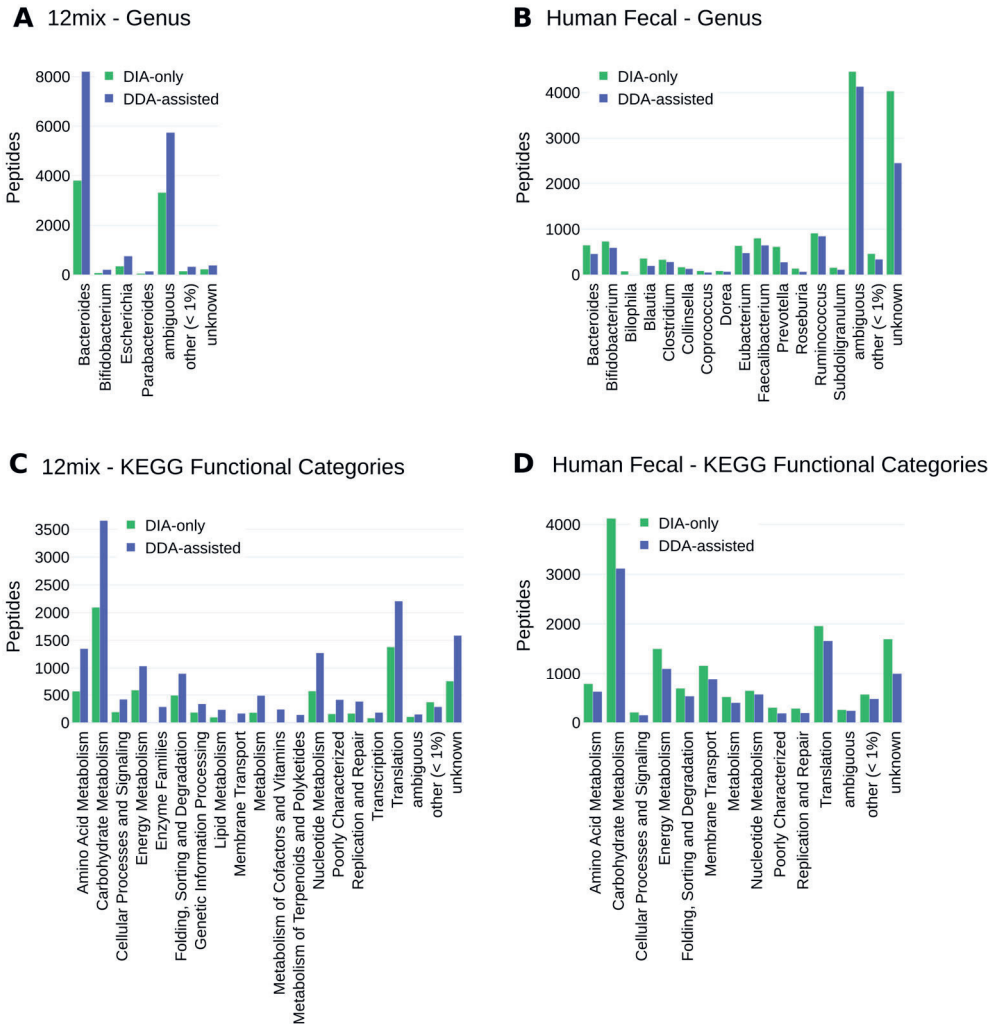

Figure 5. Genus level taxonomic profiles of the (**A**) 12mix and (**B**) human fecal samples using the DIA-only or the DDA-assisted approach and, respectively, (**C,D**) KEGG functional categories. Genera and KEGG function categories having less than 1% of the total peptides were aggregated to category *other*, as such small proportions fall below false discovery rate (FDR) thresholds used in the peptide identification process.

5.2.3 Reproducibility

Reproducibility of identifications, i.e. the proportion of common peptides to all peptides identified by repeated analyses, was measured with three replicated samples of 12mix. High reproducibility was observed by more than 97% of peptides being identified in all three technical replicates with the DIA approaches **(Figure 6A-B)**, while reproducibility of the corresponding DDA data was only 41%.

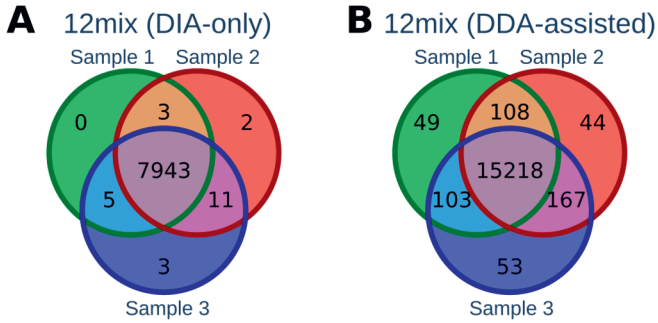

Figure 6. Intersection and difference of peptides from repeated analysis of a 12mix sample with **(A)** DIA-only, (**B**) DDA-assisted methods.

Reproducibility of quantifications was assessed by calculating Pearson correlation coefficients of the peptide intensities between each pair of technical replicates in the 12mix. The coefficients were very high ($r > 0.95$ with $p < 0.001$ in each pairwise comparison with both DIA-only and DDA-assisted approaches, **Figure 7**, indicating high reproducibility of the quantifications.

