universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Metagenomic analysis of microbial communities in anaerobic digestion

verfasst von | submitted by

Dipl.-Ing. Lisa Bauer

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 875
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Bioinformatik
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Thomas Rattei

Acknowledgements

I would like to thank my co-supervisors Ade and Daan for their openness to take on my proposed topic for this thesis. I strongly benefitted from their guidance and their dedication to support me throughout this project.

I want to thank my supervisor Thomas for giving me the possibility to work on this thesis.

I am thankful for the support and advice I have received from Petra, Joana, and Bela from the Joint Microbiome Facility.

Thank you to my fellow bioinformatics students for their team spirit throughout this journey.

I am grateful for the support I received from the competence centre BEST – Bioenergy and Sustainable Technologies GmbH, in particular the COMET funding program BEST 4.0 (FFG proposal nr. 40920195).

Special thanks go to Bernhard for his expert advice regarding biogas-related topics and for initiating the idea for this thesis. Thank you for making it possible for me to pursue this study programme alongside my work.

This work would not have been possible without my amazing colleagues at BEST. Thank you for stepping in, taking over responsibilities, and being supportive throughout my studies.

I want to thank Laakirchen Papier AG, in particular Katharina Prall and Anton Marijanovic, and Agrana Stärke GmbH, specifically Wilfried Neuhauser, for their support.

Thank you to my partner and my family for their constant support and encouragement.

Abstract

Anaerobic digestion can be used to obtain renewable energy in the form of biogas from various sources and thus, to valorise organic wastewaters or residues. Under anaerobic conditions, microorganisms break down organic matter in multiple steps to biogas, i.e. mainly methane and carbon dioxide. The microorganisms accomplishing this process and their complex interactions are not well understood yet. The aim of this thesis is to investigate the microbial community in various anaerobic digestion systems and to provide insights into understudied systems, using molecular methods such as amplicon-based sequencing and metagenomics. Samples from anaerobic digestion systems of various sizes, setups, and using a variety of substrates, are analysed by amplicon-sequencing based methods to investigate their microbial community composition. The suitability of three databases for taxonomic classification is evaluated before choosing one for taxonomic profiling of the systems. The results are contextualised with a publicly available dataset to find global patterns in the taxonomic profiles of diverse anaerobic digestion systems. Selected systems are further investigated by a metagenomic approach. Metagenome-assembled genomes (MAGs) are recovered after long-read sequencing. They are functionally annotated to obtain information about their potential roles in the investigated systems. The amplicon-sequencing based analysis of various anaerobic digestion systems revealed patterns in the community composition and first insights into the potential roles of taxa. Amplicon sequence variants (ASVs) were found to be predominantly system-specific. After long-read sequencing of samples from two systems treating industrial wastewater, 419 MAGs were obtained. In a functional analysis, pathways and metabolites were assigned to MAGs. A number of MAGs that could be responsible for the production of acidogenesis products could be identified and the results revealed interesting potential interactions between acetogenic and sulphate reducing bacteria and methanogens. The genes encoding enzymes for hydrogenotrophic and acetoclastic methanogenesis were found in all archaeal MAGs. Overall, this thesis provides insights into the microbial community and their potential functions.

Kurzzusammenfassung

Anaerober Abbau kann verwendet werden, um aus verschiedenen Quellen erneuerbare Energie in Form von Biogas zu gewinnen und damit organische Abwässer oder Reststoffe zu valorisieren. Mikroorganismen bauen in einer anaeroben Umgebung organisches Material in mehreren Schritten zu Biogas, also größtenteils Methan und Kohlendioxid, ab. Die Mikroorganismen, die diesen Prozess durchführen, und deren komplexe Wechselwirkungen sind noch nicht vollständig erforscht. Das Ziel dieser Arbeit ist es, durch die Verwendung molekularer Methoden wie amplikonbasierte Sequenzierung oder Metagenomik die mikrobielle Gemeinschaft in unterschiedlichen Biogasanlagen zu untersuchen und Einblicke in noch nicht erforschte Systeme zu vermitteln. Proben aus Biogasanlagen, die unterschiedliche Größen und Konfigurationen haben und unterschiedliche Substrate verwenden, werden durch auf Amplikonsequenzierung basierenden Methoden analysiert, um die Zusammensetzung ihrer mikrobiellen Gemeinschaften zu untersuchen. Die Eignung von drei Datenbanken für die taxonomische Klassifizierung wird bestimmt, bevor eine für die Erstellung von taxonomischen Profilen der Systeme verwendet wird. Die Ergebnisse werden durch einen öffentlich verfügbaren Datensatz kontextualisiert, um globale Muster in den taxonomischen Profilen von diversen Biogasanlagen zu finden. Ausgewählte Systeme werden durch einen metagenomischen Ansatz weiter untersucht. Metagenom-assemblierte Genome (MAGs) werden nach einer Long-Read-Sequenzierung gewonnen. Diese werden funktionell annotiert, um Informationen über ihre potenziellen Rollen in den untersuchten Systemen zu erhalten. Die auf Amplikonsequenzierung basierende Analyse der unterschiedlichen Biogasanlagen zeigte Muster in der Zusammensetzung der Gemeinschaften, sowie erste Einblicke in die potenziellen Rollen von Taxa. Amplicon Sequence Variants (ASVs) erwiesen sich als überwiegend systemspezifisch. Nach der Long-Read-Sequenzierung von Proben aus zwei Systemen, welche industrielles Abwasser behandeln, wurden 410 MAGs gewonnen. In einer funktionellen Analyse wurden Stoffwechselwege und Metabolite MAGs zugeordnet. Einige MAGs, die für die Herstellung von Acidogenese-Produkten zuständig sein könnten, konnten identifiziert werden und die Ergebnisse zeigten interessante potenzielle Interaktionen zwischen acetogenen und Sulphat-reduzierenden Bakterien und Methanogenen. Die Gene, welche die Enzyme für hydrogenotrophe und acetoklastische Methanogenese codieren, wurden in allen MAGs, welche als Archaea klassifiziert wurden, gefunden. Insgesamt gibt diese Arbeit Einblicke in die mikrobielle Gemeinschaft und ihre potenziellen Funktionen.

List of abbreviations

AD	Anaerobic digestion
ANI	Average nucleotide identity
ASV	Amplicon sequence variant
CA	Correspondence analysis
CH_4	Methane
CO_2	Carbon dioxide
COD	Chemical oxygen demand
CSTR	Continuously stirred tank reactor
DADA	Divisive Amplicon Denoising Algorithm
EC	Enzyme Commission
EGSB	Expanded granular sludge bed
GTDB	Genome Taxonomy Database
GTDB-Tk	Genome Taxonomy Database Toolkit
H_2	Hydrogen
H_2S	Hydrogen sulphide
HMM	Hidden Markov model
HPLC	High performance liquid chromatography
IC	Internal circulation
JMF	Joint Microbiome Facility
KEGG	Kyoto encyclopedia of genes and genomes
KO	KEGG orthology
MAG	Metagenome-assembled genome
MCR	Methyl-coenzyme M reductase
MiDAS	Microbial Database for Activated Sludge
NCBI	National Center for Biotechnology Information
NH_3	Ammonia
NMDS	Non-metric multidimensional scaling
ONT	Oxford Nanopore Technologies
OTU	Operational taxonomic unit
PCA	Principal components analysis

PCR	Polymerase chain reaction
pHMM	Profile hidden Markov model
RED	Relative evolutionary distance
rRNA	Ribosomal ribonucleic acid
SGB	Species-level genome bin
SRB	Sulphate-reducing bacteria
TOC	Total organic carbon
TSAD	Two-stage anaerobic digestion
TSS	Total sum scaling
UASB	Upflow anaerobic sludge blanket
VFA	Volatile fatty acid
zOTU	Zerio radius operational taxonomic unit

Table of contents

1 Introduction

1.1 Organic wastewater and residues in industry

Various industrial sectors such as chemical, mining, agro-industry, food processing, or paper depend on large amounts of water (Singh et al. 2023). In Europe, 54 % of water that is used for human activities are used for industrial purposes (Granger et al. 2019). This is accompanied by large volumes of wastewater, which is often characterised by high concentrations of organic matter and pollutants such as sulphate (Wu et al. 2025; Drosg et al. 2021).

On a global scale many counties do not have elaborated reporting systems regarding the amount of wastewater that is produced and treated (UN-Water 2024). Nevertheless, UN-Water calculated that only 27 % of the wastewater produced in 2024 were safely treated (UN-Water 2024). Discharging wastewater, which is not properly treated, to water bodies leads to environmental pollution and risks for human health, e.g. by contaminating drinking water sources (Cai et al. 2016; Singh et al. 2023). Wastewater treatment is therefore an important topic for industries. On the one hand, negative effects on the environment can be limited by treating wastewater before discharging it. On the other hand, the reuse of water can contribute to a reduction of freshwater demand which is an important issue for developments such as climate change (UN-Water 2024; Granger et al. 2019). This task can either be outsourced to urban wastewater treatment plants, or wastewater can be treated on-site, for example in biological systems (Granger et al. 2019; Singh et al. 2023).

In the paper making process, the raw material is washed and cleaned before it is stewed (Liang et al. 2023). Then, chemicals are applied to disintegrate the raw material during pulping, i.e. obtaining cellulose by separating lignin from the fibres (Liang et al. 2023). The pulp is washed and contaminating materials are removed in a screening step (Liang et al. 2023). The cleaned pulp is bleached and then used to produce paper (Liang et al. 2023). The wastewater resulting from the pulp and paper industry is usually highly biologically degradable (Möbius 2010). The main components are dissolved substances derived from wood, i.e. poly- and oligosaccharides from hemicelluloses and lignin-based substances (Möbius 2010; Lindholm-Lehto et al. 2015). Thus, the wastewater is made up of easily degradable carbohydrates such as starch, amylose, and cellulose and difficult to degrade lignin derivatives (Möbius 2010; Liang et al. 2023). One of the main sources of wastewater in the paper making process is pulping (Liang et al. 2023). The stewing step introduces black liquor to the waste stream and bleaching introduces organochlorides (Liang et al. 2023). White water, which is the wastewater from the final step of turning pulp into paper, introduces materials such as fillers and rubbers to the wastewater (Liang et al. 2023). Nevertheless, there are typically only low amounts of toxic substances in the wastewater because most chemicals that are used in the paper making process are not water soluble (Möbius 2010). However, resin acids, organochlorides and, especially in factories that recycle wastewater, sulphate can be found in the wastewater (Möbius 2010; Liang et al. 2023; Meyer and Edwards 2014; Lindholm-Lehto et al. 2015).

Similarly, wastewaters are produced along the steps of processing corn or other feedstocks to obtain starch (Drosg et al. 2021). The cleaned corn is softened in a steeping step and then germ, fibre, and gluten are separated from the starch (Eckhoff and Watson 2009). The resulting starch slurry is dried to obtain corn starch or modified to obtain modified starch products (Eckhoff and Watson 2009; Juneja et al. 2019). Starch saccharification can be done to obtain simpler sugars such as glucose from starch (Drosg et al. 2021). Steeping, i.e. mixing corn with water and sulphur dioxide, softens the cleaned corn, loosens the protein matrix that surrounds the starch, and thus prepares it for wet milling (Eckhoff and Watson 2009; Juneja et al. 2019). In the first stage of steeping, lactic acid is produced by an active *Lactobacillus* sp. fermentation (Eckhoff and Watson 2009). Water is often re-used by circulating it within the process and large

amounts of wastewater are thus avoided (Drosg et al. 2021). The main contributors to the wastewater are condensates from steeping or drying, and wash waters which yields a usually acidic wastewater with easily biodegradable compounds such as residual starch, simple sugars, and lactic acid, but also considerable amounts of sulphate, derived from the sulphur dioxide (Drosg et al. 2021; Neogi 2020; Shubhaneel et al. 2018).

Besides wastewater, a variety of residues accrue in industries such as the agro-industrial sector. Considerable amounts of wastes such as crop residues or molasses are produced by industries that process agricultural products. They can partially be turned into by-products and used as animal feed or fertilizer but as they are available in excess, they are often disposed at landfills and incinerated. This releases gaseous pollutants and valuable nutrients are lost. (Ogbu and Okey 2023)

To establish circular economies it is therefore essential to valorise both types of waste instead of disposing them. An attractive way of dealing with accumulating organic waste is anaerobic digestion (AD) (Stoyancheva et al. 2023). This technology is traditionally used to obtain methane as an energy carrier from energy crops such as corn silage (Weinrich and Nelles 2021). However, it is versatile and a wide range of substrates can be used, including complex wastes such as industrial wastewaters or residues (Weinrich and Nelles 2021; Camargo et al. 2023). AD can be seen as a sustainable waste and wastewater treatment technology, breaking down organic matter and producing renewable energy (Wu et al. 2025; Wall and O'Shea 2025). By applying AD, industries can reduce their environmental impact and become independent from fossil fuels at the same time (Stoyancheva et al. 2023; Wall and O'Shea 2025).

1.2 Principles of anaerobic digestion

AD is used worldwide to obtain renewable energy from various sources of organic matter such as energy crops, harvest residues, organic by-products, and waste (Weinrich and Nelles 2021). This is done by breaking down organic matter and capturing the stored energy in the form of methane (Weinrich and Nelles 2021; Cai et al. 2016). It outcompetes aerobic waste treatment and is an essential technology for replacing fossil energy with renewable energy to cope with the finite energy reserves and tackle the climate crisis (Weinrich and Nelles 2021).

Under anaerobic conditions, microorganisms break down organic matter and convert it to biogas in several steps (Da Costa Gomez 2013). The conversion process will be elaborated in more detail below. The biogas composition mainly depends on the substrate, with typically 45-75 % methane (CH_4) and 25-45 % carbon dioxide (CO_2) (Weinrich and Nelles 2021; Da Costa Gomez 2013). Besides these main components, substances such as water vapour, hydrogen (H_2), hydrogen sulphide (H_2S), or ammonia (NH_3) can be found at lower concentrations (Weinrich and Nelles 2021; Da Costa Gomez 2013). In many cases, the produced biogas is used directly on-site for heat and electricity generation after drying and desulphurisation (Weinrich and Nelles 2021). However, it can alternatively be fed into the gas grid to replace natural gas if it is cleaned ('upgraded') to reach a concentration of approximately 98 % methane (Weinrich and Nelles 2021; Da Costa Gomez 2013). The non-gaseous residues ('digestate') can be used as fertiliser or, in the case of applying the technology as wastewater treatment, discharged.

The AD process is influenced by multiple parameters such as substrate characteristics, reactor type, and pH (Camargo et al. 2023; Domínguez-Espinosa et al. 2023).

1.2.1 Substrates for anaerobic digestion

As mentioned above, a broad range of substrates can be used in AD. Their suitability for and impact on the process is determined by various factors. One of the most important factors is the composition of the feedstock, thus providing the nutrients needed for the metabolism of the

anaerobic microbial community (Al Seadi et al. 2013). Therefore, suitable substrates have high content of sugar, starch, proteins, or fats (Al Seadi et al. 2013).

Nutrients, including macro- and micronutrients, should be supplied in a suitable ratio. The organic matter should be easily biodegradable, as the digestibility of the substrate has an impact on both, the decomposition efficiency and time. The digestibility of some materials can be improved by pre-treatment which decreases particle sizes or breaks down complex biomass such as lignin. Regarding the components of the substrate that is fed to the AD reactor, substances with toxic or inhibitory effects on the microbial community should be avoided. Moreover, the efficiency of AD is optimal around a neutral pH and thus, substrates with a buffering capacity in this range are preferred in traditional setups. Balancing all of these factors in an optimal way will lead to a stable process and high methane yield. If one substrate does not have optimal composition, this can be achieved by mixing various substrates and running the process as a co-digestion. Another important substrate parameter is the dry matter content. However, the optimal value strongly depends on the reactor type. In general, the dry matter content should be in a range that ensures the reactor content can be properly stirred. (Al Seadi et al. 2013)

Apart from the described characteristics, availability of the substrates is important. They are provided by various sectors, including the agricultural, industrial, and municipal sector (Al Seadi et al. 2013). Renewable raw materials from the agricultural sector are the most widely used substrates in Austria and Germany (Weinrich and Nelles 2021; Al Seadi et al. 2013). In these countries, energy crops are often solely grown for this purpose due to their high methane potential (Al Seadi et al. 2013; Weinrich and Nelles 2021). However, this drastically reduces the sustainability of the process because a considerable amount of resources and energy are needed to provide the crops and their production is in direct competition with food production (Al Seadi et al. 2013). It is therefore highly preferable to use by-products or waste, as this requires fewer resources and, simultaneously, they are valorised (Al Seadi et al. 2013). In this regards, industrial waste and residues are a promising alternative, but to date, they are only used in a small number of biogas plants (Weinrich and Nelles 2021). One possible disadvantage preventing widespread adoption is that waste streams can contain pollutants that originate from the industrial processes, which can inhibit AD and make the digestate unsuitable as fertiliser (Al Seadi et al. 2013). However, they are often mainly made up of easily digestible components with high methane potential (Al Seadi et al. 2013) and their suitability for AD should be investigated in more detail.

1.2.2 Reactor technology

1.2.2.1 Reactor types

In Europe, the most widely used reactor type for AD is the continuously stirred tank reactor (CSTR) (Wall and O'Shea 2025). They are well suited for substrates with high levels of solids, as long as the required complete mixing can be ensured (Show et al. 2020; Mao et al. 2015). However, CSTRs have several disadvantages. They are operated at low organic loading rates and long hydraulic retention times, i.e. the time the wastewater spends in the reactor (Show et al. 2020; Tauseef et al. 2013; De Lemos Chernicharo 2007). This means that the substrate can only slowly be fed to the reactor which requires large reactor volumes (Show et al. 2020; Tauseef et al. 2013). Moreover, biomass, i.e. the active microbial community, is not retained in the reactor but washed out with the effluent (Show et al. 2020; Tauseef et al. 2013).

The Upflow Anaerobic Sludge Blanket (UASB) reactor which was designed by Lettinga et al. at Wageningen University (Lettinga et al. 1980) is better suitable for treating wastes with high water content, such as industrial wastewaters (Show et al. 2020; Tauseef et al. 2013). This reactor type supports granule formation and thereby retains the biomass inside the reactor,

even without carrier materials (Tauseef et al. 2013; De Lemos Chernicharo 2007). This allows treating high organic loading rates at low hydraulic retention time (Tauseef et al. 2013). Granular biomass is an important feature of this reactor type and will be discussed in more detail below. In a UASB, the wastewater feed enters the reactor through the bottom and flows upward through a sludge blanket (i.e. granular biomass) (Tauseef et al. 2013). This leads to efficient mixing of wastewater and microbial granules and thus quick degradation (Tauseef et al. 2013). On top of the reactor, the solids (mainly granules) are held back while gas and liquid leave the reactor (Tauseef et al. 2013). UASB reactors have been further developed to achieve higher organic loading rates. The large height-to-diameter ratio of the Expanded granular sludge bed (EGSB) reactor causes better mixing and ensures efficient methane production at low operating costs (Tauseef et al. 2013; Domínguez-Espinosa et al. 2023). Even high organic loads and shorter hydraulic retention times can be handled by internal circulation (IC) reactors (Tauseef et al. 2013). IC reactors can be viewed a combination of two UASB reactor compartments on top of each other to separate biogas production and solids retention (Tauseef et al. 2013; De Lemos Chernicharo 2007; Show et al. 2020).

1.2.2.2 Granular biomass

One of the main challenges in high-rate wastewater treatment by AD is that the microorganisms have low growth rates and are thus easily washed out when the retention time is low (De Lemos Chernicharo 2007). This can be overcome by making the hydraulic retention time, independent from the solid retention time, i.e. the time that the microorganisms spend in the reactor (De Lemos Chernicharo 2007). One possibility to achieve this is using granular biomass which is retained in the system while the liquids and gases are not (Show et al. 2020). As described above, granular biomass is typically used in UASB reactors and their variants (Show et al. 2020). Granules are dense and settle rapidly in bioreactors (Show et al. 2020), which is the reason why they are retained more easily than dispersed biomass (Trego et al. 2021). They are aggregates of diverse microbial cells and their structure enables efficient substrate transfer between the functionally specialized microorganisms, allowing a syntrophic symbiosis (De Lemos Chernicharo 2007; Trego et al. 2021).

The formation of granules is a complex, spontaneous microbial development that takes several months and is not yet fully understood (Show et al. 2020; Harirchi et al. 2022). Factors that influence the granule formation, their size, and their structure are, for example, substrate characteristics, pH, the presence of bivalent cations, temperature and the upflow velocity in the system (Show et al. 2020; De Lemos Chernicharo 2007). Several models suggest that methanogens are the core of the granules (Show et al. 2020; De Lemos Chernicharo 2007). The outer layers are described in different ways in various models but typically consist of acid-producing and sulphur-reducing bacteria as well as hydrogenotrophic methanogens (Show et al. 2020; De Lemos Chernicharo 2007). The granular structure makes the system as a whole more resilient to process disturbances as the microorganisms, especially those in the core, are protected against environmental changes (Trego et al. 2021; Show et al. 2020).

Granular biomass in AD systems has not yet been analysed extensively using metagenomic or metaproteomic methods (Trego et al. 2021). There have, however, been studies based on amplicon sequencing that describe the microbial community structure and response to environmental changes (Trego et al. 2021).