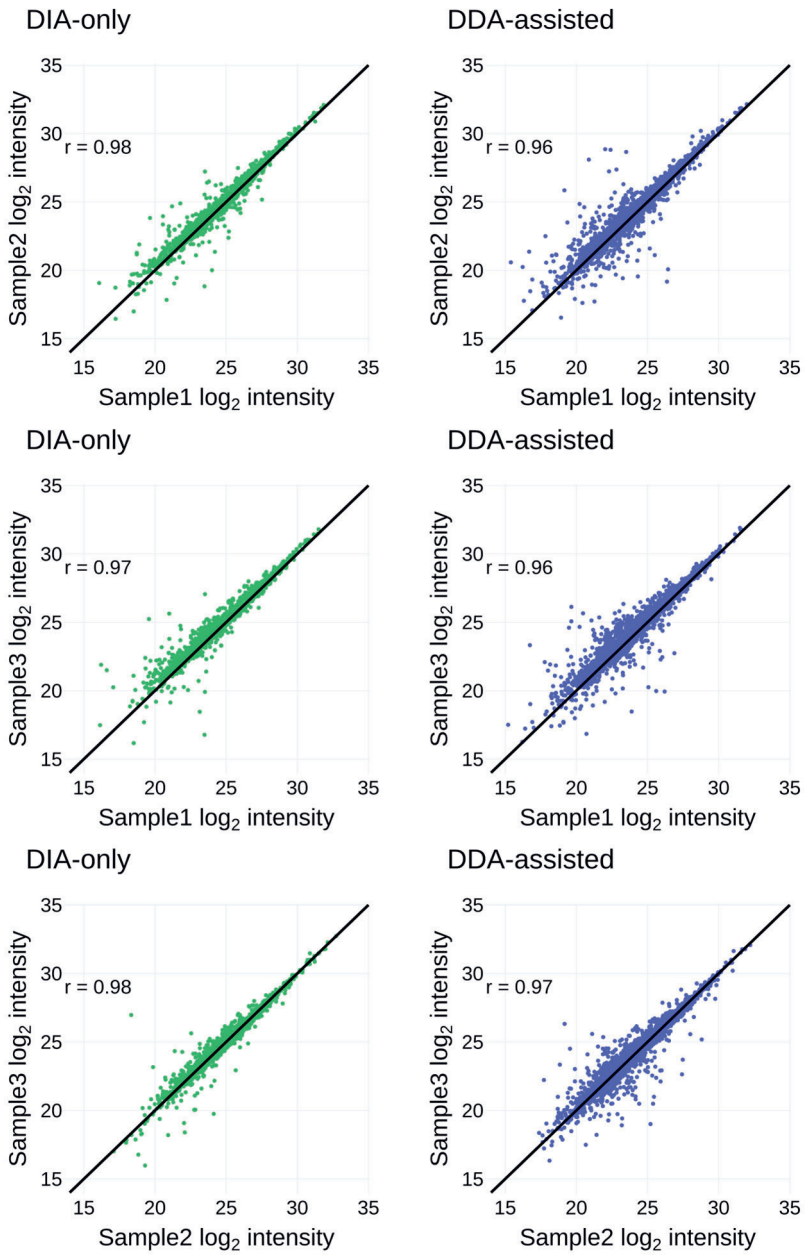

Figure 7. Pairwise correlations, and Pearson's correlation coefficients r, of the quantified peptide intensities between technical replicates of the 12mix samples using the DIA-only (green) or the DDA-assisted (blue) approach.

5.2.4 Quantification consistency between approaches

The peptide intensity consistency between the DIA-only and DDA-assisted approaches is high, but nevertheless the intensities are not identical. This is shown in both 12mix and human fecal data, **Figure 8**. Peptide quantification is defined by the set of fragment ions used to describe a peptide in the library, thus quantification difference originate in spectral and pseudospectral library composition differences.

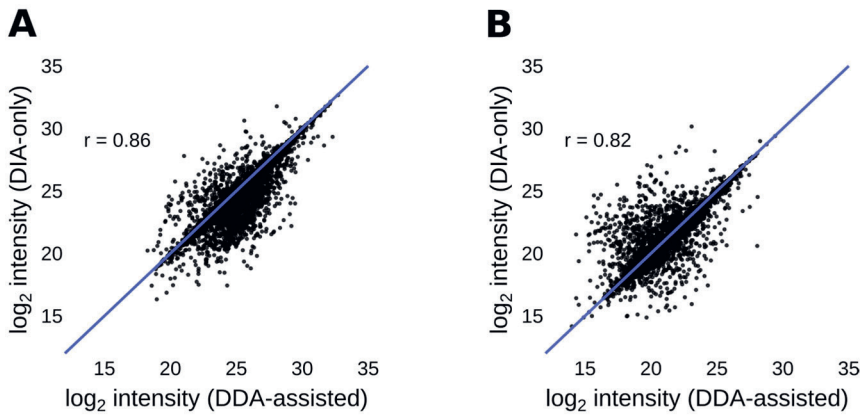

Figure 8. Correlations, and Pearson's correlation coefficients r, of the quantified peptide intensities between DIA-only or the DDA-assisted approach in A) 12mix and B) human fecal data.

5.2.5 Peptide prevalence

In the peptide intensity matrix, which contains all peptide intensities for each sample, an undetected peptide in a sample is marked with a zero-intensity value. The low complexity 12mix intensity matrix had 0.1% and 1.4 % of zero intensities with DIA-only and DDA-assisted approaches, respectively, while the human fecal matrix had 44.1% and 36.5% correspondingly.

5.3 Metaproteomic analysis of human gut microbiota

In **Publication IV** the DDA-assisted DIA metaproteomics approach was applied to profile fecal microbiota in 63 healthy adult individuals, the focus being on microbial CAZy (carbohydrate-active) enzymes involved in glycan foraging. The peptides were identified and quantified with the DIA analysis method and workflow presented in **Publications I and II**. The carbohydrate-active enzymes were identified against the IGC database and were annotated with CAZyme annotations assigned to the IGC proteins.

5.3.1 Peptide identifications and taxonomy

In total, 56571 peptides were identified with per sample identifications ranging from 5415 to 17904 peptides, the median being 11868 peptides. The proportion of zero values in the peptide intensity matrix was 82%, with sample mean percentage 79 and standard deviation 7.1, indicating heterogeneity in the sets of peptides between the samples.

An unambiguous genus level taxonomy annotation could be assigned to 45% of the peptides, which presented 10 genera with over 1% abundance, **Figure 9A**. Overall, *Bacteroides* and *Faecalibacterium* were the most prevalent genera in the gut metaproteome. Notably, a few donors had a distinguishable high amount of *Prevotella* (**Figure 9B**).

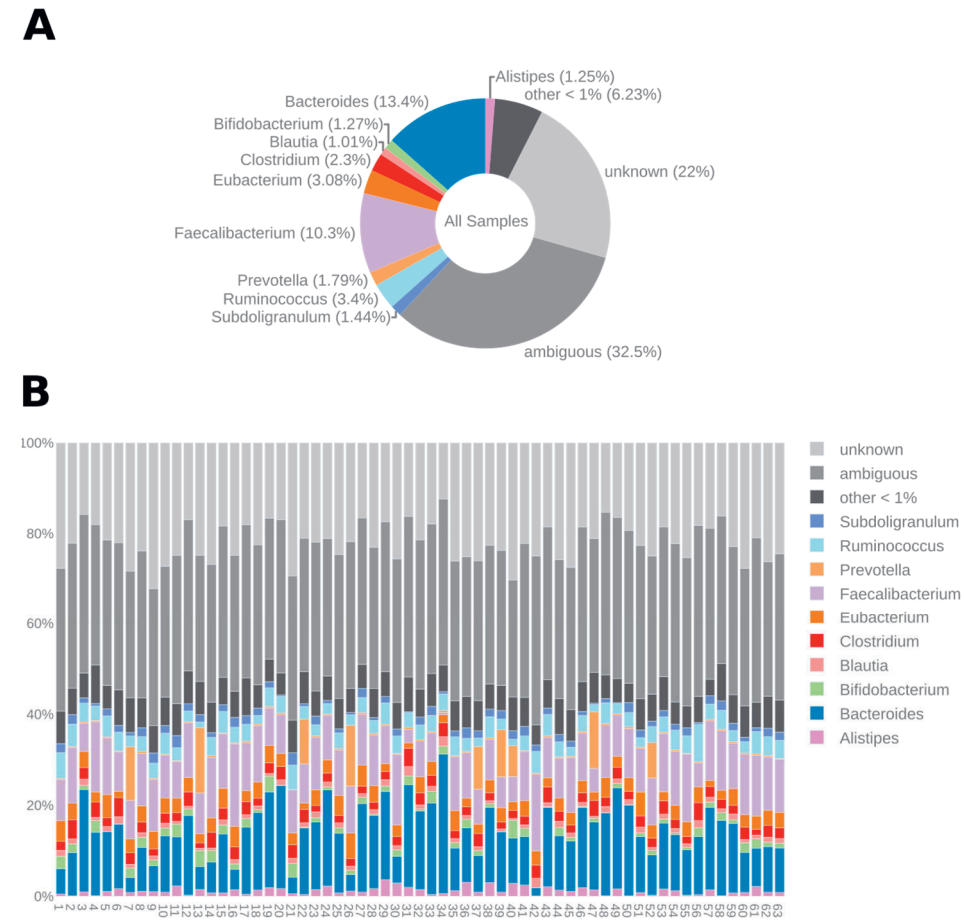

Figure 9. The peptides from the 63 fecal samples annotated at genus level shown **(A)** dataset-wide **(B)** and per-sample. Genera having less than 1% of the total peptides were aggregated to category *other*.

5.3.2 Carbohydrate-active enzyme profiles

Out of all identified peptides, 10.1 % of peptides could be assigned to a single CAZyme family while 8.7 % could be assigned to a single protein product (enzyme) annotation. In total, 11.3 % of peptides were identified as CAZymes.

Potential CAZyme associations of the *Bacteroides*, the largest genus, were profiled by dividing the samples into high and low groups by the median of *Bacteroides* 16S rRNA gene[6] abundance and choosing statistically different genus specific CAZymes between the two groups using ROPECA with modified t-test with FDR < 0.05 (Mokkala *et al.*, 2016; Suomi and Elo, 2017). While profiling *Bacteroides*, a highly pronounced *Prevotella* associated profile emerged in a group of donors having a high amount of *Prevotella* in their microbiota **(Figure 10A)**. Additional principal component analysis (PCA) of 16S rRNA gene data indicated similarities of bacterial composition in these samples **(Figure 10B)**. The *Prevotella* associated profile contained several enzymes with predicted activities in metabolism of xylan and other complex polysaccharides derived from plants, such as GH51 alpha-l-arabinofuranosidase, GH28-family endopolygalacturonase, GH43 Beta-xylosidase and GH3 Xylan-1,4-beta-xylosidase. A *Bacteroides* profile was also detectable for a group of samples, but the profile was less pronounced. Compared with *Prevotella* CAZy profle, enzymes identified within *Bacteroides* CAZy profile suggested a different metabolic specialization. The profile included glycoside hydrolase families, GH18 and GH20, which contain enzymes active against animal glycan (El Kaoutari *et al.*, 2013). The second most adundant genus, *Faecalibacterium,* was profiled respectively, but abundant *Faecalibacterium* samples did not present a distinctive CAZyme profile. The energy intake of various dietary components was also compared as well as fruit or vegetable consumption, but no differences were found to explain the *Prevotella*-CAZy profile in the particular individuals. However, the whole grain intake was lower in individuals with *Prevotella*-CAZy profile.

[6] The 16S rRNA gene abundance data was obtained from a previously conducted study (Mokkala *et al.*, 2016).

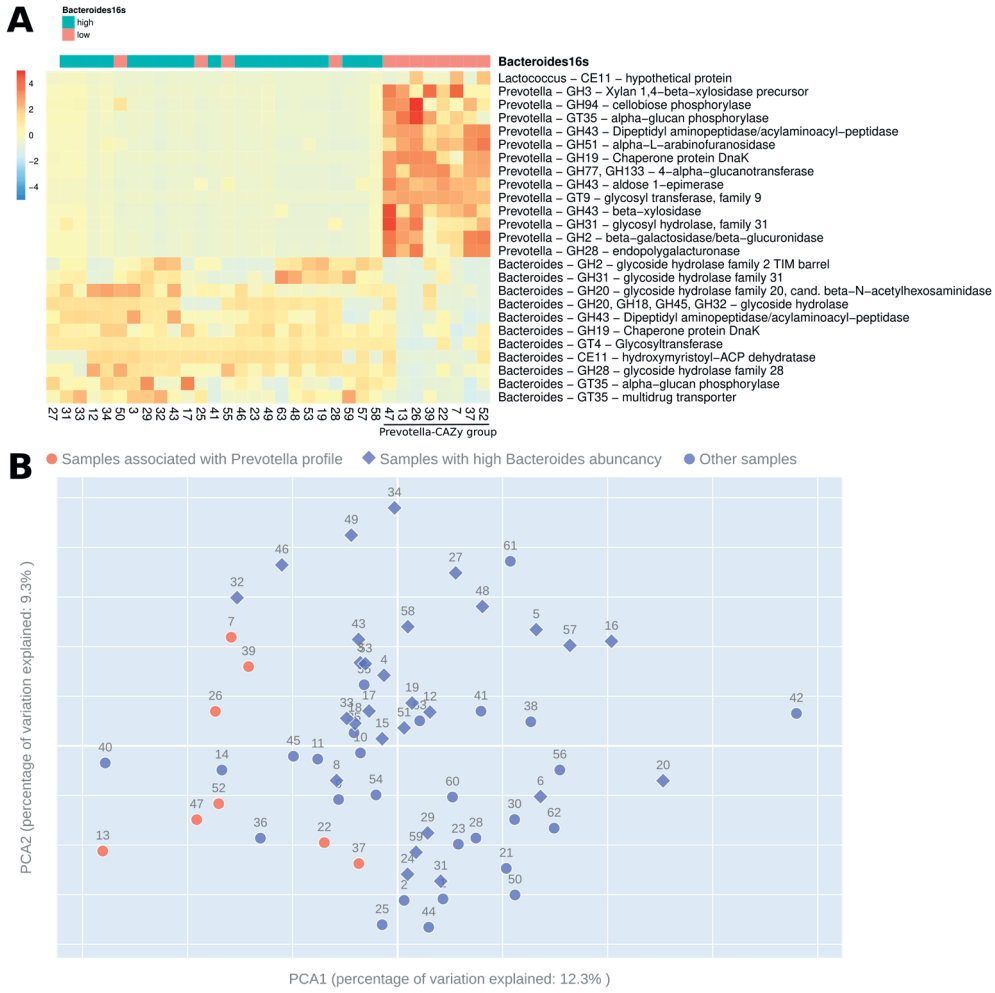

Figure 10. (**A**) A heatmap of CAZyme intensities showing *Bacteroides*- and *Prevotella* –annotated CAZyme among study subjects, identifying a group of subjects with *Prevotella*-annotated CAZyme profile. **(B)** A PCA analysis of microbiota profiles based on CLR-transformed 16S rRNA gene sequencing data with red indicating samples from **(A)** having *Prevotella* associated profile. Samples having high (above the median) abundancy of *Bacteroides* are marked with blue zalmiak.

6 Discussion

6.1 Background, novelty and importance

In recent years, human gut microbiota profiling has attracted attention with microbes being associated with human health not only by pathogenesis but by carrying out vital functions in the human body possibly governed by crosstalk between the microbes and the host body. Tiny molecules carry out functions and mediate interactions. The number and variety of such particles are staggering. Analysis of these interacting molecules requires sophisticated methods and instruments. New methods and approaches are needed to study the particles in complex and thus challenging scenarios.

At the time this study was conceived, mass spectrometer data-dependent acquisition mode (DDA) was the most commonly used method for metaproteome analysis. Data-independent acquisition mode (DIA) had not been discussed in the literature in the context of metaproteomics. The feasibility was a concern as DIA produces highly convoluted spectra in metaproteomic scenarios where there are increased amounts of co-eluting peptides. However, DIA was seen as an appealing technique as it had been found to achieve higher reproducibility and offer superior quantification over DDA in simpler proteomics scenarios (Barkovits *et al.*, 2020). To explore and advance the field of DIA metaproteomics, **Publication I** showed that DIA is a feasible analysis method for complex metaproteomic data and **Publication II** presented a workflow for the analysis.