1.2.2.3 Process monitoring

In practice, the AD process is usually monitored to be able to react to upcoming disturbances and thus maintain a stable process. The volatile fatty acids (VFA) content and profile are among the most informative parameters. VFAs are short-chained organic acids and important intermediate metabolites in the AD process (Drosg 2013). An accumulation of VFAs inhibits

methanogenesis and thus, it is important to detect increasing concentrations in an early stage (Drosg 2013). Typically, acetic acid accumulates to some degree even in a stable process while the accumulation of propionic, butyric, or valeric acid can indicate severe problems (Drosg 2013; Harirchi et al. 2022). Overall, the VFA profile only provides indirect information by indicating process disturbances, but not their reasons (Drosg 2013). To be able to understand the causes of disturbances that lead to specific profiles of VFA accumulation, it is necessary to better understand the underlying microbial metabolic processes. The pH is often tracked, but it only changes after a disturbance arises and is therefore not suitable as an early indicator for problems (Drosg 2013). For example, the pH drops after VFAs start to accumulate. However, it provides valuable information about the state of the process and can rapidly be determined (Drosg 2013). The volume and composition of the produced biogas is monitored to directly obtain information about the targeted product which is usually methane (Drosg 2013). The efficiency of the process can be determined by the ratio of organic matter in the effluent and in the influent (Holl et al. 2022).

1.2.3 Metabolic processes in anaerobic digestion

As mentioned above, complex organic matter, i.e. carbohydrates, proteins, and lipids, are broken down to methane and carbon dioxide in AD (Harirchi et al. 2022). The process can be described by four sequential, but also inter-dependent stages, namely hydrolysis, acidogenesis, acetogenesis, and methanogenesis (Venkiteshwaran et al. 2015). They are driven by complex interactions between various microorganisms who live in syntrophic relationships (De Lemos Chernicharo 2007). These relationships are based on exchanging nutrients and ensuring thermodynamic stability (Cao et al. 2025). It is essential that the microorganisms specialised in tasks associated with different stages are in ecological balance with each other to ensure a stable process that efficiently breaks down organic matter (Venkiteshwaran et al. 2015). The four stages and their interconnectedness are depicted in a schematic overview in Figure 1 and will be elaborated below.

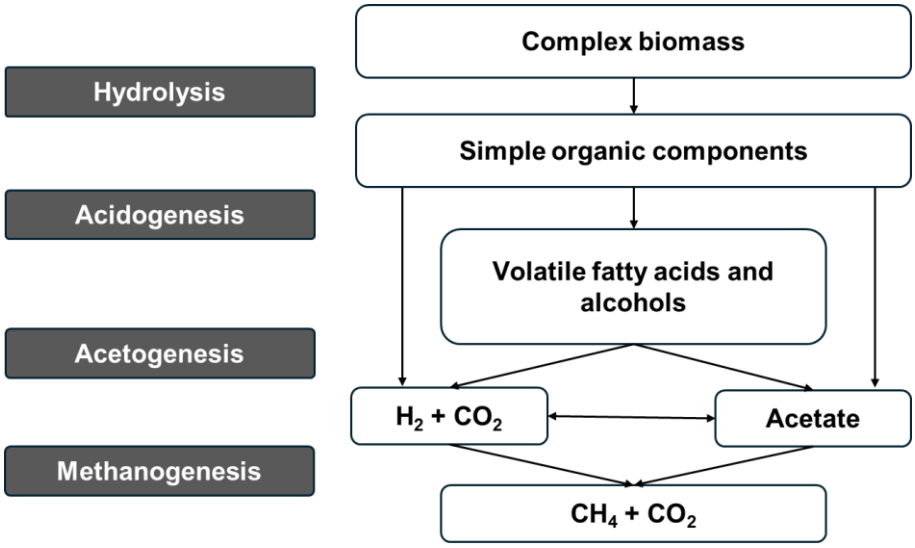

Figure 1: *Scheme of the four stages of anaerobic digestion: Hydrolysis, acidogenesis, acetogenesis, and methanogenesis.*

1.2.3.1 Hydrolysis

The hydrolysis stage is the first step that is necessary to break down complex organic matter. Polymeric substances such as carbohydrates, lipids, and proteins are hydrolysed to obtain simple soluble oligomeric or monomeric compounds such as oligo- and monosaccharides, fatty

acids, amino acids, and alcohols (Cao et al. 2025). This is achieved outside of the cell by hydrolytic bacteria who excrete exoenzymes including cellulases, amylases, lipases, and proteases, among others (De Lemos Chernicharo 2007; Venkiteshwaran et al. 2015). These convert large polymers into small molecules that can pass through the cell membrane (De Lemos Chernicharo 2007; Cao et al. 2025). Thus, hydrolysis prepares complex organic matter for further metabolization.

Hydrolytic bacteria are mainly found in the phyla Bacteroidota and Bacillota, but also Actinomycetota, Pseudomonadota and Chloroflexi (Cao et al. 2025; Venkiteshwaran et al. 2015). They are phylogenetically diverse and often many species in the microbial community can accomplish this task (Venkiteshwaran et al. 2015). This functional redundancy makes the hydrolytic community robust against environmental changes (Cao et al. 2025). Their growth rate strongly depends on substrate parameters such as particle size and digestibility, as described above (Venkiteshwaran et al. 2015).

1.2.3.2 Acidogenesis

Acidogenic bacteria, which are also a large and diverse group, convert the metabolites of hydrolytic bacteria into even simpler molecules inside their cells and then excrete them again (De Lemos Chernicharo 2007; Cao et al. 2025). Similarly, to hydrolysis, this is mainly achieved by members of the phyla Bacteroidota, Bacillota, Chloroflexota, and Pseudomonadota (Cao et al. 2025; Venkiteshwaran et al. 2015). The main products of this step are volatile fatty acids (VFAs), such as acetic, propionic, butyric, and valeric acid (De Lemos Chernicharo 2007; Cao et al. 2025). Moreover, alcohols, lactic acid, CO_2, H_2, NH_3 and H_2S are produced (De Lemos Chernicharo 2007). Acetic acid is usually the most abundant VFA and is formed from all organic compounds (De Lemos Chernicharo 2007).

The metabolic pathways involved in this stage include various fermentations, for example lactic acid, butyric acid, propionic acid, and alcohol fermentations (Heider 2014).

Organic substrates are at the same time electron donors and acceptors

In fermentations, organic substrates act as both electron donor and acceptor (Hackmann 2024). In the case of fermentations in the acidogenesis stage of AD, these organic substrates are typically monomeric products from the hydrolysis stage that are converted to fermentation products, i.e. acids and alcohols (Heider 2014). An important intermediate metabolite is pyruvate (Cao et al. 2025).

Of the acidogenesis products, acetic acid, H_2, and CO_2 are the only ones that can directly be used by methanogens (De Lemos Chernicharo 2007). The other products are further converted by acetogenic bacteria first (Hao et al. 2020). In a stable system, the VFAs are immediately consumed during acetogenesis or methanogenesis (see below). However, in case of disturbances, they accumulate and the pH drops, which further disrupts the whole process (De Lemos Chernicharo 2007; Harirchi et al. 2022).

1.2.3.3 Acetogenesis

The acetogenesis summarises various metabolic processes that yield acetate, CO_2, and H_2. It involves multiple syntrophic as well as competing pathways.

Homoacetogenic bacteria use H_2 and CO_2 in the Wood-Ljungdahl pathway to produce acetate, which can then be used by acetoclastic methanogens (Harirchi et al. 2022; Cao et al. 2025). Conversely, syntrophic acetate oxidising bacteria use the same pathway in reversed direction to oxidise acetate into H_2 and CO_2 (Cao et al. 2025).

Syntrophic fatty acid oxidising bacteria convert acidogenesis products (mainly propionic or butyric acid, but also ethanol) to acetic acid, formate, H_2, and CO_2 (De Lemos Chernicharo 2007; Harirchi et al. 2022; Hao et al. 2020; Lim et al. 2020). They are usually found at low abundances due to their low growth rates (Hao et al. 2020). These oxidation reactions are thermodynamically unfavourable with a positive Gibbs free energy at standard conditions (De Lemos Chernicharo 2007; Harirchi et al. 2022). This is only possible in a syntrophic relationship with methanogens: Hydrogenotrophic archaea use the produced H_2 to produce methane, which enables the oxidation reactions (De Lemos Chernicharo 2007; Harirchi et al. 2022). Alternative syntrophic partners who metabolise H_2 are homoacetogenic and sulphate-reducing bacteria. They compete with methanogens for H_2 (Harirchi et al. 2022). In addition, they are in a syntrophic relationship with acetate consumers such as acetoclastic methanogens (Westerholm et al. 2021). Syntrophic fatty acid oxidation has been associated with some members of the phyla Bacillota (e.g. *Syntrophomonas*) and Thermodesulfobacteriota (e.g. *Syntrophus*) (Westerholm et al. 2021).

1.2.3.4 Methanogenesis

Finally, the products of the previously described stages are converted to methane and CO_2 by methanogenic archaea (De Lemos Chernicharo 2007). Three distinct methanogenesis pathways have been described. They all have in common that they depend on coenzyme M (CoM) and methyl-coenzyme M reductase (MCR) (Harirchi et al. 2022). Methanogenic archaea are capable of one or more of these pathways. It has been observed that typically, 65-75 % of methane in anaerobic digesters are produced by the acetoclastic pathway and 25-35 % by the hydrogenotrophic pathway, while methylotrophic methanogenesis can be neglected (Domínguez-Espinosa et al. 2023; Niya et al. 2024).

Acetoclastic methanogenesis

In many anaerobic digesters, acetoclastic methanogens have been found to be the most prevalent archaea, even though they are sensitive to inhibitors and only few species belong to this group (De Lemos Chernicharo 2007; Harirchi et al. 2022). They produce methane through the cleavage of acetic acid. The methyl group is reduced to methane, and the carboxyl group is oxidized to CO_2, as described by the following equation (De Lemos Chernicharo 2007).

$$CH_3COOH \rightarrow CH_4 + CO_2$$

Thus, they are dependent on bacteria providing acetate. The CO_2 fraction in the biogas mainly originates from this pathway. They are represented by members of the orders Methanotrichales and Methanosarcinales (Cao et al. 2025).

Hydrogenotrophic methanogenesis

In the hydrogenotrophic pathway, CO_2 or formate is reduced to CH_4 using H_2 as electron donor in the following reactions (De Lemos Chernicharo 2007; Harirchi et al. 2022).

$$CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O$$

$$H_2CO_2 + 3H_2 \rightarrow CH_4 + 2H_2O$$

The pathway is found in many methanogens, for example in members of the order Methanobacteriales (Cao et al. 2025; Harirchi et al. 2022).

As described above, the quick metabolization of H_2 is important for the stability of syntrophic acetogenic pathways (De Lemos Chernicharo 2007; Cao et al. 2025). This is achieved by this group of methanogens who consume the H_2 that is produced in the other stages (De Lemos Chernicharo 2007). Hydrogenotrophic methanogens have been described as more resistant to environmental stress than acetoclastic methanogens (Harirchi et al. 2022).

Methylotrophic methanogenesis

In methylotrophic methanogenesis, methanol or other methylated compounds are converted to methane and CO_2 by demethylation (Harirchi et al. 2022). However, this pathway has not been found to be of importance in AD systems (Harirchi et al. 2022).

1.2.3.5 Sulphate reduction

The four stages described above are the main AD processes that yield methane. However, depending on the waste stream characteristics, additional processes are running in parallel. In wastewaters with high sulphate concentrations, sulphate-reducing bacteria growth is stimulated (Madden et al. 2014). They oxidise organic compounds of the wastewater to CO_2 by reducing sulphate to sulphide (De Lemos Chernicharo 2007). Thus, they use the same substrates as the microorganisms described above, but with sulphate as electron acceptor instead of CO_2. This leads to competition for substrates. Sulphate reduction has thermodynamic advantages over CO_2 reduction and thus, sulphate reducers can outcompete hydrogenotrophic and acetoclastic methanogens at excess concentrations of sulphate (Stefanie et al. 1994; Madden et al. 2014). At a COD/sulphate ratio > 0.67, the sulphate is not enough to remove all of the COD and thus, methanogenesis runs after the sulphate is used up (Wu et al. 2025). Taxa that are known to be sulphate reducers that oxidise organic substrates to CO_2 are *Desulfobacterium*, *Desulfobacter*, and *Desulfococcus* (Harirchi et al. 2022). A different group of sulphate reducers, including *Desulfobulbus*, *Desulfomonas*, and *Desulfovibrio*, is known for incomplete oxidation of substrates to acetate (De Lemos Chernicharo 2007; Harirchi et al. 2022). They are in competition with acetogenic bacteria but can live in syntrophic relationships with acetoclastic and hydrogenotrophic methanogens by degrading propionate or butyrate to acetate and H_2 (Harirchi et al. 2022). However, the end product is H_2S for both groups, which is potentially toxic and inhibitory for other microorganisms (Madden et al. 2014).

1.2.3.6 Two-stage anaerobic digestion (TSAD)

In traditional systems, all of the four described stages run in one reactor (Holl et al. 2022). For various reasons, this is not the optimal configuration for AD. The microorganisms responsible for the early stages (hydrolysis and acidogenesis) have different physiological and environmental requirements from those responsible for the late stages (acetogenesis and methanogenesis) (Holl et al. 2022). Acidogenic bacteria prefer a slightly acidic pH while acetogenic bacteria and methanogenic archaea only grow at neutral to slightly alkaline pH (Holl et al. 2022). Moreover, the early and late stages often proceed at different speeds. When easily degradable substrates are used, hydrolysis and acidogenesis can be rapid while acetogenesis and methanogenesis are usually slow (Stoyancheva et al. 2023). In this setting, volatile fatty acids (VFAs) accumulate, the pH drops, and methanogenesis is inhibited (Holl et al. 2022). On the other hand, hydrolysis can be the rate limiting step if complex biomass is used as substrate (Holl et al. 2022). Therefore, it makes sense to decouple both processes by using two separate reactors, with an acidification reactor for hydrolysis and acidification stage and a methane reactor for the acetogenesis and methanogenesis stage (Holl et al. 2022). This way, the optimal conditions can be provided to all participating microorganisms, and the process can be controlled more easily (Holl et al. 2022). The loading rates for both reactors can be selected in a way that ensure optimal VFA concentration in the methane reactor and fluctuations in the substrate can be stabilised. For these reasons, two-stage anaerobic digestion (TSAD) systems often have higher methane yields than single-stage systems and the amount of energy that is recovered from the wastewater is maximised (Holl et al. 2022; Stoyancheva et al. 2023).

1.3 Microbial communities in anaerobic digestion

The anaerobic digestion of organic matter is driven by complex interactions between multiple microorganisms. Even though the general process has been known and described by the four-stages model for a while (Speece 1996) and several key taxa have been described, a complete understanding of the microbial communities and their individual members is still lacking, and the process remains to be seen as a “black box” (Stoyancheva et al. 2023; Niya et al. 2024). To understand the relationships between, and the tasks of individual taxa, it is essential to investigate them in more detail (Stoyancheva et al. 2023). Since traditional cultivation-based approaches to characterise the microbial community are labour-intensive and limited to the cultivable fraction (Campanaro et al. 2020; Jünemann et al. 2017), molecular techniques have been adding valuable insights to our current understanding of AD.

The following section provides an overview of these techniques and is followed by a description of recent insights in this field of study.

1.3.1 Overview of molecular techniques

1.3.1.1 Amplicon-based methods

Starting in the early 2000s, amplicon-based sequencing technologies have been used to study the composition of microbial communities in AD, with a steep increase of publications from 2013.

This widely used and cost-effective approach is based on the taxonomic identification of bacteria and archaea using marker genes (Pollock et al. 2018; Pjevac et al. 2021). A short region of the selected marker gene, which typically is the 16S rRNA gene, is amplified by Polymerase Chain Reaction (PCR) and sequenced (Pollock et al. 2018). The choice of the hypervariable region, as well as of the primers to amplify it, introduce biases, as none of them cover all bacteria and archaea (Pollock et al. 2018). Thus, some taxa are missed while some are overrepresented which affects the biological interpretation (Cai et al. 2016).

The raw sequencing data is then processed to obtain operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) which are sequences that represent distinct taxa (Pollock et al. 2018; Callahan, McMurdie, et al. 2016). By comparing them to databases, they are assigned taxonomies (Pollock et al. 2018). Finally, a table with relative abundances of taxa in the analysed samples is obtained.

This approach focuses on obtaining taxonomic profiles and can be used to study shifts in the microbial community. However, the assignment of functions using methods solely based on the 16S rRNA gene can only be done indirectly and should be avoided (Cai et al. 2016). Reference sequences that are used to infer functions are usually not associated with AD but were isolated from different environments (Campanaro et al. 2018). Thus, they might have distinct functions (Campanaro et al. 2018). The analysis of potential functions and metabolic pathways should instead be done using genome-based approaches.

1.3.1.2 Metagenomic methods

Metagenomic methods have been used in studies of AD to a limited extent since around 2013, and only as recently as in 2020, their application started to increase. Such methods are suitable to study the genomic potential regarding functions in the microbial community. Metagenomic shotgun sequencing methods are not dependent on amplification (Jünemann et al. 2017), and therefore less biased for specific taxa.

Short-read or long-read sequencing is used to sequence the extracted DNA (Kim et al. 2022). To obtain contigs, i.e. longer contiguous sequences, the reads are assembled (Kim et al. 2022;

Jünemann et al. 2017). This means that overlapping regions are used to join them together. This process is facilitated by using long-reads compared to short-reads, for example because long repetitive regions in the genome can more easily be resolved (Latorre-Pérez et al. 2020). However, long-reads are still more error-prone and expensive than short reads (Latorre-Pérez et al. 2020). Early metagenomic studies mainly used gene-centric approaches (Kim et al. 2022) to study the functional potential of the microbial community as a whole by annotating genes in contigs (Jünemann et al. 2017). Recent trends have shifted to genome-centric approaches to study individual genomes and interactions between taxa instead (Kim et al. 2022). For this method, an additional step is necessary to retrieve metagenome-assembled genomes (MAGs) from the contigs. This is done by clustering the contigs into groups (bins) based on their abundance and sequence structure (Kim et al. 2022; Kang et al. 2019). High quality bins, or MAGs, represent genomes (Kim et al. 2022). This can be followed by taxonomic classification to obtain taxonomic profiles (Jünemann et al. 2017). Furthermore, the resulting MAGs can be functionally annotated to study their genomic potential regarding metabolic pathways (Kim et al. 2022)

1.3.2 Current state of research

Valuable findings have arisen from multiple amplicon-based studies. They have mainly been focusing on identifying key members of the microbial community in AD (Dueholm et al. 2024; Agostini et al. 2023; Li et al. 2018), and on the parameters that shape the composition of the microbial community (Dueholm et al. 2024; Harirchi et al. 2022; de Jonge et al. 2020). Gene-centric metagenomic studies have shed light on the prevalence of metabolic pathways in anaerobic digesters (Cai et al. 2016; Jünemann et al. 2017). More recently, genome-centric studies have populated databases with AD-specific MAGs and increasing effort is put into investigating their functions and interactions (Breton-Deval et al. 2021; Campanaro et al. 2020; Beraud-Martínez et al. 2024).

The taxonomy for many of the microorganisms specific for AD is poorly characterized and taxonomical characterisation is often only possible at high taxonomic levels (Dueholm et al. 2024; Treu et al. 2016). This suggests that many unknown species are part of the biogas microbiome (Treu et al. 2016). This limitation was addressed by the Microbial Database for Activated Sludge (MiDAS) (Dueholm et al. 2024). The project originally focused on the taxonomy and phenotypes of microorganisms in activated sludge from wastewater treatment systems (McIlroy et al. 2015). Their findings are compiled in a full-length ASV database, which is a manual curation of the SILVA taxonomy (Quast et al. 2013), available at https://www.midasfieldguide.org. Yet unknown or unclassified taxa are assigned placeholder names that allow their identification in communities, even if they have not yet been characterised. In its more recent version (MiDAS 5), the was expanded with data from 285 full-scale anaerobic digesters (Dueholm et al. 2024), making it a promising classifier for AD samples. The inclusion of the AD samples yielded many new ASVs relative to the previous version of the MiDAS database which was focused on activated sludge-specific taxa. The novelty within the phylum Bacillota stood out, while the archaeal taxa identified were mostly well-known methanogens.

In an effort to find core members of AD communities, they have often been described and compared at the phylum level. Bacillota and Bacteroidota have been found to be the most abundant phyla in a variety of anaerobic digesters (Niya et al. 2024; Treu et al. 2016; Theuerl et al. 2015). Other bacterial phyla that have been identified as important are Pseudomonadota, Chloroflexota, and Actinomycetota (Treu et al. 2016; Niya et al. 2024). The most frequently found archaea are members of the orders Methanobacteriales, Methanomicrobiales, and Methanosarcinales (Harirchi et al. 2022). Even though high overall diversity can be found in anaerobic digesters, the microbial community is often skewed to a few abundant taxa while

the vast majority of taxa are rare (Dueholm et al. 2024; Theuerl et al. 2019). The composition has been found to be mainly affected by the substrate used in AD (Dueholm et al. 2024; Harirchi et al. 2022). Another important factor is the temperature as it has been shown that the microbial community in thermophilic systems is different from that in mesophilic systems (Campanaro et al. 2018; Treu et al. 2016). For example, Bacteroidota have been found at higher abundance in mesophilic than in thermophilic systems (Treu et al. 2016). Furthermore, the inoculum initially used to start up an AD system is important (Camargo et al. 2023; Harirchi et al. 2022).

Studies analysing the metabolic potential of MAGs from anaerobic digesters have shown a funnel-like specialisation of the microbial community along the four stages of the process (Kim et al. 2022; Campanaro et al. 2016). Many MAGs of various phyla could be assigned to the initial stages while the number of assigned MAGs decreased towards methanogenesis, suggesting more specialised MAGs in the later stages (Campanaro et al. 2016). Another study observed that systems where functional redundancy is found are more stable than systems where each task is only fulfilled by a small number of taxa (Cao et al. 2025).

1.4 Thesis aims

Even though there has been increasing research effort in the field of microbial communities in AD in recent years, many of the involved species are still uncharacterised and their functions in the AD process are unknown (Theuerl et al. 2019; Dueholm et al. 2024). Moreover, research has focused on systems that use common substrates such as manure or agricultural feedstocks (Campanaro et al. 2018; Jun et al. 2025) while knowledge about systems that treat industrial wastewaters is limited (Cai et al. 2016). Investigating the differences between these kinds of systems will help to understand the dynamics of microbial communities (Cai et al. 2016). Furthermore, granular biomass in anaerobic digesters has not yet been studied in detail using metagenomic methods, but their structure regarding taxa and their functions could reveal important patterns (Trego et al. 2021). In conclusion, the application of genome-based metagenomic approaches in AD research is still limited and especially systems with less common substrates have not been characterised yet. More thorough research regarding metabolic pathways and the roles of individual microorganisms will provide a better understanding of the AD process which is the basis for improving its efficiency. Therefore, this thesis pursues the following aims.

(1) Determine the microbial community composition in a variety of AD systems by amplicon-based methods: The performance of three reference databases in the taxonomic classification of the microorganisms in this environment will be evaluated. Moreover, taxonomic profiles of anaerobic digesters with substrates such as agro-industrial residues and industrial wastewaters will be compared.

(2) Recover MAGs and conduct a functional analysis of selected systems using metagenomic methods: Based on the results of the first part, a subset of systems will be selected. The aim is to choose systems that are underrepresented in current studies. After metagenomic sequencing using the Oxford Nanopore Technologies platform, MAGs will be recovered by assembly and binning. Key pathways for AD will be defined. These pathways will be linked to the obtained MAGs after functional annotation to obtain information about their potential roles in the investigated systems.

Overall, the aim of this thesis is to investigate the microbial community in various AD systems and to provide insights into understudied systems. The main focus will be on taxonomic profiling of the systems and on the functional characterisation of individual MAGs to better understand their roles and how they interact with each other.

2 Materials and Methods

2.1 Software

The computational results of this work have been achieved using the Life Science Compute Cluster (LiSC) of the University of Vienna. The software that was used within this thesis is listed in Table 1 (recurring tasks), Table 2 (analysis of amplicon dataset), and Table 3 (metagenomic analysis). Databases that were used within this thesis are listed in Table 4.

Table 1: *List of software that was used for recurring tasks in this thesis.*

Software	Version	Reference
R	4.5.1	R Core Team 2025
RStudio	2025.09.1.401	Posit team 2025
ggplot2 (R package)	4.0.0	Wickham 2016
ggh4x (R package)	0.3.1	van den Brand 2025
tidyverse (R package)	2.0.0	Wickham et al. 2019
microViz (R package)	0.12.7	Barnett et al. 2021

Table 2: *List of software that was used for analysis of the amplicon dataset in this thesis.*

Software	Version	Reference
bcl2fastq	2.20.0.422	Illumina 2017
Cutadapt	5.1	Martin 2011
DADA2 (R package)	4.5.1	Callahan, McMurdie, et al. 2016
FastQC	0.12.1	Andrews 2010
MultiQC	1.28	Ewels et al. 2016
USEARCH	11.0.667	Edgar 2010
vegan (R package)	2.7-2	Oksanen et al. 2025

Table 3: *List of software that was used for the metagenomic analysis in this thesis.*

Software	Version	Reference
MinKNOW	25.06.16	Oxford Nanopore Technologies 2025c
dorado	7.11.2	Oxford Nanopore Technologies 2025a
Filtlong	0.3.1	Wick 2025
NanoComp	1.25.6	De Coster and Rademakers 2023
Seqkit2	2.11.0	Shen et al. 2024
MetaPhlAn 4	4.2.4	Blanco-Míguez et al. 2023
SingleM	0.20.3	Woodcroft et al. 2025
metaFlye	2.9.6	Kolmogorov et al. 2020
minimap2	2.30	Li 2018
SAMtools	1.23	Danecek et al. 2021
CoverM	0.7.0	Aroney et al. 2025
medaka	2.1.1	Oxford Nanopore Technologies 2025b
MetaBAT2	2.18	Kang et al. 2019

Table 3 (continued): *List of software that was used for the metagenomic analysis in this thesis.*

Software	Version	Reference
Quast	5.3.0	Gurevich et al. 2013
CheckM2	1.1.0	Chklovski et al. 2023
dRep	3.6.2	Olm et al. 2017
GTDB-Tk	2.6.1	Chaumeil et al. 2022
anvi'o	9	Eren et al. 2020

Table 4: *List of databases that were used within this thesis.*

Database	Version	Reference
MiDAS 5	5.3	Dueholm et al. 2024
SILVA SSU Ref NR 99	138.1, 138.2	Quast et al. 2013
GlobDB	226	Speth et al. 2025
KOfam	v2025-11-07	Aramaki et al. 2020

2.2 Experimental Design

This thesis is made up of two consecutive parts. The overall structure is depicted in Figure 2.

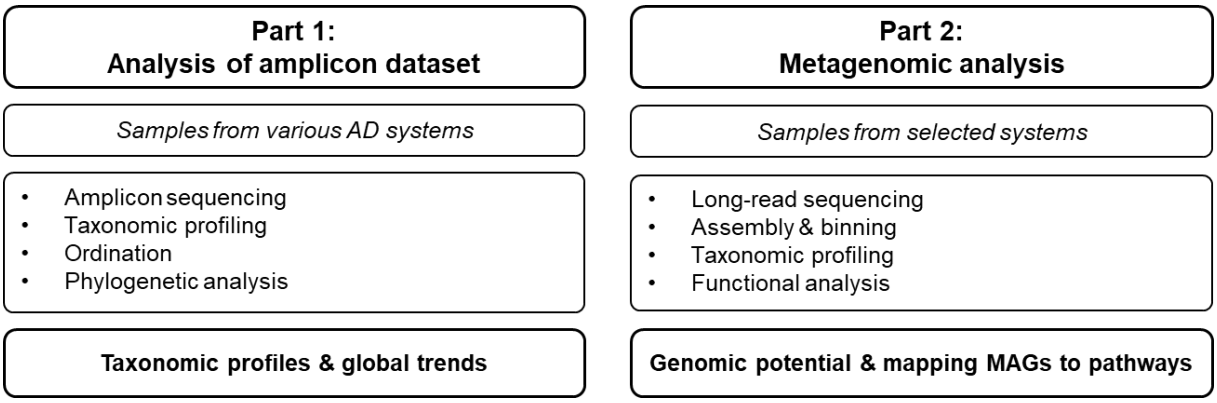

Figure 2: *Structure of the thesis. AD: Anaerobic digestion. MAGs: Metagenome-assembled genomes.*

In the first part, an amplicon sequencing dataset was analysed. The dataset covers a diverse set of samples from a variety of anaerobic systems. Additionally, it was contextualised with a publicly available dataset of comparable systems. The aim of this analysis was to find global patterns in the taxonomic profiles of diverse anaerobic digestion (AD) systems. Based on the results, the most promising systems for an in-depth analysis using metagenomics techniques were selected.

In the second part, a metagenomic analysis of samples from the selected systems was conducted. After long-read sequencing, metagenome-assembled genomes (MAGs) were recovered from the assembled reads. Taxonomic profiles of the sampled reactors were generated, and the MAGs were functionally annotated. Based on these results, the potential roles of MAGs in the process of breaking drown organic matter to biogas were investigated.

2.3 Analysis of amplicon dataset

2.3.1 Dataset description

An existing amplicon sequencing dataset was used to study patterns in the taxonomic profiles of a variety of AD systems. The sampled systems and the sequencing approach are described in this section.

2.3.1.1 Sampled systems

The majority of the analysed samples originates from AD or acidification processes with substrates from agro-industries or the pulp and paper industry. The samples originate from lab, pilot, and full-scale systems. The substrates used in these systems are either industrial wastewaters or residues. Industrial substrates are only used in a small fraction of biogas plants in Austria or Germany. Typical biogas plants, as they can be found at high frequency in this region, use residues from agriculture, such as renewable raw materials (plants) or animal manure (Weinrich and Nelles 2021). Therefore, two more samples were included to be able to specifically find potential differences caused by the substrate type. They are combined in the category 'Others (Ref)'. The sampled systems include typical single-stage systems, where the complete AD process happens in one single reactor, and two-stage systems. Here, the process is separated: In the first reactor (acidification reactor), the biomass is broken down to mainly acids in the hydrolysis and acidification stage. In the second reactor ('methane reactor'), the components of the effluent from the first reactor are further broken down via acetogenesis and methanogenesis, with the main product being methane. An overview of the samples and the sampled systems is given in Table 5. The full metadata table is available in the supplementary material (Table S2).

Table 5: *Description of sampled systems. AD: Anaerobic digestion.*

SYSTEMS THAT TREAT WASTEWATER AND RESIDUES FROM AGRO-INDUSTRIES			
Scale	**Abbreviation**	**Description**	**Nr. of samples**
Full	StWC	AD at a starch factory. The raw materials for the factory are wheat and corn. The factory's wastewater is treated in two CSTRs.	4
Full	ST	AD at a sugar factory. The raw materials for the factory are sugar beets. The factory's wastewater is treated in an EKJ reactor (Kroiss and Svardal 1988).	1
Full	SK	AD at a sugar factory. The raw materials for the factory are sugar beets. Pressed sugar beet pulp, which is a residue in the production, is treated in three CSTRs.	4
Full	StarchC	AD at a starch factory. The raw material for the factory is corn. The factory's wastewater is treated in a two-stage AD, with an acidification reactor and a CSTR methane reactor.	18
Pilot	Pilot	AD of thin stillage in a 300L pilot reactor. The reactor was inoculated with sludge from an agricultural biogas plant and thin stillage from ethanol production was used as substrate for a period of nine months.	18

Table 5 (continued): *Description of sampled systems. AD: Anaerobic digestion.*

Scale	Abbreviation	Description	Nr. of samples
Lab	Conti	AD of various agro-industrial residues in four 5L continuous lab reactors (CSTR) over 8-10 months. The reactors were inoculated with sludge from the previously described 300L pilot reactor. The four reactors were fed with different substrates: CoA: Monodigestion of thin stillage CoB: Codigestion of thin stillage and pressed sugar beet pulp CoC: Monodigestion of small sugar beet parts CoD: Codigestion of sugar beet parts and starch residues	57
SYSTEMS THAT TREAT WASTEWATER FROM PULP AND PAPER INDUSTRY			
Scale	**Abbreviation**	**Description**	**Nr. of samples**
Full	PaP	AD at a pulp and paper factory. The factory's wastewater is treated in a two-stage AD. It is made up of two separate cascades. Cascade 1 has an acidification reactor followed by two EGSB methane reactors and cascade 2 has an acidification reactor followed by two IC methane reactors.	7
OTHERS (REF)			
Scale	**Abbreviation**	**Description**	**Nr. of samples**
Full	Ref_Agri	Anaerobic digester (CSTR) that uses agricultural feedstocks (mixture of maize silage and pig manure) as substrate.	1
Full	Ref_Slaughter	Anaerobic digester (CSTR) that treats residues from a slaughterhouse.	1

2.3.1.2 Sampling and sequencing

Samples were collected over a period of 1.5 years (October 2023 to March 2025). They were processed (DNA extraction and sequencing) in multiple batches. The fresh samples were centrifuged at 2244 × g for 30 min and then stored at -20 °C until further processing. The samples were thawed and DNA was extracted using the FastDNA™ SPIN Kit for Soil (MP Biomedicals Germany GmbH, 37269 Eschwege, Germany) according to the manufacturer's instructions (MP Biomedicals 2021). Extracted DNA was stored at -20 °C until it was sent to Microsynth AG (Balgach, Switzerland) for library preparation and sequencing. Two-step, Nextera-barcoded PCR libraries using the primer pair 515F (5′-GTGYCAGCMGCCGCGGTAA-3′) and 806R (5′-GGACTACNVGGGTWTCTAAT-3′) (Klindworth et al. 2013) were prepared. The PCR libraries were sequenced on an Illumina NovaSeq 6000 platform using an SP 500 cycles kit (Illumina, Inc., San Diego, CA 92122, USA). The produced paired-end reads which passed Illumina's chastity filter were subject to de-multiplexing and trimming of Illumina adaptor residuals using Illumina's bcl2fastq software,

version 2.20.0.422 (Illumina 2017). The resulting FASTQ-files were provided by Microsynth AG and further processed as described below.

2.3.2 Data processing

The DADA2 algorithm (DADA: Divisive Amplicon Denoising Algorithm) (Callahan, McMurdie, et al. 2016) was used to obtain amplicon sequence variants (ASVs) from the provided paired-end fastq-files. The core part of this algorithm corrects for amplicon sequencing errors in order to separate them from biological variation (Callahan, McMurdie, et al. 2016). It first learns an error-model from the data and then infers sample composition by dividing amplicon reads into partitions that are consistent with that error model (Callahan, McMurdie, et al. 2016).

The data was processed using R, version 4.5.1 (R Core Team 2025) and the DADA2 package, version 1.30.0 (Callahan, McMurdie, et al. 2016), following the recommended workflow (Callahan, Sankaran, et al. 2016). The workflow is organised in a series of R-Markdown scripts, which are deposited at https://github.com/lisa94fc/ADMicrobes_supp (directory amplicon-dataset). It is briefly described below. The results were evaluated and compared between sequencing runs, based on the number of reads after each step and on the number of obtained ASVs.

2.3.2.1 Primer trimming

Before starting the DADA2 pipeline, the forward and reverse primers were removed using cutadapt, version 5.1 (Martin 2011) in paired-end mode with the relevant primer sequences. Reads in which the primers were not detected were removed from the output that was used in the consecutive steps. Otherwise, default parameters were used. This was done using the script `01-amplicon-primer-trimming.Rmd`, with following main command.

```
cutadapt -g GTGYCAGCMGCCGCGGTAA -G GGACTACNVGGGTWTCTAAT \
    -o <trimmed-dir>/<sample>_R1_trimmed.fastq.gz \
    -p <trimmed-dir>/<sample>_R2_trimmed.fastq.gz \
    --untrimmed-output <untrimmed-dir>/<sample>_R1_untrimmed.fastq.gz \
    --untrimmed-paired-output <untrimmed-dir>/<sample>_R1_untrimmed.fastq.gz \
    <data-dir>/<fwd_filename>.fastq.gz \
    <data-dir>/<rev_filename>.fastq.gz \
    > <log-dir>/cutadapt_logs/<sample>_cutadapt.log
```

2.3.2.2 Quality filtering and trimming

After removing the primer sequences, the reads were quality-filtered and trimmed using the DADA2 function `filterAndTrim` (see script `02-amplicon-filter-trim.Rmd`). Default parameters were used, except for the expected errors (`maxEE`) the truncation length (`truncLen`). Expected errors were set to 2, as suggested in Callahan, Sankaran, et al. (2016). The truncation lengths (for forward and reverse reads) were chosen to cut the reads where the quality drops, based on the quality profiles and are given in Table 6. They were chosen based on their quality profiles. These were obtained by running FastQC, version 0.12.1 (Andrews 2010) on each sample and then MultiQC, version 1.28 (Ewels et al. 2016) on the results of all samples, separately for forward and reverse reads. This was done separately for each sequencing run to be able to consider the run-specific quality profiles. To obtain the quality profiles, the following commands were used (see `scripts fastqc_multiqc.sh` and `fastqc_multiqc_slurm.sh`).

```
for file in ${dir}/*${pattern_fwd}*.fastq.gz; do
    fastqc ${file} -o ${dir_fastqc_r1}
done
for file in ${dir}/*${pattern_rev}*.fastq.gz; do
```

```
    fastqc ${file} -o ${dir_fastqc_r2}
done

multiqc ${dir_fastqc_r1} -o ${dir_multiqc} -n ${dir_fwd}
multiqc ${dir_fastqc_r2} -o ${dir_multiqc} -n ${dir_rev}
```

Table 6: *Truncation lengths used in the DADA2 step* `filterandTrim`*.*

Sequencing-run ID	truncLen (forward reads)	truncLen (reverse reads)
NGS03108	220	145
NGS03287	220	200
NGS03459	220	180
NGS03614	215	200
NGS03693	220	200
NGS03874	220	220
4401216	210	210

To check the output of this step, a random sample of error profiles before and after `filterAndTrim` was inspected using the DADA2 function `plotQualityProfile`. Both commands are given below.

```
filterAndTrim(
   fwd = <fwd-input-filelist>, filt = <fwd-output-filelist>,
   rev = <rev-input-filelist>, filt.rev = <rev-output-filelist>,
   maxEE = 2,
   truncLen = c(<fwd-truncLen>, <rev-truncLen>),
   trimLeft = 0, trimRight = 0, truncQ = 2, minLen = 50
)
plotQualityProfile(c(
  <fwd-trim-sample>, <fwd-filt-sample>,
  <rev-trim-sample>, <rev-filt-sample>
))
```

2.3.2.3 Error estimation and sample inference

The trimmed reads were dereplicated using the function `derepFastq` and the error model for the sequencing run was set up using the function `learnErrors` on the dereplicated data (see script `03-amplicon-inference.Rmd`). This model was then used in the core DADA2 function `dada`, again on the dereplicated data, to infer sequence variants. For this step, the data of all samples of one sequencing run was pooled (`pool = TRUE`), which increases computation time, but rare variants can be detected more reliably (Callahan, Sankaran, et al. 2016).

After processing the forward and reverse reads separately as described above, they were merged using the function `mergePairs` with default settings. The processed and merged paired-end data was converted to a sequence table using the function `makeSequenceTable` and the tables of all runs were combined to one large sequence table using the function `mergeSequenceTables`.

The commands used for these steps are summarized below.

```
# Dereplication
<dereplicated-files> <- derepFastq(<input-filelist>, verbose = interactive())
```

```
# Error estimation
<dError> <- learnErrors(<dereplicated-files>)

# Inference
<dada-result> <- dada(<dereplicated-files>, err=<dError>, pool = TRUE)

# Merge pairs and make sequence table
merged <- mergePairs(
  dadaF = <fwd-dada-result>, derepF = <fwd-dereplicated-files>,
  dadaR = <rev-dada-result>, derepR = <rev-dereplicated-files>,
  minOverlap = 12, maxMismatch = 0
)
seqtab <- makeSequenceTable(merged)

# Merge runs
<merged-seq-tabs> <- mergeSequenceTables(<seq-tab1>, <seq-tab2>)
```

2.3.2.4 Bimera removal

Bimeras, which are sequencing artifacts, were removed with the function `removeBimeraDenovo`, in the `script 04-amplicon-bimeras.Rmd`.

```
<seq-tabs-no-chimeras> <-
      removeBimeraDenovo(<merged-seq-tabs>, method = "consensus")
```

The output was used as ASV table in further analysis of the sampled anaerobic systems, as described below.

2.3.2.5 Taxonomic classification

The taxonomic classification of the ASVs at the domain to genus level was done using the function `assignTaxonomy`. The function implements the Naïve Bayesian Classifier algorithm, which uses k-mers of the ASV sequences to compute probabilities for the taxa in the classifier database (Wang et al. 2007). A bootstrapping step, which repeatedly uses a fraction of the k-mers for the classification, provides confidence values. Only taxonomic classifications with a confidence value of at least 0.5 (default confidence threshold; option `minBoot`) were written to the output. This comparatively low threshold was chosen because the results are supposed to only serve as first approximation and taxonomic assignments based on short regions such as the V4 region should be treated with caution in all cases. Species were assigned using the function `addSpecies`, which uses an exact-matching algorithm instead, since this is more appropriate for assigning species to short sequences such as the V4 region (Edgar 2017). In the script `05-amplicon-taxonomy.Rmd`, the following commands were used.

```
<classified-tab> <-
  assignTaxonomy(
    <seq-tab>,
    refFasta = <refdb-file>,
    minBoot = 0.5, outputBootstraps = TRUE,
    # only assign down to Genus level
    taxLevels = c("Kingdom", "Phylum", "Class", "Order", "Family", "Genus")
)

<classified-tab-species> <-
  addSpecies(
    <classified-tab>$tax,
    refFasta = <refdb-species-file>,
    allowMultiple = TRUE
  )
```

Three different databases were used for taxonomic classification of the ASVs. The SILVA database (Quast et al. 2013) was used as it as general database that is broadly used. ASVs were classified using two different versions of SILVA SSU Ref NR 99. The most recent version, 138.2, as well as an earlier version, 138.1, were used. The DADA2-compatible versions were downloaded from https://zenodo.org/records/14169026 and https://zenodo.org/records/4587955, respectively. The earlier version was included because the taxonomy used in the third database, MiDAS 5 (Dueholm et al. 2024), is better comparable to this version. MiDAS 5 is specialised on activated sludge and anaerobic digesters and was used with the aim of receiving more information about the microorganisms in the investigated systems than by using a more generic database like SILVA. Version 5.3 was downloaded from the Downloads section at https://midasfieldguide.com/.

The performance of the databases was compared regarding two aspects (see script `06-amplicon-compare-classifiers.Rmd`). Firstly, the number of ASVs with bootstrap values of 100 (i.e. 100 % confidence) were counted for each taxonomic rank and database. Secondly, the number of unclassified ASVs per taxonomic rank and database were counted. Moreover, MiDAS 5-specific placeholder names (in the format “midas_x_y”), which are assigned in MiDAS 5 to taxa that could not be assigned an existing taxonomy, were counted per rank. Based on this, one database was chosen to be used for further data analysis.

2.3.3 Data analysis

2.3.3.1 Taxonomic profiling

From the processing of sequencing data described above, ASV tables with abundances per sample and taxonomic classifications of ASVs were obtained. The abundances were normalized using Total Sum Scaling (TSS), i.e. the abundances were turned into relative abundances by dividing the number of reads per ASV in a sample by the total number of reads in that sample (Swift et al. 2023). The relative abundance tables were used to generate bar plots and bubble plots using ggplot2, version 4.0.0 (Wickham 2016) and ggh4x, version 0.3.1 (van den Brand 2025), see scripts `07-amplicon-table-formatting.Rmd` and `08-amplicon-plots.Rmd`. The data was interpreted and conclusions drawn based on these plots.