The performance difference between DDA and DIA is commonly addressed in the literature to the data acquisition mode, but it is also reported to be caused by the identification strategy (Fernández-Costa *et al.*, 2020) as DIA identification is typically spectral library based while DDA is database sequence based. Using both analysis concepts and DIA data, **Publication III** introduced a pseudospectral library approach in combination with sequence database search for metaproteomic DIA data analysis. The results obtained with these methods substantiated that the analysis was feasible and equivalent identification rate, to the rate of DDA, was reached while achieving superior reproducibility and thus reducing the requirement from multiple DDA analyses per sample to only a single DIA analysis per a sample. This work

presented the first-time analysis of metaproteomic DIA data with a description of an applicable analysis method, which is a crucial step for the further development of any line of new analyses by pointing out a solid starting point.

In **Publication IV**, for the first time, carbohydrate-active enzymes, expressed by microbial taxa, were profiled with the DIA method to study the associations of bacterial taxonomic abundances and enzymes.

6.2 Results and implications

The results suggest that DIA is a compelling alternative for DDA in metaproteomics as it achieves peptide identification rates comparable to those reported with DDA analysis (**Figure 4A**) (Kolmeder *et al.*, 2016), (Tanca *et al.*, 2017) while offering very high reproducibility (**Figure 6A,B**) and support for quantification. The improved reproducibility offered by the DIA method provides benefits for studies. The observed differences in time series samples from the same individual are more likely of biological than technical origin. Likewise, high technical overlap in profiles (e.g. healthy vs disease) improves the possibility to find meaningful biological differences.

The presence of protein (functional) homology is apparent at peptide level where the peptides are commonly shared between species of different genera, causing a large proportion of peptides to have multiple taxonomic annotations which renders the peptide annotation ambiguous at genus level (**Figure 5A,B**). For instance, in the diverse fecal samples only 41 - 42% peptides could be annotated (**Table 1**). In the less diverse bacterial mixture, 56 - 61% could be annotated (**Table 1**). These results indicate that the mass spectrometry methods, drawing properties from the peptide level data, have diminished applicability and resolution for taxonomic identification in comparison to the sequencing methods. This is understandable as peptides are very short and their analysis is primarily intended for functional analysis. The amount of unambiguous peptides could have been reduced with lowest common ancestor (LCA) approach, which would solve multiple different annotation problem by climbing up in the taxonomic levels until a consistent annotation is reached, but that would have lead into a dataset where peptides do not have the same level taxonomic annotation (Aho, Hopcroft and Ullman, 1973).

However interestingly, while peptides are commonly shared between different taxa, they are not commonly shared between different functions (**Figure 5A-D**) and proteins sharing a peptide is an indicative of the proteins likely perform the same function as well. Importantly, the functional annotation could be assigned to 87 - 89% of the fecal dataset peptides (**Table 1**). Reduced sample complexity did not seem to improve radically the functional annotation rate, like it did with taxonomic

annotations. The direct annotation assignment strategy was chosen due to the inherent property of the data for the majority of peptides mapping uniquely to a function. There is a strategy developed for metaproteomics, where peptides are used to determine metaproteins, which are groups of proteins defined by various rules such as by shared peptides and by consistent taxonomic annotation (Muth *et al.*, 2015). From an annotation point of view, these requirements appear to have similarities to our direct peptide annotation strategy, which demands unique annotation for each peptide.

Only 2% of peptides were annotated to genera not present in the bacterial mixture and the taxonomic composition of the fecal metaproteome were similar to those reported by others (Kolmeder *et al.*, 2012, 2016; Tanca *et al.*, 2017), which indicated correctness of the peptide identifications and annotation assignments. Interestingly, the distribution of peptide counts by genera differs from the taxonomic composition of the mixture possibly indicating differing levels of activity among species.

Overall, the technical evaluation of the method indicated that a metaproteomic DIA data analysis pipeline can be constructed building on existing proteomics methods, such as sequence database (Eng, Martin and Aebersold, 2005) and spectral library (Lam *et al.*, 2008) methods, as components. The results suggest the potential of unfiltered mass spectrometry data analysis where all ions are recorded and analyzed. The ability to analyze such data and make identifications in highly convoluted metaproteomic scenarios challenges the prevailing concept according to which it is necessary to produce and use incomplete records though ion selection procedures.

The metaproteomic analysis of the 63 healthy human donors identified 11.3% of peptides being CAZyme originated and out of those 78% were assigned with a single enzyme annotation, while the rest assigned with multiple annotations which rendered the annotation ambiguous. The enzyme differences of the largest genus, *Bacteroides*, were profiled by dividing the samples into two groups around *Bacteroides* 16S rRNA gene median abundance. *Bacteroides* associated enzymes were found to target animal originated molecules. While profiling *Bacteroides*, *Prevotella* associated enzyme profile became visible showing enzymes targeting plant molecules. 16S rRNA gene analysis confirmed that the samples showing *Prevotella* enzyme profile (**Figure 10A**) had a distinguishable high abundance of *Prevotella*. This showed an association between the identified enzyme profile and the taxonomic abundance of *Prevotella* in the gut microbiota. This type of profiling may help to predict personalized responses to dietary interventions based on the CAZyme profiles, which may reveal if the particular gut microbiota is able to utilize a certain type of nutrients (i.e. fiber). Importantly, the results highlight the applicability of the DIA method for biological research with potential to reveal important insights into human

health. However, it should be kept in mind that the method does not exclude the possibility of the results having included also a small number of false identifications.

6.3 Bioinformatics software and DIAtools

We are living in an era where computing is becoming embedded with every aspect of biological research. The current high throughput laboratory instruments produce vast amounts of data thus creating a demand for high performance computing and bioinformatics. This implies that computer software and algorithms have an increasing role in the research. While proper practises for software development are well established in the computer software industry, it is much less so in the field of academic research. Developing software in academic projects is challenging. Typically, academic projects do not have access to resources like software companies have to manage the various aspects of software development life-cycle, such as specialized teams for software development, testing, technical documentation and support. Most importantly, academic projects are highly constrained to a specific funding period, which is typically very short from the perspective of software development. Also the funding is strongly geared towards making new discoveries and thus rendering all other concerns, such as software quality, secondary. However, the funding and academic research are highly compatible with open source ideology. In fact, the public funding favours the idea that results, including the software, are free and open for all. Once a program is under a free and open source license, which permits anyone to modify and redistribute the program, the community can step in and contribute to the development even if the original funding is discontinued. This is a huge benefit over closed proprietary programs. In this regard, the companies developing laboratory instruments should be encouraged to be more open. For instance, mass spectrometers produce spectrum files in vendor specific proprietary raw formats and a proprietary software library is typically needed to read the files. One such library is Thermo Scientific raw file conversion library, which poses restrictions hindering its distribution as a part of any open source software.

At the time this study was started, no DIA metaproteomic studies were published and it became obvious that building a software for the analysis is a crucial step in the process for the discovery of feasible analysis methods. DIAtools was written for this purpose. The initial version had only a command line interface, but later the importance of easy usability was recognized and DIAtools was given a modern web based user interface. An open source licensing was chosen as it was determined to be the best model for an academic project. The DIA analysis workflow is rather long and has multiple steps implemented by various utilities. DIAtools is implemented

as a container which provides all the required utilities and libraries in a single package, enabling it to offer reproducible analyses even after long periods of time. DIAtools also supports the latest rootless container technology, Podman, which enables users to install and operate a container as a normal user without needing privileged user access.