2.3.3.2 Merging with MiDAS 5 samples

To put the data generated in the scope of this thesis into a broader, global context, it was merged with samples from the MiDAS project. Their full-length ASV database, which originally focused on activated sludge, was recently supplemented with ASVs that were recovered from anaerobic digester samples (Dueholm et al. 2024). The raw sequences corresponding to these samples were obtained from the NCBI Sequence Read Archive, accession number PRJNA1019951. These sequences will be referred to as ‘MiDAS dataset’ in the following sections. The amplicon dataset described above will be referred to as 'thesis dataset’.

To combine both datasets, it was necessary to ensure that they were processed identically. Due to data availability issues, it was not possible to process the raw sequences from the MiDAS dataset as described in section 2.3.1.2 to then merge ASV tables from the MiDAS dataset and from the thesis dataset. The DADA2 pipeline described above is designed for paired-end data which could not be retrieved from the MiDAS dataset. Therefore, the raw sequences of both datasets were processed together, as described in Dueholm et al. (2024). The workflow will be described in the following sections.

From the MiDAS dataset, 363 samples from 202 digesters were used for this analysis. Duplicates were available for 161 digesters and only one sample was available for the remaining 41 digesters. The remaining sequences were not used for this analysis because

they are not compatible with the thesis dataset. All reverse reads from the thesis dataset, and the single and reverse reads from the MiDAS dataset were used. They were chosen because it was the largest compatible subset. The script that was used for the complete workflow is `merge-thesis-midas.sh` In all of these sequences, the 815R primer sequence was detected at the beginning of the reads. Cutadapt, version 5.1 (Martin 2011) was used to trim the primers from the start and then trim all reads to 226 nucleotides. This was done to obtain reads with a uniform length. The following command was used.

```
cutadapt -g GGACTACHVGGGTWTCTAAT \
    --discard-untrimmed \
    <filename> \
    --length 226 \
    -o <out-dir>/${sample}.fastq \
    > <log-dir>/${sample}_cutadapt.log
```

The trimmed fastq-files were combined and the USEARCH pipeline, version 11.0.667 (Edgar 2010) was used to infer zOTUs. zOTUs are zero-radius Operational Taxonomic Units, i.e. they are similar to ASVs. They are, however, not obtained by applying a statistical error model like ASVs are, but by clustering of the sequences. Briefly, the reads were reverse complemented to match the standard orientation of the 16S rRNA gene, using `usearch -fastx_revcomp`. Then, they were quality filtered with `usearch -fastq_filter` and the option `fastq_maxee 1.0`. Using SeqKit2, version 2.11.0 (Shen et al. 2024), reads shorter than 220 nucleotides (command `seq`) as well as files with less than 8,000 reads (command `stats`) were identified and discarded. The samples were pooled in one fastq-file using `cat`, dereplicated using `usearch -fastx_uniques`, and finally denoised using the core command of USEARCH, `unoise3`, to obtain zOTUs. An abundance table was obtained from the function `usearch -otutab` and used for a multivariate analysis, which is described in the next section. The core commands used for this part are given below.

```
# reverse complement
usearch -fastx_revcomp <file> -fastqout <rev-comp-dir>/${sample}_rc.fastq

# quality filter
usearch -fastq_filter <rev-comp-dir>/${sample}_rc.fastq -fastq_maxee 1.0 \
    -relabel ${sample}. -fastqout <out-dir>/${sample}_filtered.fastq

# pool all samples
for file in filtered-files/*.fastq; do
   cat ${file} >> pooled_samples.fastq
done

# dereplicate
usearch -fastx_uniques pooled_samples.fastq -fastqout uniques.fastq \
    -sizeout -relabel Uniq.

# denoise
usearch -unoise3 uniques.fastq -zotus zotus.fa

# make ASV table (map reads to zOTUs)
usearch -otutab pooled_samples.fastq -zotus zotus.fa -strand plus -otutabout
otutab.txt
```

2.3.3.3 Multivariate analysis

To reveal patterns of similarities of the analysed samples regarding their taxonomic profiles, a multivariate analysis, specifically ordination, was carried out. In ordinations, objects are placed

in an ordination space, and similar objects are closer to each other than dissimilar objects (Ramette 2007). This reveals similarity patterns, but it does not reveal their causes (Ramette 2007). In this specific case, the objects are samples, and the similarity is based on their microbial community profile, i.e. the abundance of ASVs. Correspondence analysis (CA) and non-metric multidimensional scaling (NMDS) were chosen as ordination methods.

The aim of CA is to represent the degrees of correspondence between rows (ASVs) and columns (samples) of a community matrix in an ordination space (Ramette 2007; Paliy and Shankar 2016). A unimodal relationship between ASVs and environmental gradients is assumed, which supports the ecological niche theory (Ramette 2007; Paliy and Shankar 2016). The method was chosen because it is recommended for relative abundances of microbial communities and preferred over linear-based methods such as Principal Components Analysis (PCA) for compositional datasets with many zeroes, such as community matrices (Ramette, 2007; Paliy and Shankar, 2016).

In NMDS, pair-wise distances between samples are ranked, based on a distance matrix (Ramette 2007; Paliy and Shankar 2016). These ranks are then used to map the samples to an ordination space (Ramette 2007; Paliy and Shankar 2016). In contrast to various other ordination methods, NMDS does not assume an underlying distribution and can therefore be applied when the type of relationship between the ASVs and their environment is not known (Paliy and Shankar 2016).

The abundance table that was obtained as described in section 2.3.3.2 and contains abundances for both, the MiDAS and thesis dataset, was investigated using CA and NMDS (see script `Multivariate-analysis-w-MiDAS.Rmd`). It was decided to use both methods to have broader support for any conclusions drawn from the results. CA was selected because it is specific for microbial community data and NMDS was selected because it is a more generic method that is robust to deviations from data distribution assumptions.

The R package vegan, version 2.7-2 (Oksanen et al. 2025) was used for this analysis. The function `cca` was used to compute the CA, which is based on the method described in Legendre & Legendre (2012) and originally introduced in ter Braak (1986). The wrapper function `metaMDS` was used to compute the NMDS. The abundance table was directly passed to `metaMDS`. The function uses `vegdist` to compute a distance matrix from the abundance table, before computing the NMDS. The default distance metric, the Bray Curtis distance, was used.

The site scores were extracted from the ordination results and merged with the sample's metadata, using the following function.

```
get_scores <- function(ord.object) {
  site_scores <- as.data.frame(scores(ord.object, display = "sites")) %>%
    rownames_to_column("filename") %>%
    merge(metadata)

  return(site_scores)
}
```

These scores were then used to visualize the locations of the samples in the ordination space using ggplot2, version 4.0.0 (Wickham 2016). The sample points were coloured by metadata variables using the colour palettes from the microViz package, version 0.12.7 (Barnett et al. 2021).

2.4 Metagenomic analysis

2.4.1 Anaerobic digestion systems and sampling

A fresh set of samples from two industrial AD systems (PaP and StarchC) was taken for a metagenomic analysis. The systems were chosen based on the results of the analysis of a larger amplicon dataset, see section 2.3. Both systems are located at company sites to treat their wastewaters. Both are two-stage systems that separate the process into two parts. The organic matter in the waste streams is broken down to mainly acids in the hydrolysis and acidification stage. This takes place in the acidification reactor. In the methane reactor, the components of the effluent from the acidification reactor are further broken down via acetogenesis and methanogenesis, with the main product being methane.

The system that is abbreviated as 'PaP' within this thesis is located at a pulp and paper factory and includes six reactors that are operated in two parallel lines. The waste stream is separated into two parts of equal composition. One part goes to acidification reactor 1 (sample 'PaP_Acid1') and is then fed to two expanded granular sludge bed (EGSB) reactors (methane reactors; samples 'PaP_EGSB1', 'PaP_EGSB2'). The other part goes to acidification reactor 2 ('PaP_Acid2') and is then fed to two internal circulation (IC) reactors (methane reactors; samples 'PaP_IC3', 'PaP_ IC4'). A schematic representation of the system is shown in Figure 3A.

The system that is abbreviated as 'StarchC within this thesis is located at a starch factory and includes two reactors. The waste stream goes to the acidification reactor (sample 'StarchC_Acid') and then to the methane reactor (sample 'StarchC_Ferm'). In addition, a sample of the effluent ('StarchC_Ferm_O') was taken. A schematic representation of the system is shown in Figure 3B.

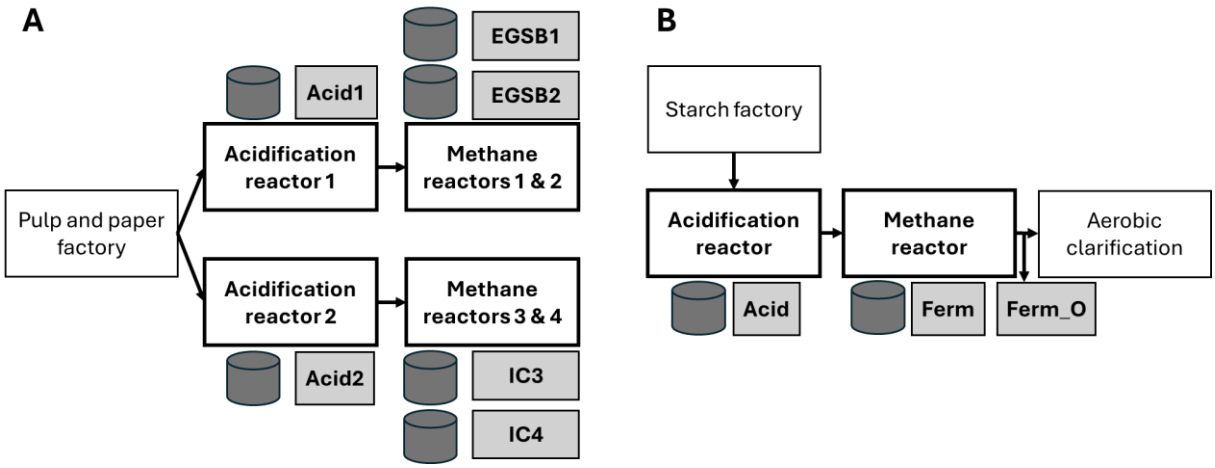

Figure 3: *Schemes of anaerobic digestion systems used for metagenomic analysis. Sample names in this thesis are used as given in the squares with grey background, with a prefix for the system.* ***A:*** *System at pulp and paper factory, abbreviated in this thesis as 'PaP'.* ***B:*** *System at starch factory, abbreviated in this thesis as 'StarchC'.*

In addition, two samples from anaerobic digesters that treat other substrates than industrial wastewaters were included to be able to find specifics of such systems. One sample from a typical agricultural system (sample 'Ref_Agri') and one from an anaerobic digester that treats slaughterhouse waste were included (sample 'Ref_Slaughter').

The samples were provided by the companies. They were frozen upon arrival and thawed and aliquoted as soon as all samples were available. A fraction of each sample was centrifuged at 2244 x g for 30 minutes to separate the liquid fraction from the biomass pellet. All samples (unseparated, liquid fraction, pellet) were frozen again until further processing.

2.4.2 Metadata analysis

2.4.2.1 Process parameters

To evaluate the stability and performance of the selected AD systems, process parameters provided by the companies were used. The efficiency of the systems was estimated using the organic matter content of the influent and effluent of the systems. Metrics to estimate the organic matter content were provided as chemical oxygen demand (COD) by the company responsible for the system 'PaP' and as total organic carbon (TOC) by the company responsible for the system 'StarchC'. The efficiency was determined using the following equation.

$$Efficiency\,[\%] = 100 - \frac{COD\ or\ TOC\ in\ influent}{COD\ or\ TOC\ in\ effluent} * 100$$

Furthermore, the fraction of methane (CH_4) in the biogas was provided by both companies. The biogas is the offgas of the AD system. A large fraction of methane in the biogas is desired because it is the main energy carrier in the biogas.

All datasets were provided for the time period 12 days before to 12 days after the sampling date.

2.4.2.2 Physicochemical analysis

Besides process parameters that were provided by the companies, physicochemical parameters of the samples were determined and provided additional context that facilitated the biological interpretation of sequencing results.

The pH was determined using a FiveEasy pH meter (Mettler-Toledo GmbH, 1230 Vienna, Austria) with an LE422 pH electrode (Mettler-Toledo GmbH, 1230 Vienna, Austria).

After cleaning up the liquid fractions of the samples by Carrez clarification (Carrez 1908), the concentration of volatile fatty acids (VFAs) and alcohols was determined using High Performance Liquid Chromatography (HPLC). A 1260 Infinity II HPLC system was used (column: Coregel ION 300-S/N 13729504, vial sampler: G7129A 1260, iso pump: G7110B 1260, column oven: 2x G7116A 1260 MCT, detector: G7162A 1260 RID; Agilent, Santa Clara, CA 95051, United States). The injection volume was 40 µL and the system was run at a flow rate of 0.325 mL/min with a pressure of 63.49 bar and at 45 °C. The eluent was 0.005 M H_2SO_4.

Ion chromatography was used to determine the concentration of anions (sulphate, nitrate, chloride, phosphate) in the liquid fractions of the samples. A Dionex ICS-900 system (AS-DV autosampler, conductivity detector, guard column: AG14A 4 x 50 mm, analytical column: AS14A 4 x 250 mm; Thermo Scientific, Waltham, MA 02451, USA). The eluent was a solution of 8 mM Na_2CO_3 and 1 mM $NaHCO_3$ an isocratic flow rate of 1 mL/min was used. A 57 mM H2SO4 solution was used to regenerate the ACRS 500 suppressor at the same flow rate.

2.4.3 Sequencing

The frozen pellets were submitted to the Joint Microbiome Facility (JMF) for DNA extraction and sequencing. The steps in section 2.4.3.1 - 2.4.3.3 were carried out by the JMF.

2.4.3.1 DNA extraction

DNA was extracted using the FastDNA™ SPIN Kit for Soil (MP Biomedicals Germany GmbH, 37269 Eschwege, Germany) according to the manufacturer's instructions (MP Biomedicals

2021). Additionally, DNA was extracted using the DNeasy PowerSoil Pro extraction kit (Qiagen GmbH, 40724 Hilden, Germany) following the manufacturer's instructions. The fragment lengths of the resulting DNA extracts were analysed using a TapeStation Instrument (Agilent, Santa Clara, CA 95051, United States).

2.4.3.2 Amplicon sequencing

Amplicon sequencing of the V4 region of the 16S rRNA gene of the DNA was conducted. The DNA extract obtained from using the FastDNA™ SPIN Kit for Soil was used. The aim was to analyse the stability of the systems by comparing the results to those obtained from previous samples (see section 3.2). Moreover, the complexity of the community was evaluated based on the results of this sequencing run to determine the approximate necessary sequencing depth for long-read sequencing (see section 2.4.3.3). Library preparation and sequencing was done using a dual barcoding method (Pjevac et al. 2021) with adapted V4 primers V4-EXT (Hu et al. 2025). This primer pair became available and was used because it has been shown that it covers more of the biological diversity than the standard V4 primer pair (Hu et al. 2025).

2.4.3.3 Long-read metagenomic sequencing

In the past, most metagenomic studies had been conducted using short but highly accurate reads, produced by the Illumina platform (Latorre-Pérez et al. 2020). However, the *de novo* assembly of such reads is mainly complicated by repetitive regions that are longer than the reads, which leads to fragmented genome assemblies (Latorre-Pérez et al. 2020). This limitation can be overcome by using longer reads. Technologies that yield such long-reads have been developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) (Trigodet et al. 2026; Latorre-Pérez et al. 2020). This comes at a cost of higher error rates (Latorre-Pérez et al. 2020). However, read accuracy has been increasing over the last years (Trigodet et al. 2026). Additionally, ONT provides cost-effective platforms for real-time sequencing which is why they have become increasingly popular (Latorre-Pérez et al. 2020). Thus, the ONT PromethION platform was used for long-read sequencing for metagenomic analyses of the selected samples.

The DNA that was extracted with the PowerSoil Pro extraction kit was used. Samples were prepared for sequencing using the SQK-NBD114.96 (Oxford Nanopore Technologies, Oxford OX4 4DQ, United Kingdom). About 53fmol of a 26Kb library was loaded on an R10.4.1 flowcell (FLO-PRO114M, Oxford Nanopore Technologies, Oxford OX4 4DQ, United Kingdom) and sequenced for 72h on a PromethION P2 solo (Oxford Nanopore Technologies, Oxford OX4 4DQ, United Kingdom) using MinKNOW, version 25.06.16 (Oxford Nanopore Technologies 2025c). Raw signal was live basecalled using the super accuracy model (model: r1041_e82_400bps_sup_v5.2.0) using dorado, version 7.11.2 (Oxford Nanopore Technologies 2025a). Quality filtering of the reads was part of the basecalling process, with the criteria Phred score (Q) ≥ 15 and minimum length of 1,000 nucleotides. At the same time, adapters and barcodes were removed. The resulting reads were provided as FASTQ files by the JMF.

Three of the samples (all from the StarchC system) were sequenced in another run, after an additional clean-up step using the DNA Clean & Concentrator kit (Zymo Research, Irvine, CA 92614, United States) kit to increase the sequencing depth.

2.4.4 Data analysis

The analysis of the metagenomic dataset was conducted using multiple scripts. These are deposited at https://github.com/lisa94fc/ADMicrobes_supp (directory ont-dataset). The main command of each step is given below.

2.4.4.1 Amplicon sequencing data

The raw amplicon sequencing results were processed by the JMF. Their workflow using DADA2 (Callahan, McMurdie, et al. 2016; Callahan, Sankaran, et al. 2016) was similar to that described in section 2.3.2, but it was adapted to be compatible with the dual barcoding sequencing method. An ASV table was provided by the JMF, which was then taxonomically classified using the MiDAS 5 database, version 5.3 (Dueholm et al. 2024), as described in section 2.3.2.5. Taxonomic profiles were obtained as described in section 2.3.3.1.

2.4.4.2 Metagenomics data

Quality control and filtering

The worst 10% of the reads were removed using Filtlong, version 0.3.1 (Wick 2025). NanoComp, version 1.25.6 (De Coster and Rademakers 2023) and SeqKit2, version 2.11.0 (Shen et al. 2024) were used for quality control of the raw and filtered reads. The following commands were used (see scripts `01_filtlong.sh` and `02_QC_ont.sh`).

```
# filter
filtlong --keep_percent 90 <file>

# QC
seqkit stats <path-to-data>/*.fastq.gz -Ta > seqkit_stats.tsv
NanoComp --tsv_stats --fastq <path-to-data>/*.fastq.gz --outdir <out-dir>
```

Read-based taxonomic profiling

Before assembly, read-based taxonomic profiles were investigated. Taxonomic profiles can be generated from metagenomic sequencing reads by various methods. Unlike classifiers which classify every read by comparing it to representative sequences in a database, taxonomic profilers use marker genes to generate taxonomic profiles. Since only marker genes are used for classification, a large fraction of the reads remains unclassified. The relative abundances of taxa in the metagenome are thus estimated based on the results and provided as taxonomic profiles. Besides the choice of the profiler itself, the choice of the reference database is crucial for performance.

MetaPhlAn 4 (Blanco-Míguez et al. 2023) classifies reads based on species-level genome bins (SGBs). They were defined from a collection of microbial genomes (from isolates) and MAGs by the developers. After annotation and identification of core genes, each SGB got assigned a set of unique marker genes, specific for that SGB. Additionally, a taxonomic label was assigned to each SGB, based on the National Center for Biotechnology Information (NCBI) taxonomy. This label can either be defined from the taxonomic rank of kingdom down to species, or it can be undefined at certain taxonomic ranks. Metagenomic long-reads are mapped against the SGB marker gene database using minimap2 (Li 2018), assigned to an SGB, and given the taxonomic classification of the corresponding SGB. From this, the coverage of each clade is inferred and normalised, yielding relative abundances. Reads that cannot be assigned to an SGB are not classified at any taxonomic level and summarized in ‘unclassified reads’. (Blanco-Míguez et al. 2023)

Even though SingleM (Woodcroft et al. 2025) is based on marker genes as well, the method differs from the MetaPhlAn algorithm. The marker genes in SingleM are not specific for a taxon but conserved single-copy marker genes are used. Reads are matched to conserved regions in the marker genes using hidden Markov models (HMMs) to cluster them into operational taxonomic units (OTUs). The OTUs are matched to species representatives from the Genome Taxonomy Database (GTDB) (Parks et al. 2022) to assign species if possible. Genus-level or higher taxonomy is assigned to the remaining OTUs by alignment in amino acid space using DIAMOND blastx (Buchfink et al. 2021). Finally, the classifications of all marker genes are

condensed into a consensus taxonomic profile. In SingleM, custom databases, organised in metapackages, can be used which enables users to make it more specific for the environment of interest. (Woodcroft et al. 2025)

For cross-validation purposes, both tools, MetaPhlAn 4 and SingleM, were used for taxonomic profiling of the metagenomic reads. MetaPhlAn 4, version 4.2.4 (Blanco-Míguez et al. 2023) was used with the database mpa_vJan25_CHOCOPhlAnSGB_202503. The taxonomic profiles of all samples were merged to an abundance table of all samples, using the utility script `merge_metaphlan_tables.py`. The following commands were used (see script `03_metaphlan_batch.sh`).

```
# run MetaPhlAn 4 for every sample
metaphlan <read-file>.fastq.gz --input_type fastq \
    --mapout <outdir>/mappings/<sample>.minimap2.bz2 \
    --long_reads --nproc 5 -o <outdir>/profiles/<sample>-profile.txt \
    --db_dir <db-dir> --verbose

# merge tables
merge_metaphlan_tables.py <outdir>/profiles/*.txt \
    > <outdir>/merged_abundance_table.txt
```

SingleM, version 0.20.3 (Woodcroft et al. 2025) was used with the metapackage linked to the GlobDB taxonomy, release 226 (Speth et al. 2025), which is available on the GlobDB website (https://fileshare.lisc.univie.ac.at/globdb/globdb_r226/taxonomic_profiling/globdb_r226_SingleM_metapackage.tar.gz). The GlobDB is a comprehensive database of species-dereplicated microbial genomes from 14 genomic catalogues. The taxonomic profiles obtained with the command `pipe` were transformed into a species-by-site relative abundance table with the command `summarise`. The script `04_singlem.sh` was used and the main commands are given below.

```
# run SingleM for every sample
singlem pipe -1 <read-file>.fastq.gz -p <outdir>/<sample>_profile.tsv \
    --otu-table <outdir>/<sample>_otu.tsv

# merge tables
singlem summarise --input-taxonomic-profile <outdir>/*_profile.tsv \
    --output-species-by-site-relative-abundance-prefix <outdir>/singlem-globdb
```

The results of both tools were compared regarding the fraction of unclassified reads. The results of the tool with less unclassified reads were visualised and compared with the results of amplicon sequencing. This was done to find out if the same taxa are identified by amplicon and metagenomic sequencing. Even if primers are designed to universally detect bacteria and archaea, they do not bind to the DNA of all taxa with the same affinity due to mismatches (Apprill et al. 2015). Therefore, the results are biased and some taxa are not detected in amplicon sequencing approaches (Apprill et al. 2015). In contrast, metagenomic long-read sequencing is independent of DNA amplification and not as selective for specific taxonomic groups, although DNA extraction biases can apply. On the other hand, because the complete genome is targeted instead of a short region, the sequencing depth must be much larger in metagenomic than in amplicon sequencing to obtain similar taxonomic resolution.