6.4 Limitations and future research

The peptide identification and quantification performance of the presented DIA analysis method, including reproducibility, is likely generalizable to gut microbiota datasets produced with similar laboratory protocols and instruments. More research is needed to study the method with different laboratory protocols, instruments and types of datasets. The method is not inherently limited to the analysis of gut microbiota and can be applied in various research fields.

More research is needed to improve the identification rate and the consistency of the DIA-only approach as it should ideally outperform DDA-assisted approach in all scenarios.

The comparison of taxonomic identifications from 16S rRNA gene data and metaproteomic data was not carried out in this research. Such a comparison is recommended for future studies.

Counterintuitively, more peptides were identified from the 12mix in comparison to more complex human fecal samples (**Figure 4A**). Presumably, the reasons, which hinder the identification from the more complex samples, can originate from sample preparation to data-analysis phase. However, the datasets were not designed to be comparable in this manner and there are differences in the processing protocols between the datasets which renders the comparison unpractical.

The applicability of the method for the analysis of more diverse communities, such as soil samples, might be hindered by the usage of a single dataset-wide library. In certain cases it might be beneficial to switch to per sample libraries, which are more specialized. However, it is not known how much this would reduce the comparability between the samples as libraries are likely to have different sets of peptides and even the corresponding peptides can potentially have different sets of fragments in the library that are used for identification and quantification.

Another approach for increasing the peptide yield is to target the library more precisely by performing the search step twice. The initial search is done without FDR and all identified peptides are used as the sequence search space for the second search where a FDR threshold is applied. This approach is called MetaPro-IQ (Zhang *et al.*, 2016). Overall, technical improvements of the methods are needed to increase the number of identifications, which would help to push forward the field of

metaproteomics as the number of identified peptides with current methods is much lower than the diversity suggests (Yang *et al.*, 2009; Tierney *et al.*, 2019). Increasing the identification performance might be obtainable by replacing the equal intensities of theoretical spectra with predicted intensities or by building per-sample libraries instead of dataset-wide libraries and by even determining the included fragments for a library in a more sophisticated manner than using a fixed number of fragments.

Overall, the FDR calculation strategies should be further studied in the context of metaproteomic data. This topic was not thoroughly explored, instead this work relied on existing practices and protocols.

In future research, the necessity of a pseudospectral library as such should be evaluated in the presented workflow and possibly shifted away from the paradigm where peptides are identified and quantified from the DIA-data using a library. Instead, more direct approach would first identify and quantify candidate peptide spectra from DIA-data and subsequently assign amino acid sequences for the identified spectra. This would remove the need to process through the DIA-spectra twice.

When this study was conceived and the technical datasets of this study were designed, the comparison of the DIA and DDA was not considered as an aim of this study. For this reason, the samples for such a direct comparison are somewhat lacking mainly because the technical datasets have DDA data only from the pooled samples. This is a limitation of the available technical datasets that were designed only for the validation of the DIA analysis method. Overall, there is a demand for benchmark datasets to better compare DDA and DIA in complex metaproteomic scenarios.

Peptide identification has been the focus in this work. Further research is needed on quantification as it is generally considered as a strength of DIA analysis. There is especially demand for benchmark datasets. Currently, assessing the quantification accuracy is hindered by a lack of benchmark datasets.

The peptide identification methods (Eng, McCormack and Yates, 1994; Perkins *et al.*, 1999; Craig and Beavis, 2004) deployed in this work are computationally intensive with large metaproteomic sequence databases where the amount of comparison of empirical to theoretical spectra can be enormous. Calculation time can be speeded up by parallel comparisons. Currently applied tools use Central Processing Unit (CPU) threading technique which enables parallel processing hundreds of spectra, while switching to Graphics Processing Unit (GPU) accelerated processing (Y. Li *et al.*, 2014) would enable parallel processing tens of thousands of spectra.

7 Summary/Conclusions

Human gut microbiota is a diverse community of microorganisms having complex interactions with each other and with the human host. Understanding the functions of the microorganisms is essential for gaining understanding on their role in human health. Novel methods for metaproteomics are needed for the discovery of the functions the microbes are carrying out. The aim of this work was to study the feasibility of DIA data analysis for analyzing complex metaproteomes, and to demonstrate its applicability for analysis of human gut microbiota proteins and particularly, of proteins with relevance to metabolic activities of competing microbes. To facilitate this aim, first a DDA assisted DIA data analysis approach was introduced (**Publications I** and **II**). Once, the feasibility was established, the method was improved (**Publication III**) to require solely DIA data. The refined approach introduced by this work, combined pseudospectral library approach with sequence database search. The performance of the pseudospectral library, built directly from DIA data, showed potential to render the need for DDA data obsolete. The DIA method was demonstrated with a bacterial mixture and with human fecal samples. The DIA approach achieved very high reproducibility, hence the analysis can be carried out with only a single DIA analysis per sample. This gives an advantage to metaproteomic DIA data analysis over to the de facto DDA, which typically requires multiple analysis per sample. The yield of identified peptides was found to be comparable to those reported by DDA studies with similar laboratory methods (Kolmeder *et al.*, 2012, 2016; Tanca *et al.*, 2017). The taxonomic annotations of peptides corresponded to those that can be expected by the dataset design. Finally, the applicability of the DIA method for the study of human gut microbiota was shown by applying the method to study the association of Carbohydrate-active enzymes and microbial taxa abundance (**Publication IV**).

It was concluded that DIA SWATH-MS is technically a feasible method for metaproteomic data analysis providing benefits over the currently popular DDA method. This challenges the prevailing concept according to which the ion filtering of DDA is necessary. It is also concluded that the DIA method should be applied to metaproteomics studies more widely than is done currently and the method should

be further developed. This is facilitated by the DIAtools software (**Publication I,III**) being implemented as an easy to use open source software package.

Acknowledgements

This work was conducted at the Medical Bioinformatics Centre of the Turku Bioscience Centre and at the Institute of Biomedicine of the Faculty of Medicine, University of Turku.

I am deeply grateful to my supervisors Prof. Laura Elo and Adj. Prof. Arno Hänninen, who not only have provided crucial help and guidance throughout the whole process from doing academic research to publishing the results, but also have made the whole study possible by funding it.

I wish to express my gratitude for all the co-authors: Dr. Juhani Aakko, Dr. Tomi Suomi, MSc. Mehrad Mahmoudian, Dr. Anne Rokka, Dr. Raine Toivonen, Dr. Petri Kouvonen, Dr. Kati Mokkala and Assoc. Prof. Kirsi Laitinen.

I also want to acknowledge the official reviewers Adj. Prof. Reetta Satokari and Asst. Prof. Jarkko Salojärvi for their constructive and valuable comments.

I want to thank Assoc. Prof. Tapio Pahikkala and Prof. Erkki Eerola for their support as follow-up committee members. Special thanks for Eeva Valve (doctoral programme coordinator) and Outi Irjala (chief academic officer), who have been there to answer to all my questions throughout the whole thesis process.

Last but not least, I am very grateful to my family for providing supportive environment for doing academic work and for especially understanding for the long working days it sometimes requires.