Assembly

To recover metagenome-assembled genomes (MAGs) from the sequencing data, the reads were first assembled into contigs. Contigs are contiguous sequences that are obtained by finding overlaps between the reads and stitching them together. Assembly algorithms typically construct graphs from the sequencing reads and compute assemblies by finding paths through

them. Long-read assemblers must be able to deal with different kinds of challenges than short-read assemblers and their development is still an active area of research (Kolmogorov et al. 2020; Trigodet et al. 2026). Short-read assemblies are usually based on de Bruijn graphs, which are constructed by using all k-mers in the sequencing reads (Kolmogorov et al. 2019). This graph structure relies on exact overlaps of k-mers and is thus unsuitable for error-prone long-reads. The long-read assembler Flye (Kolmogorov et al. 2019) uses repeat graphs instead, which can deal with approximate matches. Additional challenges arise for metagenome assemblies, compared to single-genome assemblies. The uneven bacterial composition and coverage, as well as highly similar genomes make the process more complex. This is addressed by metaFlye (Kolmogorov et al. 2020), a version of Flye specifically developed for metagenomes.

Draft assemblies were computed using metaFlye, version 2.9.6 (Kolmogorov et al. 2020). At first, single-sample assemblies were computed for each sample. To maximise the sequencing depth available for the assembly and to be able to use differential abundance information for binning in the next step, co-assemblies were considered. In co-assemblies, the metagenomic reads of multiple samples are combined for the assembly. The same command was used for both approaches, see script `05_flye.sh`.

```
# run flye in metagenomics mode for each sample/set of samples
flye --nano-hq <read-file1>.fastq.gz [<read-file2>.fastq.gz ...] --meta \
    --out-dir <outdir>/<assembly-name>
```

The draft assemblies obtained from metaFlye were polished using medaka, version 2.1.1. (Oxford Nanopore Technologies 2025b). Assembly polishing is used to correct any errors in the sequence that were introduced during the assembly, by mapping the reads back to it and finding a consensus. Medaka was developed specifically for polishing assemblies that are based on ONT reads and is recommended by the developers to be used in combination with Flye (Oxford Nanopore Technologies 2025b). It maps the sequencing reads to the draft assembly, using minimap2 (Li 2018) and saves a ‘base counts table’ in so-called ‘pileups’. These pileups are used as input for a neural network which infers a consensus sequence. It scores all possible bases at each position. The models for the neural networks are specific for the chemistry used during sequencing. (‘Introduction_to_how_ONT’s_medaka_works’, n.d.; Oxford Nanopore Technologies 2024)

The main commands for polishing the assemblies are given below (see script `08_medaka.sh`).

```
# for co-assemblies, concatenate the reads of all samples
cat <read-file1>.fastq.gz <read-file2>.fastq.gz [<read-file3>.fastq.gz ...] \
    > reads_combined.fastq.gz

# run medaka for each assembly
medaka_consensus -i <read-file>.fastq.gz -d <assembly-file>.fa \
    -o <outdir>/<assembly-name> -m r1041_e82_400bps_sup_v5.2.0
```

In co-assemblies, information regarding variations between the samples is lost to some extent. Thus, co-assemblies should only be used if it can be assumed that there is little to no variation in the genomes across the samples. To determine if this is a valid assumption by estimating the fraction of reads of each sample that can be found in each assembly, the reads of all samples were mapped to all single-sample assemblies first. The reads were mapped back to the assemblies using minimap2, version 2.30 (Li 2018). Using SAMtools, version 1.23 (Danecek et al. 2021), the resulting SAM-files were filtered to only keep primary alignments, and sorted. CoverM, version 0.7.0 (Aroney et al. 2025) is a tool that calculates coverage statistics for contigs or genomes. It was used in `contigs` mode to determine the coverage of

the contigs in an assembly across all samples, based on the pre-calculated alignments. The main commands given below were run separately for each single-sample assembly (see scripts `06_map_reads_back.sh` and `07_coverm_contig.sh`). For this analysis, no threshold for a minimum identity as used (i.e. `--min-read-percent-identity = 0`).

```
for READS in <read-files-dir>/*.fastq.gz; do
    # map reads to the assembly
    minimap2 -ax map-ont <assembly-file>.fasta ${READS} > <sam-file>.sam

    # sort SAM-file
    samtools view -b <sam-file>.sam | samtools sort > <bam-dir>/<sample-name>.bam
done

# determine fraction of mapped reads for all BAM files
coverm contig --bam-files <bam-dir>/*.bam \
    -o <outdir>/<sample-name>_<percent-identity>.txt \
    --min-read-percent-identity <percent-identity>
```

The results were visualised using ggplot2, version 4.0.0 (Wickham 2016). Based on the results and on the design of the AD systems, samples to be combined for co-assemblies were chosen.

Binning

After assembling metagenomic reads to contigs, they are separated into bins. The aim is to obtain one complete genome per bin.

Overall, six assemblies were used for binning. Of these, three are co-assemblies. In short, samples from acidification reactors were combined within a system and samples from methane reactors were combined within a system. Table 7 lists the assemblies that were used for binning, and the samples that were used to compute the assemblies.

Table 7: *Assemblies that were used for binning. The second column specifies the samples that were used to compute the assembly. The third column specifies it is a co-assembly or a single-sample assembly.*

Assembly name	Samples	Co-assembly
PaP_acids	PaP_Acid1 PaP_Acid2	Yes
PaP_fermenters	PaP_EGSB1 PaP_EGSB2 PaP_IC3 PaP_IC4	Yes
StarchC_Acid	StarchC_Acid	No
StarchC_fermenter	StarchC_Ferm StarchC_Ferm_O	Yes
Ref_Agri	Ref_Agri	No
Ref_Slaughter	Ref_Slaughter	No

Binning algorithms use differential coverage or structural information of the sequences (such as tetranucleotide frequencies) or both to group contigs into bins. MetaBAT2, version 2.18 (Kang et al. 2019) was used for binning of the assemblies. As the name suggests (MetaBAT: Metagenome Binning with Abundance and Tetra-nucleotide frequencies), this tool uses both characteristics. It calculates pairwise distances for all contig pairs, based on abundances and on tetranucleotide frequencies, and then clusters the contigs by a graph-based method.

Tetranucleotide frequencies are obtained within the main algorithm. Differential coverage information was obtained as follows, before running the main algorithm.

BAM-files obtained from mapping the reads of all samples to the assemblies, as described above (section 'Assembly') were used as input for the utility script `jgi_summarize_bam_contig_depths`, which is part of MetaBAT2. This script extracts coverages of the contigs in all samples. A threshold of 85 % was used for the minimum end-to-end identity of the alignments to be used in the coverage calculation (parameter `--percentIdentity`). This threshold was identified by mapping the reads back to the assemblies at various thresholds, as described above (section 'Assembly'). Otherwise, default parameters were used. The result was used as input in the main MetaBAT2 command. This was done separately for each assembly. The main commands of this process are given below, as used in the scripts `09_jgi_summarize.sh` and `10_metabat.sh`.

```
# calculate depths
jgi_summarize_bam_contig_depths --percentIdentity 85 \
    --outputDepth <depth-file>.tsv <bamfiles-dir>/*.bam

# run MetaBAT2 for every assembly
metabat2 -i <assembly-dir>/<assembly-name>.fasta \
    -a <depth-file>.tsv -o <outdir>/<assembly-name>_bin
```

Quality control

Quast, version 5.3.0 (Gurevich et al. 2013) was used to obtain basic statistics of the assemblies. The quality of the assemblies was evaluated based on the Quast reports. Default parameters were used in the following command (see script `11_quast.sh`).

```
quast.py --output-dir <outdir> <assembly-dir>/*.fa.gz
```

In addition to using basic statistics which were obtained directly from binning with MetaBAT2, the quality of the bins was evaluated based on their completeness and contamination. As the name suggests, the completeness score is an estimation of the completeness of the genome represented by the bin. The contamination score is an estimation of the presence of sequences that belong to a different genome. To obtain both scores, CheckM2, version 1.1.0 (Chklovski et al. 2023) was used. In contrast to most quality assessment tools for bins, CheckM2 does not use single-copy marker genes to estimate completeness and contamination. It was developed to avoid the introduction of biases which leads to inaccurate results for novel lineages and for reduced genomes. CheckM2 uses machine learning methods which take various factors into account, such as multi-copy genes, biological pathways, amino acid counts, and the number of coding sequences. (Chklovski et al. 2023)

The following command was used for each assembly (see script `12_checkm2.sh`).

```
checkm2 predict --input <bin-dir> -x .fa --output-directory <outdir>
```

To select genomes as species representatives, the GTDB (Parks et al. 2022) uses the equation `completeness-5*contamination > 50` (see: https://gtdb.ecogenomic.org/stats/r226#quality-of-gtdb-representative-genomes). This equation takes both scores into account. The same threshold was applied in this thesis to identify and remove low-quality bins.

The remaining bins were dereplicated at an average nucleotide identity (ANI) of 95 %. Sets of bins with ANI ≥ 95 %, over a minimum aligned fraction of 10 %, were identified using dRep, version 3.6.2 (Olm et al. 2017) with default parameters. dRep does pairwise comparisons of all bins using a fast but inaccurate algorithm (MASH) in the first step to obtain similarity

estimations for all bin pairs. In the second step, only pairs with high similarity estimations are compared by using a more accurate algorithm (fastANI) to obtain more accurate ANI scores. The following command was used.

```
dRep compare <outdir> -g <bin-dir>
```

The output was manually inspected and the bin with the highest quality score in each set, as it was defined above, was identified. The other bins in the set were removed from the final dataset.

The so obtained final set of metagenome-assembled genomes (MAGs) was used for further analysis.

Taxonomic profiling

The Genome Taxonomy Database Toolkit (GTDB-Tk), version 2.6.1 (Chaumeil et al. 2022) was used to taxonomically classify the obtained MAGs. GTDB-Tk classifies genomes by phylogenetic placement in a reference tree. After gene calling using Prodigal (Hyatt et al. 2010), bacterial and archaeal marker genes are identified using HMMER (Eddy 2011) to assign the MAGs to a domain (bacteria or archaea). They are then placed into domain-specific GTDB reference trees for taxonomic classification using pplacer (Matsen et al. 2010). In GTDB-Tk v2, the placement is done in two steps. The MAG is first placed into a tree with family-level representatives and then into a class-level subtree with species representatives. The relative evolutionary distance (RED) to reference genomes in the tree is used to resolve ambiguous classifications and the ANI is used to assign species level taxonomy. (Chaumeil et al. 2020, 2022)

The `classify_wf` workflow with default parameters was used with the following command (see script `13_gtdbtk.sh`).

```
gtdbtk classify_wf --genome_dir <bin-dir> --out_dir <outdir> -x fa
```

CoverM, version 0.7.0 (Aroney et al. 2025) was used to estimate the relative abundances of MAGs in the analysed samples (`genome` mode). The metagenomic read files were directly used as input for CoverM which first aligns them to the provided MAGs using minimap2 (flag `-p minimap2-ont`) (Li 2018) and then uses the alignments for calculating several metrics, including the relative abundance of MAGs in each sample. The following command was used in the script `14_coverm_genome.sh`.

```
coverm genome --single <reads-dir> -d <bin-dir> -x fa -p minimap2-ont \
    --bam-file-cache-directory <bam-dir> --discard-unmapped \
    -o <outdir>/mag_abundance_coverm.tsv
```

The relative abundances were visualised using ggplot2, version 4.0.0 (Wickham 2016) and ggh4x, version 0.3.1 (van den Brand 2025).

Functional annotation

For the functional annotation of the MAGs, anvi'o, version 9 (Eren et al. 2020) was used. Anvi'o is a community-driven open software ecosystem for 'omics data analyses. It contains programs that can be combined in a modular way to fulfil various tasks. The workflow used for this analysis is based on the tutorial available at the URL https://anvio.org/tutorials/fmt-mag-metabolism.

The functional analysis of the obtained MAGs is based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000; Kanehisa et al. 2014, 2017). KEGG is a database that enables the biological interpretation of genomes by providing multiple resources.

The KEGG orthology (KO) database defines KOs as functional orthologs of genes and proteins. Each KO is assigned a KO identifier, or K number. The database additionally contains Enzyme Commission (EC) numbers which classify enzymes. KEGG pathway maps provide networks of metabolic reactions, combining information about molecules and enzymes. The KEGG MODULE database contains modules that specify combinations of enzymes that are needed for specific metabolic pathways.

The data required by anvi'o programs that work with KEGG was downloaded (snapshot version `v2025-11-07`) and set up using the program `anvi-setup-kegg-data`. This program sets up profile hidden Markov models (pHMMs) from the KOfam database (Aramaki et al. 2020) which is a database of KOs. The program was run using the following command.

```
anvi-setup-kegg-data --kegg-data-dir anvio/module-dbs/kegg
```

After setting up the KEGG data, the MAGs were imported into anvi'o. To be able to import them, the headers in the FASTA files were uniformly formatted and special characters and spaces were removed using SeqKit2, version 2.11.0 (Shen et al. 2024) as follows.

```
for FILE in <bin-dir>/*.fasta; do
        FILE_PATH=${FILE%.fasta}
        seqkit replace ${FILE} -p "\s.+" > tmp1.fasta
        seqkit replace tmp1.fasta -p "\." > tmp2.fasta
        seqkit replace tmp2.fasta -p ".fa" > ${FILE_PATH}.fa
        rm tmp1.fasta
        rm tmp2.fasta
done
```

Each MAG was imported into anvi'o by generating an anvi'o contigs database from the FASTA file, using the program `anvi-gen-contigs-database`. A contigs database is the basis for all further analyses in anvi'o and stores information about the underlying MAG that is obtained by running anvi'o programs. Most importantly, the program identifies open reading frames in the MAG using pyrodigal-gv, version 0.3.2 (Hyatt et al. 2010; Larralde 2022) with additional models for viruses (Camargo et al. 2024).

The program `anvi-run-kegg-kofams` was used to annotate the MAGs in the contigs databases using the KOfam database (Aramaki et al. 2020). The program annotates the contigs database with pHMM hits from KOfam by checking all gene calls against the previously set up pHMMs using HMMER (Eddy 2011). It eliminates weak hits based on thresholds provided by KEGG, but applies a heuristic that relaxes this threshold in order to keep valid hits that would otherwise be removed. Finally, the program stores the KO hits as well as information about the modules they are associated with.

The commands for setting up a contigs database and populating it with general HMM and KOfam hits are given below (see script `15_anvio_kegg.sh`).

```
anvi-gen-contigs-database -f <bin-file>.fa -o <outdir>/<bin-name>/contigs.db \
    -n '<bin-name>'

anvi-run-kegg-kofams -c <outdir>/<bin-name>/contigs.db \
    --kegg-data-dir anvio/module-dbs/kegg
```

After annotating all MAGs, the program `anvi-estimate-metabolism` (Veseli et al. 2025) was used to collect information from the results. This program can be run on multiple contigs databases at the same time. It determines the completeness of pathways in each of the provided genomes, as defined in the KEGG modules and based on the KO hits. Additionally, it collects and outputs all KO hits of all input contigs databases. When adding the flags

`--matrix-format` and `--include-metadata`, the output is given in matrix format and metadata such as descriptions of the KOs and associated EC numbers are added to the matrix.

The program was run on all annotated MAGs at the same time by providing a list of paths to their contigs databases. The commands used for creating the path list and running the program are given below.

```
# create a list of paths to contigs dbs for all MAGs
anvi-script-gen-genomes-file --input-dir <contigs-dbs-dir> \
    --output-file my-genomes.txt --include-subdirs

# run anvi-estimate-metabolism
anvi-estimate-metabolism -e my-genomes.txt --kegg-data-dir anvio/module-dbs/kegg/ \
    --matrix-format -O estimate-metabolism --include-metadata
```

The main output of the program that was used for further data analysis is a matrix with hit numbers for KOs in the MAGs (estimate-metabolism-KEGG_ID_hits-MATRIX.txt).

Functional analysis

Selected key pathways for AD were investigated. They were manually defined using multiple resources. If available, the pathway description in Heider (2014) was used as basis. The metabolites and reactions of the pathways were sketched out and KEGG resources were used to assign EC numbers to every reaction. This was mainly done using KEGG carbohydrate and energy metabolism pathway maps. The reactions of the pathways of interest were identified by the participating molecules in the KEGG pathway maps. EC numbers associated with catalysing them were obtained from KEGG and added to the sketch. The sketches were used to define the reactions of the pathway in a spreadsheet by substrate, product, and EC numbers. Pathways for butyrate and propionate oxidation were added to the spreadsheet, as defined in Hao et al. (2020). Besides pathways, the formation of acetate and ethanol was investigated in a more general manner. Reactions associated with the formation of these substances were retrieved from KEGG pathway maps and associated EC numbers were added to the spreadsheet and used to for the investigation. The spreadsheet with pathway definitions can be found in the supplementary material (Table S18). The investigated pathways are lactic acid fermentations (heterofermentative, Bifidobacterium fermentation, homofermentative), propionate metabolism (methylmalonyl and acryloyl pathway, oxidation), butyrate oxidation, acetate formation (from acetyl-CoA − direct, via acetaldehyde, via acetyl-P; from lactate), ethanol formation, acetogenesis, methanogenesis (methylotrophic, acetoclastic, hydrogenotrophic), sulphate reduction.

KO annotations of the MAGs that were obtained from anvi'o as described above were used to identify MAGs that encode the enzymes for the investigated pathways. The relevant K numbers were obtained from the spreadsheet with pathway definitions. In addition, MAGs associated with formate formation were identified by manually looking for formate dehydrogenases using the `description` field in the KO annotation file obtained from anvi'o. Using R, version 4.5.1 (R Core Team 2025) and RStudio, version 2025.09.1.401 (Posit team 2025), MAGs with hits for the enzymes were retrieved from the dataset. The hits per MAG and enzyme were visualised in heatmaps using ggplot2, version 4.0.0 (Wickham 2016). Based on these, the end products of the respective pathways were assigned to the MAGs. A final heatmap that visualises the investigated metabolites and the assigned MAGs (only those with abundance >1 % in at least one sample) was created using ggplot2, version 4.0.0 (Wickham 2016).

Based on the individual pathways, an overall pathway map that connects all main metabolites was drawn. This map separates the pathways into what is assumed to take place in the acidification reactor and what is assumed to take place in the methane reactor. The metabolites which were assigned to MAGs were assigned to arrows in the pathway map. In addition, the

phyla associated to each of the metabolites were visualised in bar plots using ggplot2, version 4.0.0 (Wickham 2016). One bar plot per metabolite and reactor was made, using the relative abundances of the MAGs. They were visualised in combination with the pathway map to obtain information about the potential roles of the identified MAGs in AD systems by mapping taxa to pathways.

The assignment of MAGs to metabolites were done using the script `Pathway_mapping.Rmd` and the visualisations were done using the script `Pathways_plots.Rmd`.

3 Results

In this thesis, the microbial communities in anaerobic digestion (AD) systems are analysed. The microbial community composition in a variety of AD systems is determined by amplicon-sequencing based methods. Moreover, a functional analysis of selected systems using metagenomic methods is conducted.

3.1 Taxonomic nomenclature

The databases that were used for taxonomic classification in this thesis do not use uniform taxonomy. After the rank of phylum was added to the International Code of Nomenclature of Prokaryotes in 2023, major changes in taxonomy nomenclature were adopted to follow the rule that phylum names should have the suffix -ota (Oren et al. 2023). The old and new phylum names as updated by the NCBI in 2023 are listed at the URL https://ftp.ncbi.nih.gov/pub/taxonomy/Major_taxonomic_updates_2023.txt.

The database SILVA SSU Ref NR 99 SSU, version 138.1 (Quast et al. 2013) was released before the update and thus uses the previous nomenclature for phyla. The database MiDAS 5, version 5.3 (Dueholm et al. 2024) partially uses the previous nomenclature. The databases SILVA SSU Ref NR 99 SSU, version 138.2 (Quast et al. 2013), GlobDB, release 226 (Speth et al. 2025), and the underlying database in MetaPhlAn 4 use the updated nomenclature.

In this thesis, the updated nomenclature will mainly be used. However, when presenting results that were obtained with databases that use the previous nomenclature it is used in figures, but both names will be used in the text for clarification. The colours in all figures are chosen in a uniform way to represent the same phylum, independent of the used nomenclature.