7.12.2021
Sami Pietilä

References

Aggarwal, S. and Yadav, A. K. (2016) "False Discovery Rate Estimation in Proteomics," *Methods in Molecular Biology*, pp. 119–128. doi: 10.1007/978-1-4939-3106-4_7.

Aho, A. V., Hopcroft, J. E. and Ullman, J. D. (1973) "On finding lowest common ancestors in trees," *Proceedings of the fifth annual ACM symposium on Theory of computing - STOC '73*. doi: 10.1145/800125.804056.

Aitchison, J. (1982) "The statistical analysis of compositional data," *Journal of the Royal Statistical Society*. Wiley, 44(2), pp. 139–160.

Alberts, B. (2017) *Molecular Biology of the Cell*. Garland Science.

Armengaud, J. *et al.* (2021) "Taxonomical and functional changes in COVID-19 faecal microbiome are related to SARS-CoV-2 faecal load." doi: 10.21203/rs.3.rs-414136/v1.

Asnicar, F. *et al.* (2020) "Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0," *Nature communications*. Springer Science and Business Media LLC, 11(1), p. 2500.

Barkovits, K. *et al.* (2020) "Reproducibility, Specificity and Accuracy of Relative Quantification Using Spectral Library-based Data-independent Acquisition," *Molecular & cellular proteomics: MCP*, 19(1), pp. 181–197.

Bashiardes, S., Zilberman-Schapira, G. and Elinav, E. (2016) "Use of Metatranscriptomics in Microbiome Research," *Bioinformatics and biology insights*, 10, pp. 19–25.

Bäumler, A. J. and Sperandio, V. (2016) "Interactions between the microbiota and pathogenic bacteria in the gut," *Nature*, 535(7610), pp. 85–93.

Bergström, A. *et al.* (2014) "Establishment of intestinal microbiota during early life: a longitudinal, explorative study of a large cohort of Danish infants," *Applied and*

environmental microbiology. American Society for Microbiology, 80(9), pp. 2889–2900.

Bhattacharya, T., Ghosh, T. S. and Mande, S. S. (2015) "Global profiling of carbohydrate active enzymes in human gut microbiome," *PloS one*. Public Library of Science (PLoS), 10(11), p. e0142038.

Blattmann, P., Heusel, M. and Aebersold, R. (2016) "SWATH2stats: An R/Bioconductor Package to Process and Convert Quantitative SWATH-MS Proteomics Data for Downstream Analysis Tools," *PLOS ONE*, p. e0153160. doi: 10.1371/journal.pone.0153160.

Castellarin, M. *et al.* (2012) "Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma," *Genome research*, 22(2), pp. 299–306.

Clooney, A. G. *et al.* (2016) "Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis," *PloS one*, 11(2), p. e0148028.

Cole, J. R. *et al.* (2014) "Ribosomal Database Project: data and tools for high throughput rRNA analysis," *Nucleic acids research*. Oxford University Press (OUP), 42(Database issue), pp. D633-42.

Cordero, O. X. *et al.* (2012) "Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance," *Science*, 337(6099), pp. 1228–1231.

Craig, R. and Beavis, R. C. (2004) "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics* , 20(9), pp. 1466–1467.

Crick, F. H. (1958a) *On Protein Synthesis"*. Edited by F. K. Sanders. Number XII: The Biological Replication of Macromolecules. Cambridge University Press, pp. 138–163.

Crick, F. H. (1958b) "On protein synthesis," *Symposia of the Society for Experimental Biology*, 12, pp. 138–163.

De, R. (2017) "Metagenomics: The revolution in the biomedical world," *British journal of research*. Scitechnol Biosoft Pvt. Ltd. doi: 10.21767/2394-3718.100035.

Dethlefsen, L. and Relman, D. A. (2011) "Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation," *Proceedings of the National Academy of Sciences of the United States of America*. Proceedings of the National Academy of Sciences, 108 Suppl 1(Supplement_1), pp. 4554–4561.

Deutsch, E. W. *et al.* (2010) "A guided tour of the Trans-Proteomic Pipeline," *Proteomics*, 10(6), pp. 1150–1159.

Deutsch, E. W. *et al.* (2017) “Proteomics Standards Initiative: Fifteen Years of Progress and Future Work,” *Journal of proteome research*, 16(12), pp. 4288–4298.

Dobin, A. *et al.* (2013) “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics* . Narnia, 29(1), pp. 15–21.

El Kaoutari, A. *et al.* (2013) “The abundance and variety of carbohydrate-active enzymes in the human gut microbiota,” *Nature reviews. Microbiology*. Springer Science and Business Media LLC, 11(7), pp. 497–504.

Elias, J. E. and Gygi, S. P. (2007) “Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry,” *Nature Methods*, pp. 207–214. doi: 10.1038/nmeth1019.

Elias, J. E. and Gygi, S. P. (2010) “Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics,” *Methods in Molecular Biology*, pp. 55–71. doi: 10.1007/978-1-60761-444-9_5.

Eng, J. K., Jahan, T. A. and Hoopmann, M. R. (2013) “Comet: an open-source MS/MS sequence database search tool,” *Proteomics*, 13(1), pp. 22–24.

Eng, J. K., Martin, D. B. and Aebersold, R. (2005) “Tandem mass spectrometry database searching,” *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. doi: 10.1002/047001153x.g301204.

Eng, J. K., McCormack, A. L. and Yates, J. R. (1994) “An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database,” *Journal of the American Society for Mass Spectrometry*, pp. 976–989. doi: 10.1016/1044-0305(94)80016-2.

Escher, C. *et al.* (2012) “Using iRT, a normalized retention time for more targeted measurement of peptides,” *Proteomics*, 12(8), pp. 1111–1121.

Fernández-Costa, C. *et al.* (2020) “Impact of the Identification Strategy on the Reproducibility of the DDA and DIA Results,” *Journal of proteome research*, 19(8), pp. 3153–3161.

Francino, M. P. (2015) “Antibiotics and the human gut microbiome: Dysbioses and accumulation of resistances,” *Frontiers in microbiology*. Frontiers Media SA, 6, p. 1543.

Franzosa, E. A. *et al.* (2018) “Species-level functional profiling of metagenomes and metatranscriptomes,” *Nature methods*, 15(11), pp. 962–968.

Garcia-Perez, I. *et al.* (2020) “Dietary metabotype modelling predicts individual responses to dietary interventions,” *Nature Food*. Springer Science and Business Media LLC, 1(6), pp. 355–364.

Gensollen, T. *et al.* (2016) “How colonization by microbiota in early life shapes the immune system,” *Science*, 352(6285), pp. 539–544.

Gessulat, S. *et al.* (2019) “Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning,” *Nature methods*. Nature Publishing Group, 16(6), pp. 509–518.

Han, X., Aslanian, A. and Yates, J. R., 3rd (2008) “Mass spectrometry for proteomics,” *Current opinion in chemical biology*, 12(5), pp. 483–490.

Homann, C.-M. *et al.* (2021) “Infants’ first solid foods: Impact on gut Microbiota development in two intercontinental cohorts,” *Nutrients*. MDPI AG, 13(8), p. 2639.

Hu, A., Noble, W. S. and Wolf-Yadlin, A. (2016) “Technical advances in proteomics: new developments in data-independent acquisition,” *F1000Research*, 5. doi: 10.12688/f1000research.7042.1.

Hu, Q. *et al.* (2005) “The Orbitrap: a new mass spectrometer,” *Journal of mass spectrometry: JMS*, 40(4), pp. 430–443.

Huang, T. *et al.* (2012) “Protein inference: a review,” *Briefings in bioinformatics*. Oxford Academic, 13(5), pp. 586–614.

Inman, M. (2011) “How bacteria turn fiber into food,” *PLoS biology*, p. e1001227.

Integrative HMP (iHMP) Research Network Consortium (2019) “The Integrative Human Microbiome Project,” *Nature*. Springer Science and Business Media LLC, 569(7758), pp. 641–648.

Johnson, C. H., Ivanisevic, J. and Siuzdak, G. (2016) “Metabolomics: beyond biomarkers and towards mechanisms,” *Nature reviews. Molecular cell biology*, 17(7), pp. 451–459.

Johnson, E. L. *et al.* (2017) “Microbiome and metabolic disease: revisiting the bacterial phylum Bacteroidetes,” *Journal of molecular medicine*. Springer Science and Business Media LLC, 95(1), pp. 1–8.

Johnson, J. S. *et al.* (2019) “Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis,” *Nature communications*. Nature Publishing Group, 10(1), pp. 1–11.

Jones, A. R. *et al.* (2009) “Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines,” *Proteomics*, 9(5), pp. 1220–1229.

Keller, A. *et al.* (2002) “Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search,” *Analytical chemistry*, 74(20), pp. 5383–5392.

Kho, Z. Y. and Lal, S. K. (2018) "The Human Gut Microbiome – A Potential Controller of Wellness and Disease," *Frontiers in Microbiology*. doi: 10.3389/fmicb.2018.01835.

Kleiner, M. (2019) "Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities," *mSystems*, 4(3). doi: 10.1128/mSystems.00115-19.

Kolmeder, C. A. *et al.* (2012) "Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions," *PloS one*, 7(1), p. e29913.

Kolmeder, C. A. *et al.* (2016) "Faecal Metaproteomic Analysis Reveals a Personalized and Stable Functional Microbiome and Limited Effects of a Probiotic Intervention in Adults," *PloS one*, 11(4), p. e0153294.

Kolmeder, C. A. and de Vos, W. M. (2014) "Metaproteomics of our microbiome - developing insight in function and activity in man and model systems," *Journal of proteomics*. Elsevier BV, 97, pp. 3–16.

Krastanov, A. (2010) "Metabolomics—The State of Art," *Biotechnology & Biotechnological Equipment*, pp. 1537–1543. doi: 10.2478/v10133-010-0001-y.

Kumar, D., Yadav, A. K. and Dash, D. (2017) "Choosing an Optimal Database for Protein Identification from Tandem Mass Spectrometry Data," *Methods in molecular biology* , 1549, pp. 17–29.

Kumar, S., Filipski, A. and Greenbaum, D. (2004) "Maximum Parsimony Principle (Parsimony, Occam's Razor)," *Dictionary of Bioinformatics and Computational Biology*. doi: 10.1002/9780471650126.dob0419.pub2.

Lam, H. *et al.* (2007) "Development and validation of a spectral library searching method for peptide identification from MS/MS," *Proteomics*, 7(5), pp. 655–667.

Lam, H. *et al.* (2008) "Building consensus spectral libraries for peptide identification in proteomics," *Nature methods*. Nature Publishing Group, 5(10), pp. 873–875.

Langille, M. G. I. *et al.* (2013) "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences," *Nature biotechnology*, 31(9), pp. 814–821.

LeBlanc, J. G. *et al.* (2013) "Bacteria as vitamin suppliers to their host: a gut microbiota perspective," *Current opinion in biotechnology*, 24(2), pp. 160–168.

Ley, R. E. (2016) "Gut microbiota in 2015: Prevotella in the gut: choose carefully," *Nature reviews. Gastroenterology & hepatology*. Springer Science and Business Media LLC, 13(2), pp. 69–70.

Li, J. *et al.* (2014) “An integrated catalog of reference genes in the human gut microbiome,” *Nature biotechnology*, 32(8), pp. 834–841.

Li, Y. *et al.* (2014) “Accelerating the scoring module of mass spectrometry-based peptide identification using GPUs,” *BMC bioinformatics*, 15, p. 121.

Li, Y. *et al.* (2015) “Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files,” *Nature methods*, 12(12), pp. 1105–1106.

Linares-Pastén, J. A. *et al.* (2021) “Novel xylan degrading enzymes from polysaccharide utilizing loci of Prevotella copri DSM18205,” *Glycobiology*. Oxford University Press (OUP). doi: 10.1093/glycob/cwab056.

Lombard, V. *et al.* (2014) “The carbohydrate-active enzymes database (CAZy) in 2013,” *Nucleic Acids Research*, pp. D490–D495. doi: 10.1093/nar/gkt1178.

Long, S. *et al.* (2020) “Metaproteomics characterizes human gut microbiome function in colorectal cancer,” *npj biofilms and microbiomes*. Springer Science and Business Media LLC, 6(1), p. 14.

Ludwig, C. *et al.* (2018) “Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial,” *Molecular systems biology*, 14(8), p. e8126.

Maier, T., Güell, M. and Serrano, L. (2009) “Correlation of mRNA and protein in complex biological samples,” *FEBS Letters*, pp. 3966–3973. doi: 10.1016/j.febslet.2009.10.036.

Marcobal, A. and Sonnenburg, J. L. (2012) “Human milk oligosaccharide consumption by intestinal microbiota,” *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*. Elsevier BV, 18 Suppl 4, pp. 12–15.

Marteau, P. (2009) “Bacterial flora in inflammatory bowel disease,” *Digestive diseases* , 27 Suppl 1, pp. 99–103.

Mokkala, K. *et al.* (2016) “Gut Microbiota richness and composition and dietary intake of overweight pregnant women are related to serum zonulin concentration, a marker for intestinal permeability,” *The journal of nutrition*. Oxford University Press (OUP), 146(9), pp. 1694–1700.

Moos, W. H. *et al.* (2016) “Microbiota and neurological disorders: A gut feeling,” *BioResearch open access*. Mary Ann Liebert Inc, 5(1), pp. 137–145.

Moya, A. and Ferrer, M. (2016) “Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance,” *Trends in microbiology*, 24(5), pp. 402–413.

Muth, T. *et al.* (2015) “The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation,” *Journal of proteome research*, 14(3), pp. 1557–1565.

Muth, T. and Renard, B. Y. (2018) “Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification?,” *Briefings in bioinformatics*, 19(5), pp. 954–970.

NCBI Resource Coordinators (2018) “Database resources of the National Center for Biotechnology Information,” *Nucleic acids research*, 46(D1), pp. D8–D13.

Nesvizhskii, A. I. *et al.* (2003) “A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry,” *Analytical Chemistry*, pp. 4646–4658. doi: 10.1021/ac0341261.

Nesvizhskii, A. I. and Aebersold, R. (2005) “Interpretation of shotgun proteomic data: the protein inference problem,” *Molecular & cellular proteomics: MCP*, 4(10), pp. 1419–1440.