3.2 Amplicon dataset

3.2.1 Data processing

To study patterns in the taxonomic profiles of a number of AD systems, an amplicon sequencing dataset was analysed. The dataset comprises samples from seven AD systems, treating wastewater or residues from the pulp and paper or agro-industry. Laboratory-, pilot-, and full-scale systems are represented by the dataset. In addition, one sample from a full-scale system treating slaughterhouse residues and one from a full-scale system that uses agricultural feedstocks as substrate are part of the dataset.

An amplicon dataset, obtained from sequencing the V4 region of the 16S rRNA gene of the DNA from multiple samples from a variety of anaerobic digesters, was processed using DADA2 to obtain an ASV table. The average, minimum, and maximum number of reads per sample obtained in each sequencing run is given in Table 8.

The fraction of retained reads per sample for all analysed sequencing runs, tracked along the processing pipeline are presented in Table 9 (averages per run) and Figure 4 (values per sample). The full table of numbers of reads per sample in the raw data and after each processing step is provided in the supplementary material (Table S3).

Table 8: *Average, minimum, and maximum number of reads per sample in each amplicon sequencing run.*

		Nr. of reads per sample		
Run-ID	**Nr. of samples**	**Average**	**Minimum**	**Maximum**
4401216	25	127,803	114,814	138,549
NGS03108	15	39,618	26,771	42,421
NGS03287	7	135,813	132,191	141,791
NGS03459	20	112,471	105,096	121,204
NGS03614	22	120,811	103,443	130,453
NGS03615	1	132,783	132,783	132,783
NGS03693	9	135,920	132,166	140,760
NGS03874	34	132,328	95,305	141,768

Table 9: *Average fraction of reads per sample in each sequencing run, after each step of the processing pipelines, with reference to the raw data.*

	% of reads after each step, with reference to raw data		
Run-ID	**Trimming**	**Quality filtering**	**Merging and removing bimeras**
4401216	97.51	91.76	89.72
NGS03108	97.32	86.88	84.52
NGS03287	98.19	95.47	94.42
NGS03459	98.02	95.19	93.69
NGS03614	97.64	87.82	85.89
NGS03615	97.89	95.45	92.84
NGS03693	97.11	93.66	91.91
NGS03874	96.8	84.53	82.31

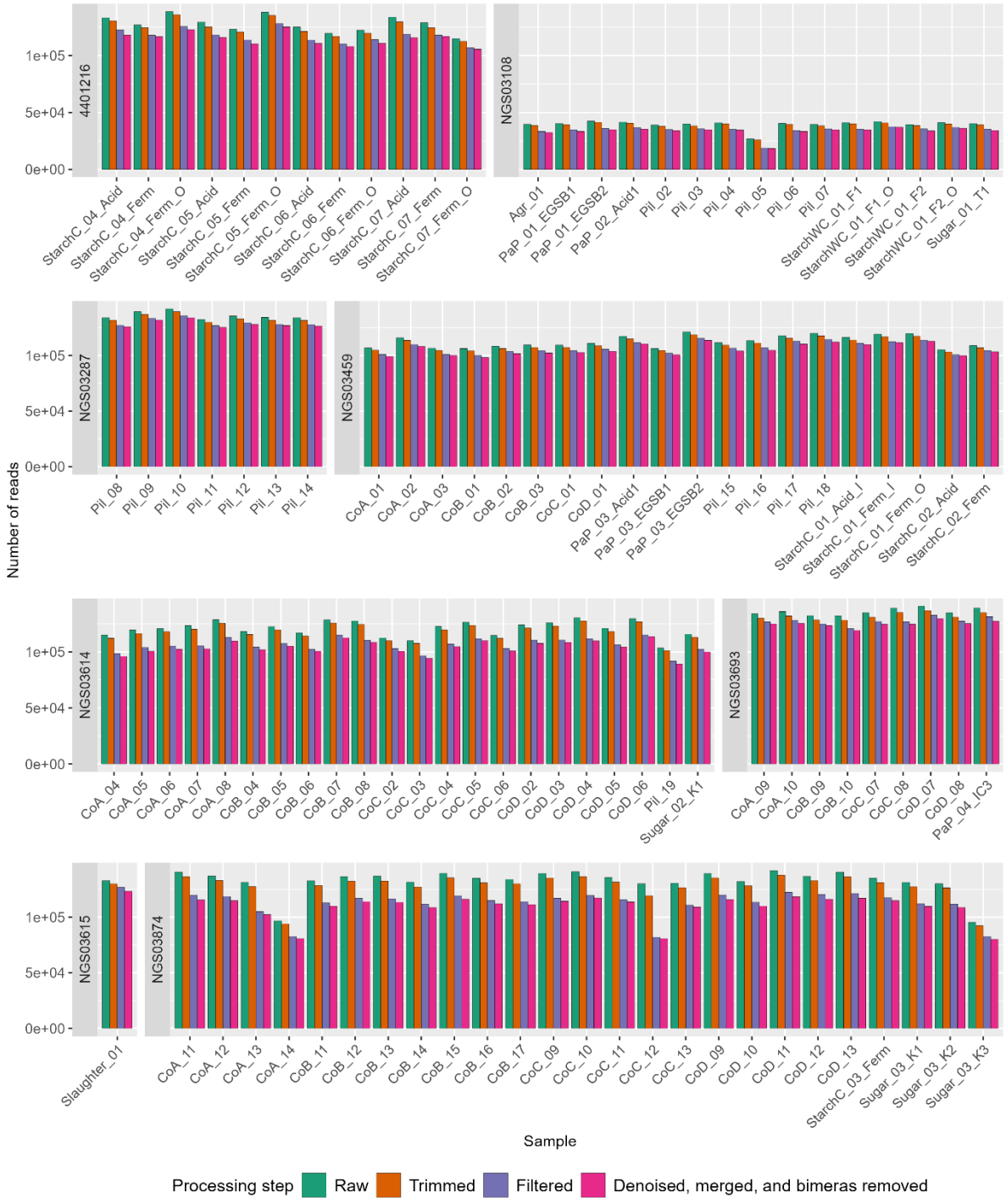

Figure 4: *Number of reads (y-axis) per sample (x-axis) across the pipeline that processes raw reads to ASVs. Four bars per sample show the number of reads after each processing step, as indicated by the colours. Grey fields on the y-axis organise the samples by sequencing runs.*

There are only minor deviations in the number of raw reads per sample within each sequencing run. Most of the runs yielded approximately 150,000 reads per sample. The sequencing run NGS03108 is an exception, where around 40,000 reads per sample were obtained. When no primer sequence is detected in a read, it is discarded during trimming. A large fraction of the reads is retained in this step. Less than 5 % are discarded. After trimming, low-quality reads are discarded. This step removes the largest fraction of reads. CoC_12 in run NGS03874 sticks out in Figure 4 because a higher fraction is discarded due to low quality, compared to the majority of the samples. Moreover, the runs NGS03108, NGS03614, and NGS03874

apparently have overall slightly worse quality than the other runs, as the fraction of reads that pass the first two steps is below 90 %. Nevertheless, most of the data passes quality filtering for all samples. Small fractions are lost during merging the paired-end reads due to insufficient overlap or mismatches between forward and reverse reads and during bimera removal. Overall, 82-94 % of the reads pass the complete pipeline. The high number of reads in the raw data and the small fraction of discarded data suggest that the results reflect most of the microbial community. Therefore, the resulting table with 9530 ASVs, was used for further analysis.

3.2.2 Comparison of classifiers

3.2.2.1 Confidence of taxonomic classification

For taxonomic classification of the ASVs, three different databases were used, namely MiDAS 5, version 5.3, and SILVA SSU Ref NR 99 SSU, versions 138.1 and 138.2. The full classification results are available in the supplementary material (Tables S4, S5, S6). Their performance regarding the classification of the identified ASVs is evaluated based on confidence of the classification and on the fraction of unclassified ASVs.

Confidence

ASVs with confidence values in specified ranges were counted at each taxonomic rank and for each database. The counts are given in Table 10. In all databases, a trend of decreasing confidence values for lower taxonomic ranks can be observed, as expected. However, at all taxonomic ranks, more taxa are classified with a confidence of 100 % when using the MiDAS 5 database than when using either of the SILVA databases. A high confidence can be obtained if the query sequence is almost identical to the sequence in the database and thus, the classification is unambiguous. Therefore, this result suggests that even the short sequence of the V4 region of the 16S rRNA gene is specific for the AD environment. Apparently, the obtained ASV sequences are more similar to those in the MiDAS database than in the SILVA databases.

Table 10: *Number of ASVs per rank in specified ranges of confidence values for taxonomic classification with three databases.*

Confidence [%]	Database	Kingdom	Phylum	Class	Order	Family	Genus
0-50	MiDAS 5.3	13	90	158	276	413	1250
	SILVA 138.1	2	230	411	654	849	1450
	SILVA 138.2	3	224	405	651	815	1489
51- 75	MiDAS 5.3	8	74	105	213	341	899
	SILVA 138.1	4	214	267	448	622	1186
	SILVA 138.2	6	223	277	450	614	1231
76 - 90	MiDAS 5.3	10	106	152	243	388	724
	SILVA 138.1	10	369	426	565	780	1165
	SILVA 138.2	10	366	414	585	773	1125
91 - 99	MiDAS 5.3	22	487	609	865	1099	1571
	SILVA 138.1	147	1233	1345	1725	1967	2242
	SILVA 138.2	153	1231	1358	1702	1968	2203
100	MiDAS 5.3	9477	8773	8506	7933	7289	5086
	SILVA 138.1	9367	7484	7081	6138	5312	3487
	SILVA 138.2	9358	7486	7076	6142	5360	3482

Fraction of unclassified ASVs

The fraction of unclassified ASVs (‘*NA*’) after taxonomic classification with three databases is shown in Figure 5.

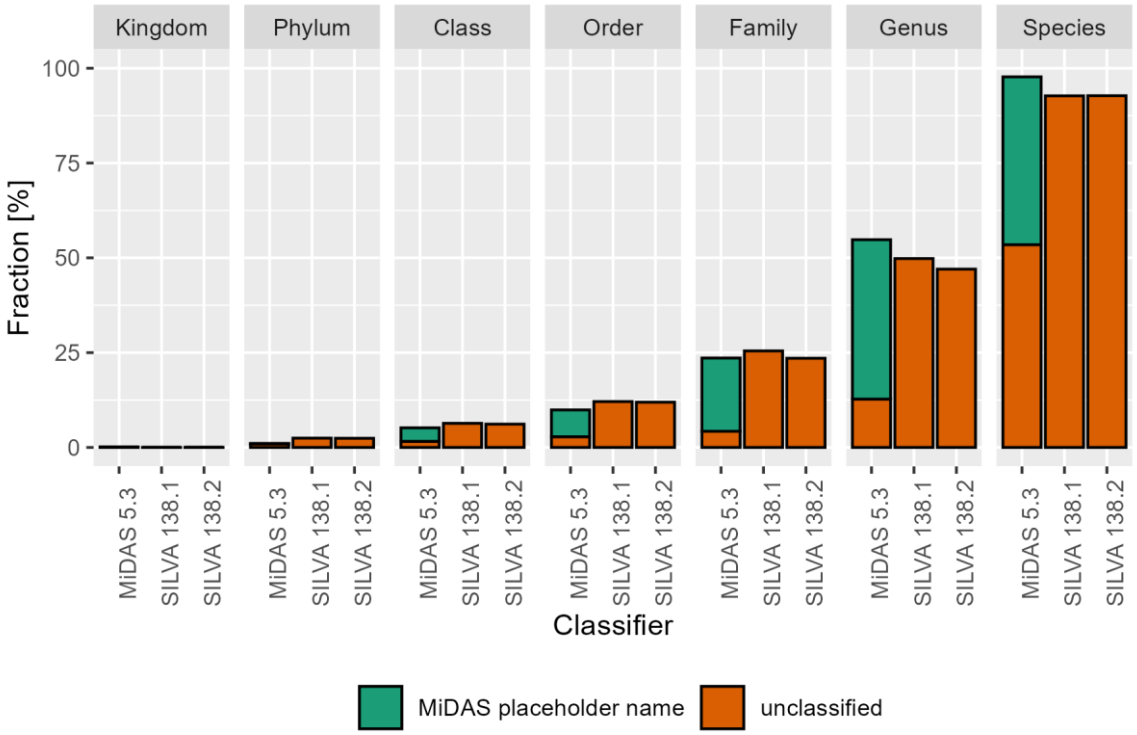

Figure 5: *Fraction of unclassified ASVs after taxonomic classification with the databases MiDAS 5.3, SILVA SSU Ref NR 99 138.1 and 138.2 and fraction of ASVs that were assigned MiDAS placeholder names.*

As expected, the fraction of unclassified ASVs increases with lower taxonomic ranks for all databases. Even though the values are similar at all levels for both SILVA versions, there are slightly less unclassified taxa in SILVA 138.2 than 138.1, especially at family and genus level. This reflects that more data is available in the more recent version (138.2) and thus, more ASVs can be classified.

The increase of unclassified ASVs is much steeper for the SILVA databases than for the MiDAS database. Using the MiDAS database, it is possible to classify more than 95 % at the family rank, 87 % at the genus rank, and more than 50 % at the species rank. Using the SILVA database, less than 10 % are classified at the species rank.

In the curation of the MiDAS database, taxa that could not be assigned a taxonomic classification are labelled by placeholder names in the format “midas_x_y”, where x corresponds to the taxonomic rank and y is a number. These names make it possible to compare taxa across datasets that cannot yet be classified, by giving them an identifier. Recurring taxa, and thus key taxa for AD, can be identified by these placeholder names. (Dueholm et al. 2024) The fraction of taxa that got assigned a MiDAS placeholder name is depicted in Figure 5. They approximately fill the gap of unclassified taxa between MiDAS and SILVA at all ranks. It can be concluded that there is a large overlap between the ASVs in the MiDAS database and those identified in this thesis. Most of the otherwise unclassified taxa obtain a placeholder name when classified based on the MiDAS database.

However, at genus and species rank, the sum of unclassified ASVs and ASVs with placeholder names exceeds the unclassified ASVs in the SILVA results. This can be explained by the MiDAS database not being an extension of the SILVA database, but a manual curation of it.

MiDAS 5.3 does not contain the complete SILVA database, but only the taxa that correspond to the full-length ASVs they found in the samples that were sequenced in the MiDAS project. However, the fraction of taxa that were detected in the systems that are analysed in this thesis but not in the MiDAS project is small. Based on these results, it seems that the MiDAS database covers the main taxa in AD systems, as intended by the curators.

It can be concluded that some information from the SILVA databases is lost when using the MiDAS database, but a lot of valuable information is added. This suggest that the database MiDAS 5.3 is best suitable to classify microorganisms when working with AD samples. Therefore, the results obtained using the MiDAS 5.3 database were used for further analyses.

3.2.3 Taxonomic profiling

After inferring ASVs from the amplicon sequencing dataset and classifying them taxonomically, the results were visualised in bar plots and bubble plots to observe patterns of the microbial communities in the analysed samples. A table showing the relative abundances of all ASVs in all samples can be found in the supplementary material (Table S7).

A bar plot showing the microbial community composition of all analysed samples at phylum level is shown in Figure 6. It provides a first insight into the similarity and dissimilarity of the samples with a focus on differences between system type and substrate type.

Some core phyla that are present in all or most of the samples can be identified, as well as some distinct features of the systems. These will be discussed in the following paragraphs.

Some trends can be observed from the phylum level distribution of taxa across various systems. The various substrates seem to drive different patterns of microbial community profiles. Especially the systems Conti, Pilot, and SK, which use residues as substrates, show similar taxonomic composition at phylum level. Dissimilar compositions can be observed for the various system types (i.e. single-stage system, acidification reactor or methane reactor). Chloroflexi are most abundant in single-stage systems treating wastewater and in methane reactors. At the same time, high relative abundances of Bacillota (Firmicutes) and Bacteroidota can be observed for single-stage systems treating residues and acidification reactors. Thus, the distribution of phyla suggests similarities between acidification reactors and single-stage systems treating presumably complex organic residues, and between methane reactors and single-stage systems treating presumably easier to degrade industrial wastewater.

The phyla Bacillota (previously Firmicutes) and Bacteroidota are identified as core phyla that are detected in almost every sample, regardless of the system, substrate, and system type. Both have previously been described as highly abundant key phyla in anaerobic digesters in several studies (Theuerl et al. 2015; Campanaro et al. 2016; Cai et al. 2016; Treu et al. 2016). Notably, they are most prevalent in acidification reactors and single-stage systems that treat residues. These results suggest that both phyla are important for the first stages of the AD process. This is in line with previous findings that assigned members of the phylum to the initial steps in breaking down organic matter, i.e. hydrolysis and acidification (Treu et al. 2016; Venkiteshwaran et al. 2015).

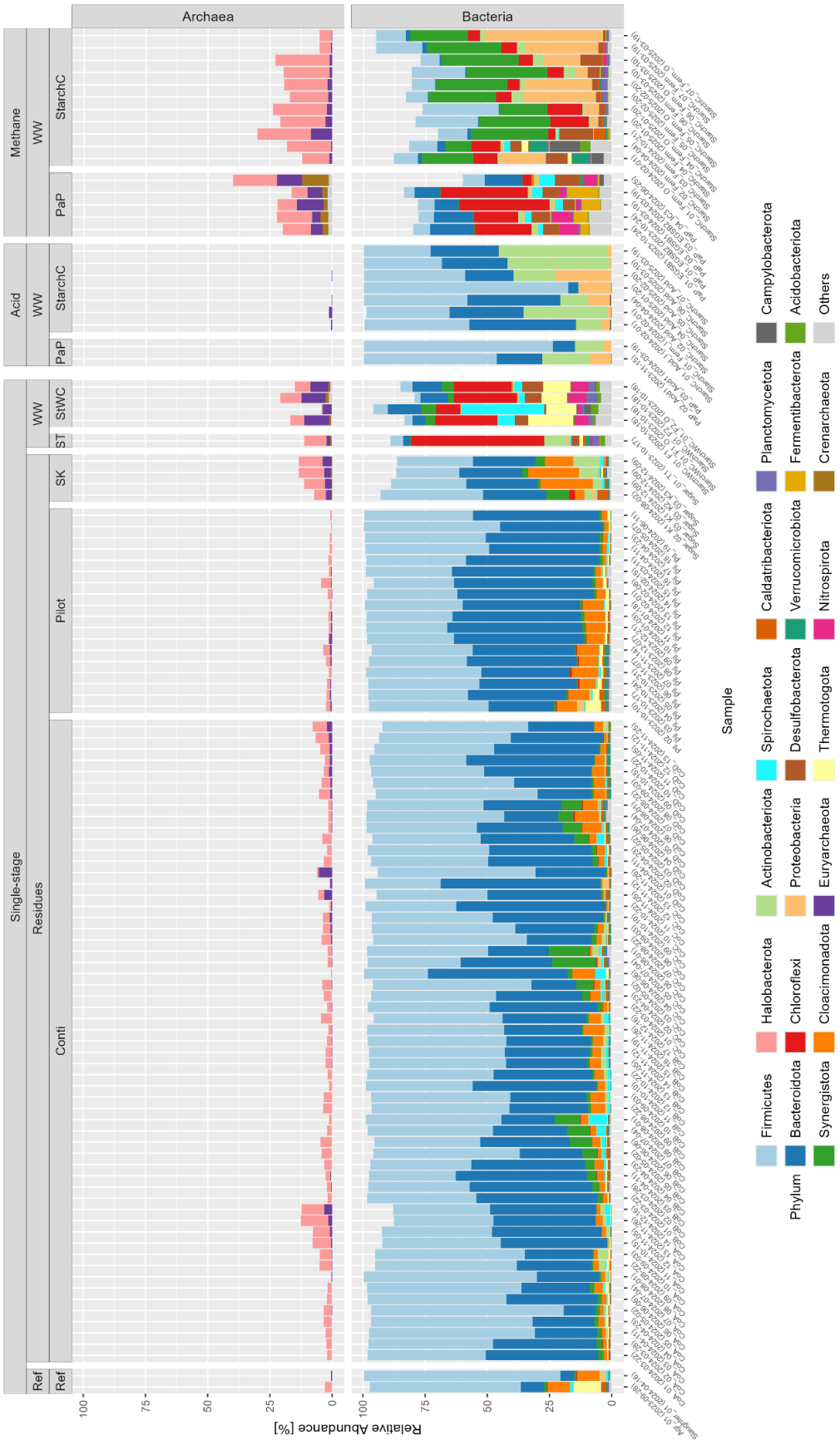

Figure 6: *Bar plot showing the composition of the microbial community in all analysed samples at phylum level. The bars are coloured by phylum. The 20 most abundant phyla are shown, and the rest are summarised in 'Others'.* ***Vertical layers*** *separate archaea and bacteria. The samples (x-axis) are organised by three horizontal layers.* ***Top layer:*** *Type of system they originate from. 'Single-stage': Typical single-stage systems; 'Acid': Acidification reactor in two-stage systems. 'Methane': Methane reactor in two-stage systems.* ***Middle layer:*** *Type of substrate that is used in the process. 'Ref': Reference samples; 'Residues': Agro-industrial residues; 'WW': Industrial wastewaters.* ***Bottom layer:*** *System names as used in Table 5 (section 2.3.1.1).*

Alongside Bacteroidota and Bacillota (previously Firmicutes), Actinomycetota (previously Actinobacteriota) and Pseudomonadota (previously Proteobacteria) are among the main phyla in the acidification reactors of two-stage systems treating industrial wastewater. However, they are not detected at high abundance in one-stage systems treating agro-industrial residues. Instead, they are found in the methane reactors of the two-stage systems. Thus, they might be specific for the substrates that are treated in these systems. However, according to previous findings, Pseudomonadota (previously Proteobacteria) are prevalent in anaerobic digesters using manure as substrate (Campanaro et al. 2016; Treu et al. 2016) while Actinomycetota (previously Actinobacteriota) have been associated with substrates such as sludge and wastewater (Treu et al. 2016).