Palleja, A. *et al.* (2018) “Recovery of gut microbiota of healthy adults following antibiotic exposure,” *Nature microbiology*, 3(11), pp. 1255–1265.

Pérez-Cobas, A. E., Gomez-Valero, L. and Buchrieser, C. (2020) “Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses,” *Microbial genomics*. Microbiology Society, 6(8). doi: 10.1099/mgen.0.000409.

Perkins, D. N. *et al.* (1999) “Probability-based protein identification by searching sequence databases using mass spectrometry data,” *Electrophoresis*, pp. 3551–3567. doi: 10.1002/(sici)1522-2683(19991201)20:18<3551::aid-elps3551>3.0.co;2-2.

Pitt, J. J. (2009) “Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry,” *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists*, 30(1), pp. 19–34.

Precup, G. and Vodnar, D.-C. (2019) “Gut Prevotella as a possible biomarker of diet and its eubiotic versus dysbiotic roles: a comprehensive literature review,” *The British journal of nutrition*. Cambridge University Press (CUP), 122(2), pp. 131–140.

Qin, J. *et al.* (2010) “A human gut microbial gene catalogue established by metagenomic sequencing,” *Nature*, 464(7285), pp. 59–65.

Reiter, L. *et al.* (2009) “Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry,” *Molecular & cellular proteomics: MCP*, 8(11), pp. 2405–2417.

Reiter, L. *et al.* (2011) "mProphet: automated data processing and statistical validation for large-scale SRM experiments," *Nature Methods*, pp. 430–435. doi: 10.1038/nmeth.1584.

Rinninella, E. *et al.* (2019) "What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases," *Microorganisms*, 7(1). doi: 10.3390/microorganisms7010014.

Röst, H. L. *et al.* (2014) "OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data," *Nature Biotechnology*, pp. 219–223. doi: 10.1038/nbt.2841.

Röst, H. L. *et al.* (2016) "TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics," *Nature methods*, 13(9), pp. 777–783.

Schroeder, B. O. (2019) "Fight them or feed them: how the intestinal mucus layer manages the gut microbiota," *Gastroenterology report*. Oxford University Press (OUP), 7(1), pp. 3–12.

Schubert, O. T. *et al.* (2015) "Building high-quality assay libraries for targeted analysis of SWATH MS data," *Nature protocols*, 10(3), pp. 426–441.

Segers, K. *et al.* (2019) "Analytical techniques for metabolomic studies: a review," *Bioanalysis*, pp. 2297–2318. doi: 10.4155/bio-2019-0014.

Sender, R., Fuchs, S. and Milo, R. (2016) "Revised Estimates for the Number of Human and Bacteria Cells in the Body," *PLoS biology*. Public Library of Science, 14(8), p. e1002533.

Shteynberg, D. *et al.* (2011) "iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates," *Molecular & cellular proteomics: MCP*, 10(12), p. M111.007690.

Shteynberg, D. *et al.* (2013) "Combining results of multiple search engines in proteomics," *Molecular & cellular proteomics: MCP*, 12(9), pp. 2383–2393.

Silva, R. *et al.* (2021) "geneRFinder: gene finding in distinct metagenomic data complexities," *BMC bioinformatics*, 22(1), p. 87.

Sonnenburg, E. D. *et al.* (2016) "Diet-induced extinctions in the gut microbiota compound over generations," *Nature*, 529(7585), pp. 212–215.

Sturm, M. *et al.* (2008) "OpenMS – An open-source software framework for mass spectrometry," *BMC Bioinformatics*. doi: 10.1186/1471-2105-9-163.

Sung, J. *et al.* (2017) "Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis," *Nature communications*, 8, p. 15393.

Suomi, T. and Elo, L. L. (2017) “Enhanced differential expression statistics for data-independent acquisition proteomics,” *Scientific reports*, 7(1), p. 5869.

Tamboli, C. P. *et al.* (2004) “Dysbiosis in inflammatory bowel disease,” *Gut*. BMJ, 53(1), pp. 1–4.

Tanca, A. *et al.* (2017) “Potential and active functions in the gut microbiota of a healthy human cohort,” *Microbiome*, 5(1), p. 79.

Tatusov, R. L. *et al.* (2000) “The COG database: a tool for genome-scale analysis of protein functions and evolution,” *Nucleic acids research*. Oxford University Press (OUP), 28(1), pp. 33–36.

Thursby, E. and Juge, N. (2017) “Introduction to the human gut microbiota,” *Biochemical Journal*, 474(11), pp. 1823–1836.

Tierney, B. T. *et al.* (2019) “The Landscape of Genetic Content in the Gut and Oral Human Microbiome,” *Cell host & microbe*, 26(2), pp. 283-295.e8.

Tremaroli, V. and Bäckhed, F. (2012) “Functional interactions between the gut microbiota and host metabolism,” *Nature*, 489(7415), pp. 242–249.

Tsou, C.-C. *et al.* (2015) “DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics,” *Nature methods*, 12(3), pp. 258–64, 7 p following 264.

Turnbaugh, P. J. *et al.* (2006) “An obesity-associated gut microbiome with increased capacity for energy harvest,” *Nature*, 444(7122), pp. 1027–1031.

Turnbaugh, P. J. *et al.* (2009) “A core gut microbiome in obese and lean twins,” *Nature*, 457(7228), pp. 480–484.

Ubeda, C., Djukovic, A. and Isaac, S. (2017) “Roles of the intestinal microbiota in pathogen protection,” *Clinical & translational immunology*, 6(2), p. e128.

UniProt Consortium (2021) “UniProt: the universal protein knowledgebase in 2021,” *Nucleic acids research*, 49(D1), pp. D480–D489.

Vancamelbeke, M. and Vermeire, S. (2017) “The intestinal barrier: a fundamental role in health and disease,” *Expert review of gastroenterology & hepatology*, 11(9), pp. 821–834.

Verberkmoes, N. C. *et al.* (2009) “Shotgun metaproteomics of the human distal gut microbiota,” *The ISME journal*. Springer Science and Business Media LLC, 3(2), pp. 179–189.

Vogel, C. and Marcotte, E. M. (2012) “Insights into the regulation of protein abundance from proteomic and transcriptomic analyses,” *Nature Reviews Genetics*, pp. 227–232. doi: 10.1038/nrg3185.

Wang, Y. *et al.* (2020) “Metaproteomics: A strategy to study the taxonomy and functionality of the gut microbiota,” *Journal of Proteomics*, p. 103737. doi: 10.1016/j.jprot.2020.103737.

Xiong, W. *et al.* (2015) “Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota,” *Proteomics*, 15(20), pp. 3424–3438.

Yang, X. *et al.* (2009) “More than 9,000,000 unique genes in human gut bacterial community: estimating gene numbers inside a human body,” *PloS one*, 4(6), p. e6074.

Yates, J. R., 3rd *et al.* (1998) “Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis,” *Analytical chemistry*, 70(17), pp. 3557–3565.

Zhang, X. *et al.* (2016) “MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota,” *Microbiome*, 4(1), p. 31.

Zhang, X. *et al.* (2018) “Assessing the impact of protein extraction methods for human gut metaproteomics,” *Journal of proteomics*, 180, pp. 120–127.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

PunaMusta Oy, Turku, Finland 2021

ISBN 978-951-29-8740-5 (pdf)
ISSN 0355-9483 (print)
ISSN 2343-3213 (online)