In the methane reactors of the systems PaP, StarchC, StWC, and ST, Chloroflexota (previously Chloroflexi) are prevalent. It appears that they are specific for the systems that treat industrial wastewater, as described previously (Treu et al. 2016). In comparison, Synergistota are only prevalent in the methane reactor of the system StarchC which treats the wastewater from a starch factory.

As expected, archaea are almost exclusively detected in the methane reactors of two-stage systems and in one-stage systems. They are responsible for the last step in the AD process, i.e. converting substrates that are produced by bacteria to methane. Archaea are more abundant in the methane reactors of two-stage systems than in one-stage systems, where all four stages are conducted and thus a larger fraction of bacteria is expected.

In addition to the differences and similarities between the systems, it can be observed that the composition inside each system is stable over time. The biggest fluctuations within a system can be observed for ‘StarchC’, but the samples were collected over a long time period (approximately one year). It is known that the wastewater that the system receives is not stable due to seasonal fluctuations in the production process in the factory, which appears to affect the microbial community.

As described, distinct microbial community patterns can be observed even at the phylum level. However, lower taxonomic ranks are more informative and might reveal otherwise hidden trends. Thus, a more detailed investigation was done at the ASV level. Bubble plots that show the abundance of individual ASVs in the samples are given in Figure 7 (acidification reactors), Figure 8 (methane reactors and single-stage systems that treat industrial wastewater) and Figure 9 (methane reactors and single-stage systems that treat agro-industrial residues).

In Figure 7, the relative abundances of ASVs in acidification reactors of two-stage systems are shown. Both systems, namely ‘PaP’ and ‘StarchC’, treat industrial wastewater. Even though Bacillota (previously Firmicutes) are prevalent in both systems, the diversity within the phylum is higher in the system ‘StarchC’, as inferred from the number of ASVs (12 ASVs in ‘PaP’ compared to 35 ASVs in ‘StarchC’). Moreover, the abundant ASVs within the phylum are not shared between the systems. A similar pattern can be observed for Bacteroidota. In ‘PaP’, the genus *Cloacibacterium* is most abundant while *Bacteroides* and *Prevotella* are most abundant in ‘StarchC’, alongside some uncharacterised genera (identified by MiDAS 5-specific placeholder names). ASV2939 (*Prevotalla_9*) can be identified in both systems. Within the phylum Actinomycetota (previously Actinobacteria), the main genus in both systems is *Bifidobacterium*. However, of those, only ASV2939 is abundant in ‘PaP’ and ‘StarchC’. The members of the phylum Pseudomonadota (previously Proteobacteria) are mainly from the genus *Aeromonas* in ‘StarchC’, while other genera are detected in ‘PaP’.

Figure 7: *Bubble plot for acidification reactors of systems that treat wastewater. The size of the bubbles indicates the relative abundance of the ASV (y-axis) in the sample (x-axis). Only ASVs with minimum relative abundance of 1 % in any of the included samples are shown. The bubbles are coloured by phylum.* ***Vertical layers*** *separate archaea and bacteria. The samples (x-axis) are organised by three horizontal layers.* ***Top layer:*** *Type of system they originate from. 'Acid': Acidification reactor in two-stage systems.* ***Middle layer:*** *Type of substrate that is used in the process. 'WW': Industrial wastewaters.* ***Bottom layer:*** *System names as used in Table 5 (section 2.3.1.1).*

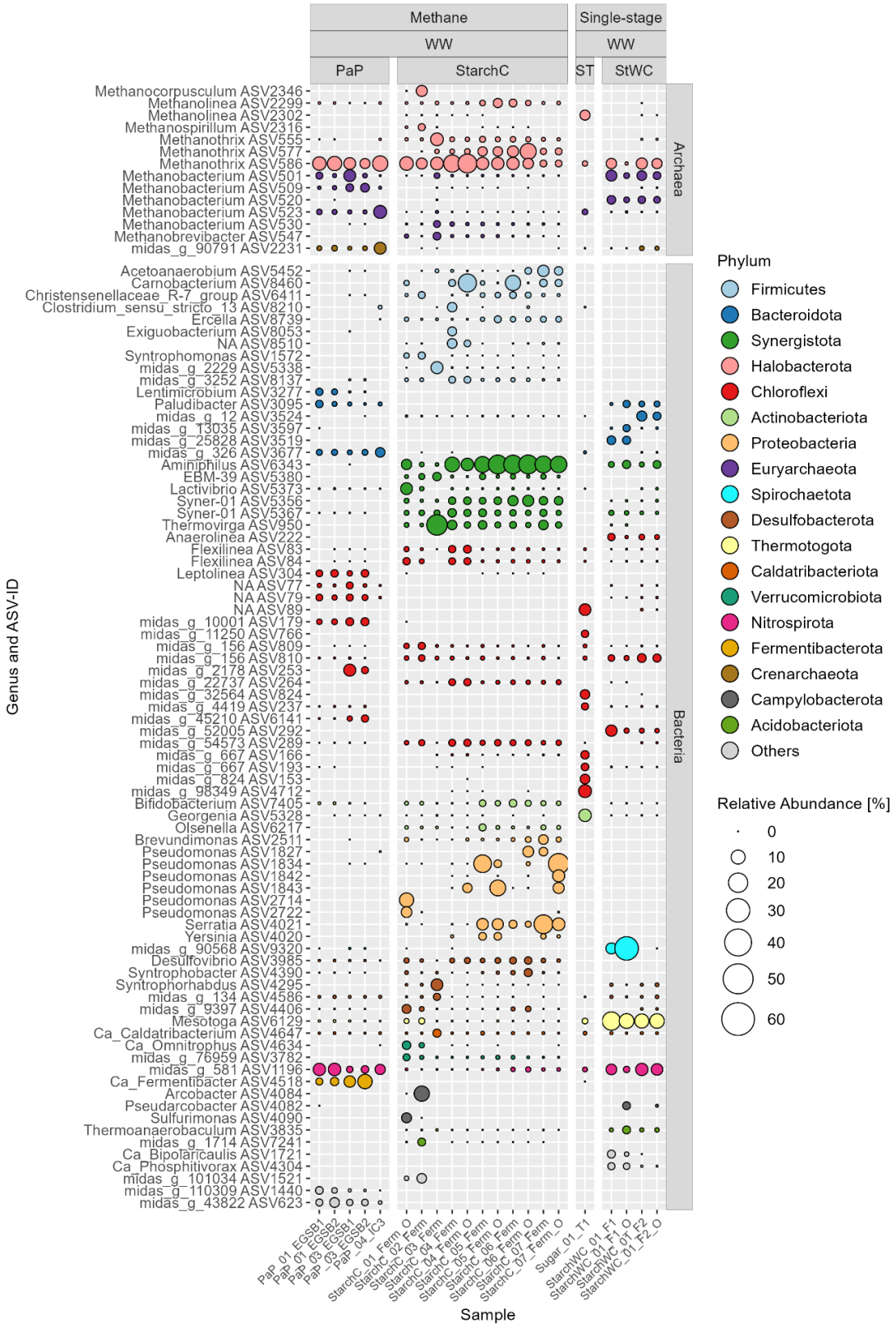

***Figure 8**: Bubble plot for single-stage systems and methane reactors of systems that treat wastewater. The size of the bubbles indicates the relative abundance of the ASV (y-axis) in the sample (x-axis). Only ASVs with minimum relative abundance of 2 % in any of the included samples are shown. The bubbles are coloured by phylum. **Vertical layers** separate archaea and bacteria. The samples (x-axis) are organised by three horizontal layers. **Top layer:** Type of system they originate from. 'Single-stage': Typical single-stage systems. 'Methane': Methane reactor in two-stage systems. **Middle layer:** Type of substrate that is used in the process. 'WW': Industrial wastewaters. **Bottom layer:** System names as used in Table 5 (section 2.3.1.1).*

Figure 9: *Bubble plot for single-stage systems that treat agro-industrial residues and reference samples. The size of the bubbles indicates the relative abundance of the ASV (y-axis) in the sample (x-axis). Only ASVs with minimum relative abundance of 2 % in any of the included samples are shown. The bubbles are coloured by phylum.* ***Vertical layers*** *separate archaea and bacteria. The samples (x-axis) are organised by three horizontal layers.* ***Top layer:*** *Type of system they originate from. 'Single-stage': Typical single-stage systems.* ***Middle layer:*** *Type of substrate that is used in the process. 'Ref': Reference samples; 'Residues': Agro-industrial residues.* ***Bottom layer:*** *System names as used in Table 5 (section 2.3.1.1). For the systems 'Conti' and 'Pilot', a subset of samples is shown instead of showing the complete dataset.*

Relative abundances of ASVs in methane reactors and single-stage systems that treat industrial wastewater are shown in Figure 8. There is no biomass retention in the acidification reactors which means that the biomass is transferred to the methane reactor in the effluent. Therefore, taxa that are detected in the acidification reactors may also be expected in the corresponding methane reactors initially. There is indeed an overlap of phyla between acidification and methane reactors, as the phyla Bacillota (previously Firmicutes) and Bacteroidota are detected in both. However, most genera and ASVs do not overlap.

The only genera that are detected in acidification and methane reactors are *Bifidobacterium* and *Olsenella*. Within the genus *Bifidobacterium*, ASV7405 is found in the acidification and methane reactors of 'PaP' and 'StarchC'. The relative abundance is higher in the acidification reactor for both systems (up to 12 % in acidification and up to 3 % in methane reactor for 'StarchC'; up to 2 % in acidification and up to 0.2 % in methane reactor for 'PaP'). ASV6217 (genus *Olsenella*) is detected in the acidification (up to 9 %) and the methane reactor (up to 2 %) of 'StarchC'. Among Pseudomonadota (previously Proteobacteria), *Aeromonas* is the main genus in the acidification reactor, but *Pseudomonas* is the main genus in the methane reactor in 'StarchC' and no ASVs overlap. This suggest that even though microorganisms are transferred from one reactor to the other, they are overgrown by other taxa and are thus detected at lower abundances or not at all. In the system 'PaP', this effect might also be attributed to the granular nature of the biomass in the methane reactors. Microorganisms that are introduced with the effluent from the acidification reactor do presumably not attach to the granules and are quickly washed out. These insights emphasize the importance of investigating lower taxonomic ranks.

Nevertheless, phylum level differences are observed as well. The phylum Synergistota is only prevalent in the methane reactor of 'StarchC' and in 'StWC'. It is dominated by ASV6343 (genus *Aminiphilus*) in both systems. All ASVs that belong to Chloroflexota (previously Chloroflexi) are apparently not well characterised. They are either identified by a MiDAS 5-specific placeholder or not at all at the genus level. A similar observation was made by Dueholm et al. (2024) who found that 43 % of all Chloroflexota ASVs they detected in a large set of anaerobic digesters are novel. In addition, this phylum is not identified in acidification reactors or in systems that treat other substrates than industrial wastewater, providing clues to their role in the system. Notably, the ASVs of this phylum are specific for the individual systems. The only overlaps are observed for ASVs that occur in 'StarchC' and 'StWC', as observed for the phylum Synergistota. They are the only systems in the dataset that obtain the wastewater from starch production. Possibly, the taxa that are specific for both systems are associated with metabolites that are obtained from starch breakdown. The phylum Desulfobacterota is detected in all systems but most abundant in 'StarchC'. The wastewater that is treated in this system is known for its high sulphate concentration, originating from the starch production process. Sulphate reduction is prevalent in the Desulfobacterota phylum.

Judging solely from the phylum level, microbial communities appear to be similar in single-stage systems that use agro-industrial residues as substrates (Figure 9) and in acidification reactors (Figure 7), see also Figure 6. Both have large fractions of Bacillota (previously Firmicutes) and Bacteroidota. However, there is almost no overlap at genus or ASV level within Bacillota (previously Firmicutes). Moreover, the genera are different among the agro-industrial single-stage systems, except some of the most abundant ASVs of the genus *Fastidiospilia*. The genera *Christensenellaceae*, *Clostridium sensu stricto*, and *Syntrophomonas* are detected in systems that treat agro-industrial residues as well as in those that treat industrial wastewaters but again, there are no overlapping ASVs. Furthermore, a large fraction of the ASVs of this phylum that are detected in the agro-industrial systems are only identified by MiDAS 5-specific placeholder names. In contrast, genera from the same phylum are mainly known in the other systems. Similarly, to acidification reactors, a large group of Bacteroidota

got assigned MiDAS 5-specific placeholder names at genus level. However, the MiDAS identifiers are not the same for ASVs in acidification reactors and for those in the agro-industrial single-stage systems. Bacteroidota are less abundant in both samples of the 'Ref' systems. The phylum Cloacimonadota is detected in all agro-industrial and 'Ref' systems, but not in the systems that treat industrial wastewater.

Overall, even though the systems seem to be similar when analysing them at phylum level, a detailed investigation at genus and ASV level reveals that there are only little overlaps between the systems. Similarities are mainly observed for the same type of system, i.e. single-stage, acidification reactor, or methane reactor, and for systems with similar substrates. Therefore, taxonomic profiling suggests that substrate and system type are the main drivers for shaping the microbial community. It has been described in previous studies that the substrate is an important parameter for the microbial community composition (Campanaro et al. 2020; Dueholm et al. 2024).

3.2.4 Multivariate analysis

The samples analysed in this thesis and described in the previous sections were merged with a larger, publicly available dataset, comprising samples from the MiDAS 5 database (Dueholm et al. 2024). This was done to put the data into a broader context, as this database hosts samples from a broad range of anaerobic digesters. The aim is to reveal patterns of similarities regarding the taxonomic profiles of the samples. This dataset contains 363 samples from 202 anaerobic digesters from the MiDAS 5 database. After merging the samples, the final dataset is made up of 474 samples. A sample list including metadata is available in the supplementary material (Table S8).

After processing the merged dataset in USEARCH, 15,504 zOTUs and their relative abundances in 474 samples were obtained (see supplementary Table S9). Two different ordination methods were applied to the abundance table to observe global trends. Figure 10 shows the ordination plots for both methods, with the correspondence analysis (CA) result in Figure 10A and the non-metric multidimensional scaling (NMDS) result in Figure 10B.

In the CA plot, the so-called 'arch effect' is observed, which is a common issue in CA (Ramette 2007). It results from the mathematical procedure that is used to obtain uncorrelated axes and occurs when the first axis is enough to maximally separate the species (Ramette 2007). The obtained plots are difficult to interpret ecologically but the position of samples along the arch is meaningful (Ramette 2007). The effect can be observed in the CA plot in Figure 10A. The results of both ordinations are therefore investigated in a combined way to cross-validate any conclusions. Trends that are supported by both methods are more likely to reflect real biological patterns.

One limitation of this analysis is that the substrate in a large subset of the MiDAS samples is unknown ('MiDAS: n.d.'). The samples are spread across a large area of the plots but they are not informative. Moreover, systems with wastewater sludge, i.e. activated sludge from wastewater treatment plants, make up another large subset. This can be attributed to the original focus of the MiDAS project, which is not on AD systems but on activated sludge (McIlroy et al. 2015). This introduces a bias to the analysis. However, samples from systems that use industrial substrates or various residues are part of the dataset as well.

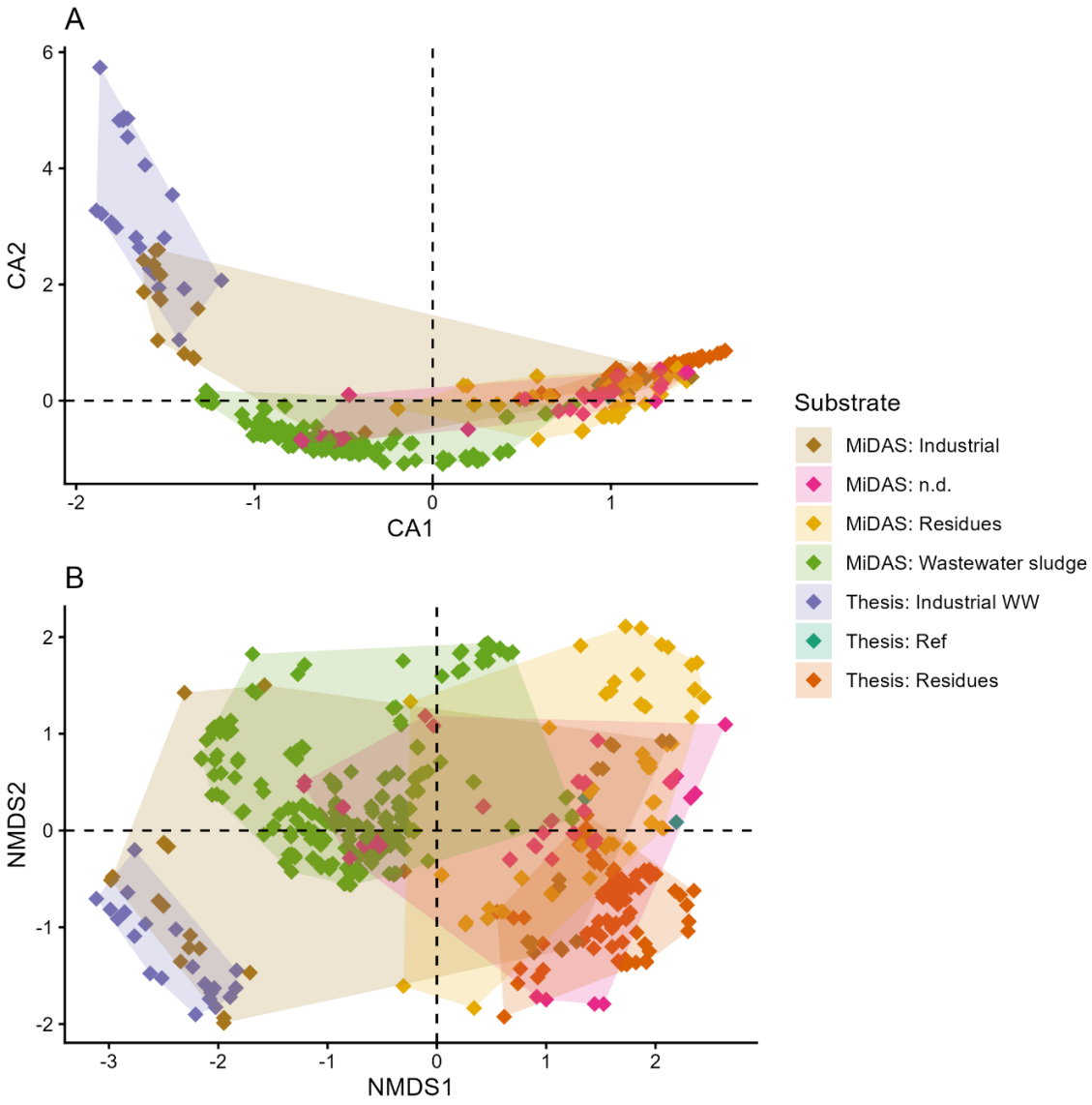

Figure 10: *Ordination plots for combined dataset (MiDAS and thesis). The ordinations are based on the relative abundances of 15,504 zOTUs in 474 samples from the dataset analysed within this thesis ('Thesis') and from a publicly available database ('MiDAS'). Each square represents a sample. The samples are coloured by the substrate that is used in the anaerobic digestion system they originate from.* ***A:*** *Correspondence analysis (CA).* ***B:*** *Non-metric multidimensional scaling (NMDS; stress: 0.1528).*

The same overall trends are detected in both plots. Two separate clusters can be observed, even though the samples within them are spread over a large area. Judging from the colours that indicate the substrates used in the systems, the same samples are clustered together in both plots. The cluster on the right is larger and more widespread than the one on the left. All MiDAS samples from systems that use residues or wastewater sludge as substrate and all thesis samples that use residues are part of this cluster. Moreover, the MiDAS samples with unknown substrates can be found in the same area of the plots. Samples from the thesis dataset are at the edge of the cluster which suggests that they are similar to the MiDAS samples but make up a distinct group within them.

MiDAS samples from systems with industrial substrates are the only ones that are distributed across both clusters. Interestingly, some of the systems are located at similar industries as the ones in the thesis dataset, i.e. corn processing, paper industry. However, metadata does not further specify the nature of the industrial substrates of all samples. It is possible that the samples in the cluster on the right are from systems that use industrial residues and those in

the cluster on the left are from systems that use industrial wastewaters. This hypothesis would be in line with the positions of the samples from the thesis dataset in the ordination, but it cannot be confirmed.

This analysis shows that the samples investigated in this system are generally comparable to a more global dataset. However, it can be concluded that they are underrepresented in the MiDAS 5 database. Especially industrial anaerobic digesters make up a small part of the publicly available dataset. Moreover, the ordination shows that their taxonomic composition is most distinct from the other samples. Thus, it is expected that new insights into an underrepresented type of AD systems will be obtained from characterising them by metagenomic methods. Moreover, the industrial systems ‘StarchC’ and ‘PaP’ are two-stage systems. This means that the stages of the AD process are separated by using an acidification (hydrolysis and acidogenesis stage) and a methane reactor (acetogenesis and methanogenesis stage). This facilitates functional analysis of the taxa in the distinct reactors. Based on these findings and considerations, samples from the systems ‘StarchC’ and ‘PaP’ were chosen for the metagenomic analysis in this thesis.

3.3 Metagenomic analyses

To gain insights into the functional potential of the taxa that are present in anaerobic digestion (AD) systems, metagenomic analyses were conducted. A subset of the systems in the previously obtained dataset analysed above (section 3.2) was investigated. For this analysis, a new set of samples was used. After evaluating the stability of the systems by amplicon sequencing, long-read sequencing was conducted. Read-based taxonomic profiles were obtained before recovering metagenome-assembled genomes (MAGs) to analyse the potential of the microorganisms to execute key AD metabolic pathways.

3.3.1 Sample processing

To analyse the stability of the selected systems (PaP, StarchC, Ref), amplicon sequencing was conducted and the results are compared to the previously obtained results (see section 3.2). This was followed by long-read sequencing to obtain MAGs for a functional analysis.

The fragment lengths after extraction with two different kits are analysed to evaluate the suitability of DNA extracts for long-read sequencing, see Figure 11 and supplementary Table S11.

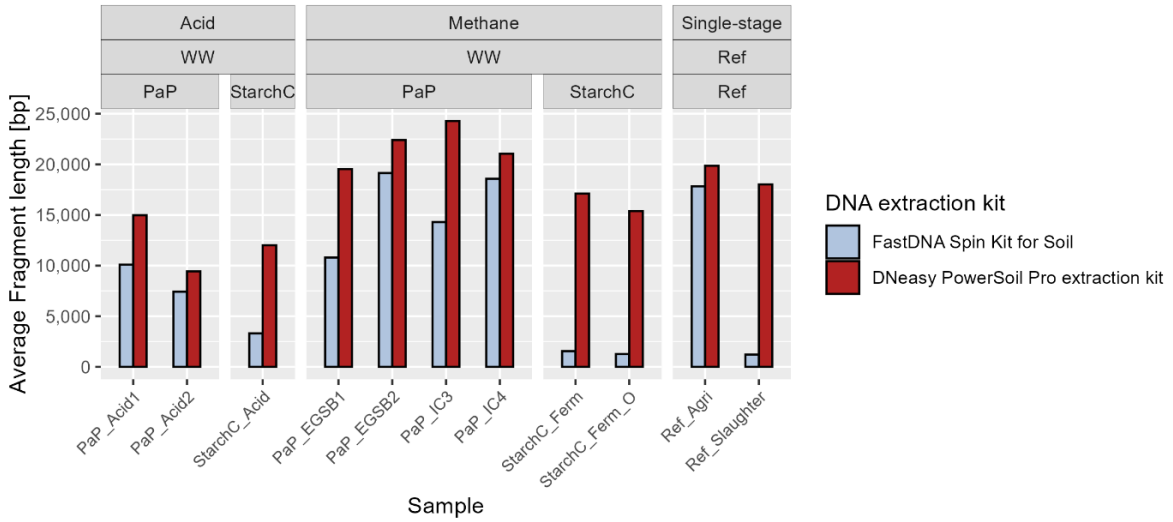

Figure 11: *DNA fragment lengths after DNA extraction using the FastDNA™ Spin Kit for Soil and the DNeasy PowerSoil Pro extraction kit for DNA extraction.*

The obtained fragment lengths differ for the two DNA extraction kits that were used. The average fragment lengths produced by the FastDNA™ Spin Kit for Soil are short for some samples (Ref_Slaughter and all in 'StarchC'; < 4,000 bp). A possible explanation could be that the shearing forces of high-speed bead beating disrupted the DNA and led to fragments that are undesirably short for long-read sequencing. The DNeasy PowerSoil Pro extraction kit uses vortex bead beating to disrupt the cells instead which is a gentler method. Switching to DNA extraction with this kit yields longer DNA fragments (close to or longer than 10,000 bp) which are suitable for long-read sequencing. Thus, it was decided to use the DNA extracts obtained from the DNeasy PowerSoil Pro extraction kit for long-read sequencing with the Oxford Nanopore Technologies (ONT) method. Amplicon sequencing was done using the DNA extracts obtained from the FastDNA™ Spin Kit for Soil to ensure comparability with the previously obtained dataset analysed above (section 3.2).

3.3.2 Amplicon and long-read sequencing

After sequencing using the Illumina (amplicon sequencing) and ONT platform (long-read sequencing), the quality of the obtained data was analysed. The main parameters of interest are the sequencing yield for both sequencing technologies, i.e. number of reads, and read quality, i.e. mean read length quality score, for long-read sequencing. They are given for all samples in Table 11. The full results for the ONT dataset can be found in the supplementary material (Table S12). Due to sequencing issues in long-read sequencing, the samples with the prefix 'StarchC' were sequenced twice. The results shown here and in the entire thesis are for the merged dataset of both runs.

Table 11: *Quality control results of amplicon and long-read sequencing. The shown results for 'StarchC' are for the merged dataset of both sequencing runs.*

	Amplicon sequencing	Long-read sequencing		
Sample	**Nr. of reads**	**Nr. of reads**	**Mean read length [bp]**	**Mean quality score**
PaP_Acid1	7,713	857,805	7,662	21
PaP_Acid2	4,428	955,667	5,488	21
PaP_EGSB1	4,043	1,574,700	10,099	20
PaP_EGSB2	9,163	1,262,132	10,064	20
PaP_IC3	6,702	960,415	11,341	21
PaP_IC4	5,894	915,716	10,977	21
StarchC_Acid	5,069	325,175	3,268	21
StarchC_Ferm	5,984	254,791	5,285	21
StarchC_Ferm_O	5,387	266,249	5,612	21
Ref_Agri	3,647	858,781	8,042	20
Ref_Slaughter	2,568	2,012,421	8,289	20
Average	5,509	931,259	7,830	21
Minimum	2,568	254,791	3,268	20
Maximum	9,163	2,012,421	11,341	21

For the amplicon sequencing datasets, the sequencing depth (i.e. number of reads per sample) is much lower than for the previous dataset (5,509 vs. 117,193 reads per sample on average). Nevertheless, it is expected that the main taxa can be detected because the microbial community is skewed to a low number of high abundance taxa.

The sequencing issues regarding samples from the system ‘StarchC’ in long-read sequencing mentioned above are reflected by the number of reads per sample. In the first sequencing run, 50,000 to 150,000 reads were obtained (data not shown), compared to >800,000 reads for the other samples. A DNA clean-up step was done to remove potential inhibitors, and the cleaned DNA was sequenced. The number of reads for the sample StarchC_Acid is 181,136 after clean-up, opposed to 144,039 without the clean-up. For StarchC_Ferm the clean-up increased the number of reads from 54,944 to 199,847, and for StarchC_Ferm_O the number of reads increased from 144,701 to 121,548. After merging both runs, approximately 300,000 reads per sample are obtained which is still less than half of the yield obtained for the other samples (see Table 11). The reason for this dramatic difference was not found, but inhibitory substances related to the wastewater which is treated in the ‘StarchC’ system might be present in the samples. Since this trend is not observed for amplicon sequencing, the inhibitors did appear to be problematic for ONT sequencing specifically. ONT sequencing is based on the principle that the DNA strand is pulled through a nanopore (Deamer et al. 2016). Inhibitory substances might either hinder the DNA from attaching to the pores, or they might clog the pores and thus the DNA cannot be drawn through them.

Due to the low sequencing yield, it is expected that genomes of low abundance are not detected in ‘StarchC’ samples. In contrast, the high number of reads in the other samples suggest that low abundance genomes can be detected. Therefore, a direct comparison of the systems ‘StarchC’ and ‘PaP’ is not possible. It was decided to focus the analysis on ‘PaP’ and consider this observation when comparing individual observations.

Despite the low yield, the quality of reads in ‘StarchC’ samples is comparable to that of the other samples. Mean read lengths per sample are 7,830 bp on average. The longest reads are obtained from ‘PaP’ methane reactors (Pap_EGSB1, PaP_EGSB2, PaP_IC3, PaP_IC4), with a mean read length of approximately 10,000 bp. Overall, the reads are approximately half as long as the DNA fragments that were used for library preparation. The quality scores are uniformly at 20 or 21, which correspond to an error rate of 99 % or 99.2 %, respectively. This is an expected error rate for modern ONT long-reads (MacKenzie and Argyropoulos 2023).

3.3.3 Taxonomic profiling

Read-based taxonomic profiling was conducted for the various sequencing approaches used for the samples obtained for metagenomic analysis. The aim is to not only evaluate the biological stability of the systems but also the comparability of the various approaches.

Firstly, taxonomic profiles for the amplicon sequencing results are investigated to evaluate the biological stability of the systems by comparing them to the taxonomic profiles obtained from the previously obtained dataset analysed above (section 3.2).

Secondly, read-based taxonomic profiles for long-read sequencing results are analysed to determine if the same main taxa can be identified with this approach as with amplicon sequencing. Two different tools (MetaPhlAn 4 and SingleM) were used to obtain the taxonomic profiles, and the results are compared to be able to choose the most suitable tool for this dataset.

The taxonomic profiles obtained with all of the three described approaches are shown in Figure 12. The full results can be found in the supplementary material (Table S13: MetaPhlAn; Table S14: SingleM).

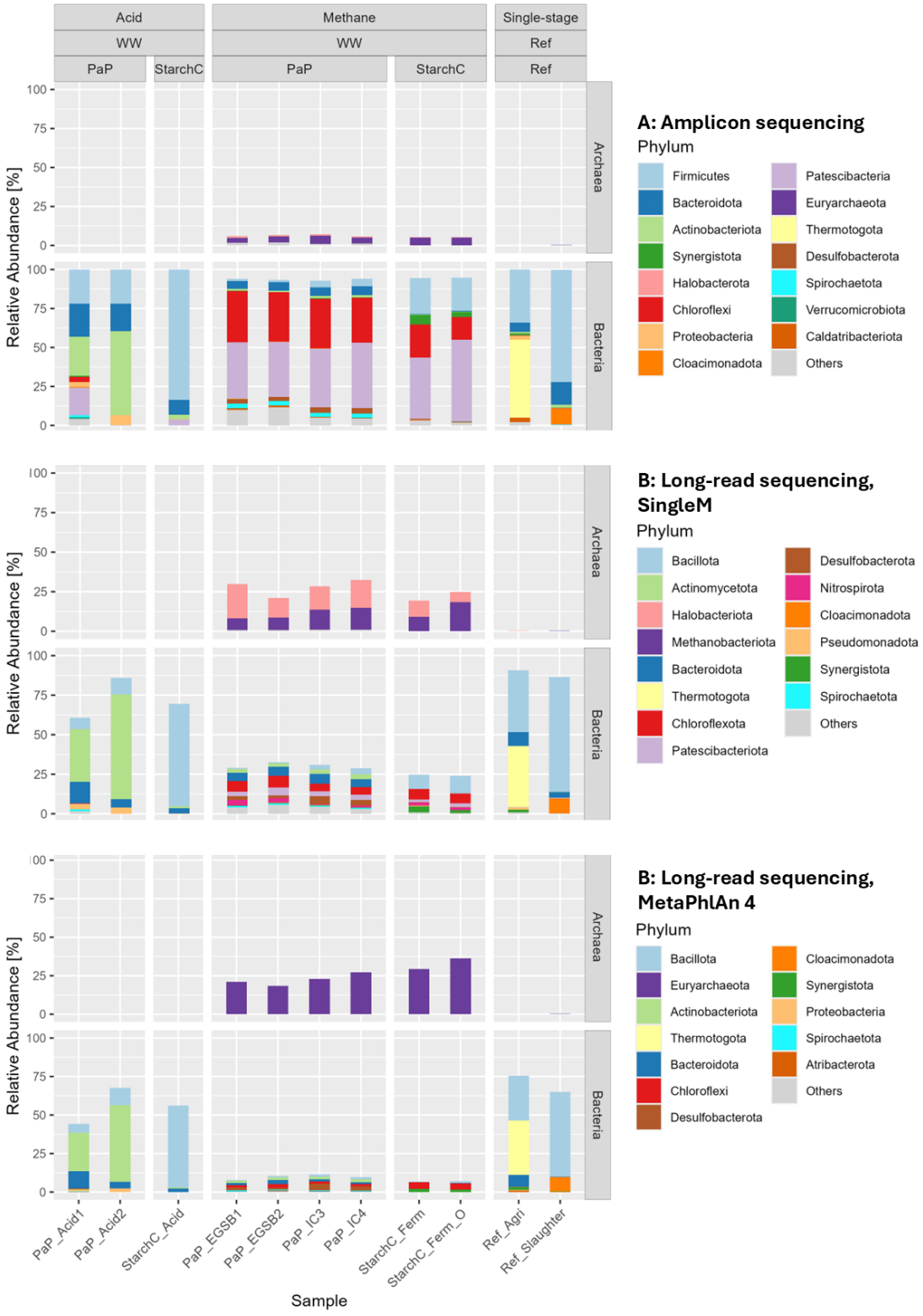

Figure 12: *Taxonomic profiles obtained from classification of amplicon* ***(A)*** *and ONT* ***(B, C)*** *sequencing reads of samples for metagenomic analysis. Reads were classified using different tools; A: DADA2 with MiDAS 5.3 database. B: SingleM with GlobDB. C: MetaPhlAn 4. The same colours are chosen to reflect the same phyla in all plots, despite of different taxonomies in the databases.* ***Vertical layers*** *separate archaea and bacteria. The samples (x-axis) are organised by three horizontal layers.* ***Top layer:*** *Type of system they originate from. 'Acid': Acidification reactor in two-stage systems. 'Methane': Methane reactor in two-stage systems. 'Single-stage': Typical single-stage systems.* ***Middle layer:*** *Type of substrate that is used in the process. 'WW': Industrial wastewaters. 'Ref': Reference samples.* ***Bottom layer:*** *System names as used in Table 5 (section 2.3.1.1).*

3.3.3.1 Evaluation of the biological stability of the investigated systems

The taxonomic profiles obtained from amplicon sequencing of the previous dataset (Figure 6 in section 2.3.3.1) and the dataset of fresh samples that is used for metagenomic analysis (Figure 12A) are compared. The sample and data processing was similar for both approaches: The same DNA extraction kit was used and the V4 region of the 16S rRNA gene was amplified and sequenced using the Illumina platform. ASVs were obtained and taxonomically classified using DADA2 and the database MiDAS 5, version 5.3 (Dueholm et al. 2024). The main difference between both approaches is the primer pair used for amplification of the V4 region. For the previous dataset, the standard V4 primer pair 515F and 806R (Klindworth et al. 2013) was used. For the fresh samples, it was decided to use the V4-EXT primer pair which became available because it has been shown that it covers more of the biological diversity than the standard V4 primer pair (Hu et al. 2025).

The most striking difference between the results of both approaches is the large fraction of Patsecibacteria that is observed in the fresh samples while the phylum is not among the 20 most abundant phyla in the previous dataset. It has been described that the V4-EXT primer pair improves the detection of Patescibacteria (Hu et al. 2025). This suggests that the observed difference does not reflect biological instability but is caused by the utilisation of different primer pairs.

Apart from Patescibacteria, the same phyla are found in the related samples from both datasets. Differences in their relative abundance are observed. However, the perception of relative abundances is distorted by the large fraction of Patescibacteria which is only detected in the previously obtained dataset analysed above (section 3.2). When normalising the data after removing ASVs from this phylum, the relative abundances are similar in both datasets.

The composition of the microbial community in the sample from one acidification reactor in the system ‘PaP’ (PaP_Acid1) however is surprising. In addition to the phyla that are detected in the other acidification reactor (PaP_Acid1), the phyla that are specific for the methane reactors of the same system (PaP_EGSB1, PaP_EGSB2, PaP_IC3, PaP_IC4) are observed. The composition thus appears to be a mixture of the typical taxa in the acidification and methane reactors. Based on the observations described here and in section 3.3.6 (metadata analysis), it is hypothesized that the process in this reactor has shifted to a single-stage process with the complete AD process taking place in one reactor. Biological interpretations regarding metabolic processes specific for the acidification reactor in this system should therefore rely mainly on the sample PaP_Acid2.

It is concluded that the systems are biologically stable and the samples collected for metagenomic analysis are representative for the reactors, except for PaP_Acid1.

3.3.3.2 Taxonomic profiling based on metagenomic long-reads

Taxonomic profiling results are strongly dependent on the tool and database that are used for the classification. Two tools and databases, namely MetaPhlAn 4 with its integrated database (Blanco-Míguez et al. 2023) (results shown in Figure 12C) and SingleM with the GlobDB, release 226 (Woodcroft et al. 2025; Speth et al. 2025) (results shown in Figure 12B), were used and the classification performance is evaluated to choose the more suitable approach for the analysed dataset.

The tools differ in the fraction of reads they are able to classify, see Figure 13.

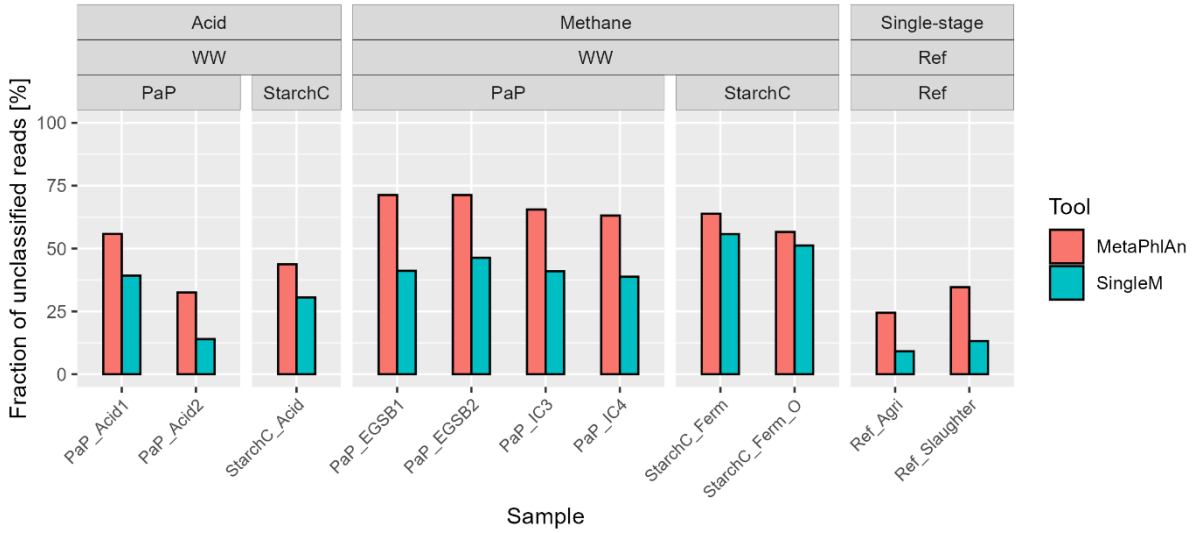

Figure 13: *Fraction of unclassified reads after read-based taxonomic classification (taxonomic rank of species) using MetaPhlAn and SingleM. The samples (x-axis) are organised by three horizontal layers.* ***Top layer:*** *Type of system they originate from. 'Acid': Acidification reactor in two-stage systems. 'Methane': Methane reactor in two-stage systems. 'Single-stage': Typical single-stage systems.* ***Middle layer:*** *Type of substrate that is used in the process. 'WW': Industrial wastewaters. 'Ref': Reference samples.* ***Bottom layer:*** *System names as used in Table 5 (section 2.3.1.1).*

The fraction of unclassified reads fluctuates across the dataset for both tools. When using MetaPhlAn 4 with the integrated database, 24-75 % of the reads remain unclassified. The authors of the tool observed similar results for samples from freshwater, ocean, or soil (Blanco-Míguez et al. 2023). The tool focuses on host-associated datasets, for which it is able to classify a larger fraction of reads than for environmental datasets (Blanco-Míguez et al. 2023). This can be attributed to an underrepresentation of environmental datasets, including AD, in the MetaPhlAn 4 sequencing database (Blanco-Míguez et al. 2023).

In contrast, SingleM focuses on the classification of reads from novel lineages for which no complete or draft genomes exist (Woodcroft et al. 2025). Moreover, the GlobDB, which was used as database for SingleM, is a broad database which represents microbial genomes from 14 sources, some of which focus on environmental datasets (Speth et al. 2025). The application of SingleM with the GlobDB is therefore expected to be more suitable for analysing uncharacterised environments such as anaerobic digesters. This hypothesis is supported by the obtained results. SingleM is able to classify more reads than MetaPhlAn 4 in all samples, with unclassified fractions of 4-56 %. Moreover, even if no species can be assigned by SingleM, the tool classifies the reads at higher taxonomic ranks, if possible, in contrast to MetaPhlAn 4. Less than 1 % remains unclassified at the phylum rank when using SingleM with the GlobDB. The results obtained from SingleM are therefore used to compare the long-read based to the amplicon read based taxonomic profiles.

At the phylum level, the taxonomic profiles are similar for both datasets (Figure 12A and Figure 12B), despite the utilisation of different DNA extraction kits. However, after long-read sequencing, the sample PaP_Acid1 appears more similar to PaP_Acid2 than to the methane reactor samples. A larger fraction of the metagenomic reads is classified as archaea than for amplicon reads. Especially the phylum Halobacteriota is better represented by the metagenomic reads. The bacterial phylum Nitrospirota is only detected in the metagenomic dataset as well. When comparing the results at species level (see supplementary Tables S4, S7, S17), more diversity can be observed within the phyla Actinomycetota (previously Actinobacteriota) and Desulfobacterota in the metagenomic dataset. These observations could be the result of the primer bias effect in the amplicon dataset.

The effect of the low sequencing yield for the samples from the system ‘StarchC’ becomes clear when doing a species level comparison of the taxonomic profiles. Apart from taxa mentioned above, only the most abundant species of the amplicon dataset are found in the metagenomic dataset. A similar effect can be observed for the samples from the system ‘PaP’ but the larger sequencing depth lowers the detection limit. In both ‘Ref’ samples, most of the species are detected in the amplicon as well as the metagenomic dataset. This can be attributed to the lower complexity of the samples, i.e. less ASVs were detected, and they are distributed more evenly compared to the other samples.

3.3.4 Assembly

After long-read sequencing, the metagenomic reads were assembled into contigs using metaFlye and medaka. Single-sample assemblies were obtained for all samples. In addition, co-assemblies were obtained for the acidification reactors of the system ‘PaP’ as well as for the samples from methane reactors of both systems (PaP and StarchC).

To assess the results, metrics that describe them were obtained from using Quast. The main results are given in Table 12.

Table 12: *Quality control results of single-sample assemblies and co-assemblies. Asm = assemblies.*

	Sample	Nr. of contigs	Mean contig length [bp]	Nr. of contigs >50,000 bp	N50
Single-sample asm.	PaP_Acid1	7,694	34,713	1,003	69,604
	PaP_Acid2	3,960	27,729	346	77,123
	PaP_EGSB1	14,368	59,452	3,948	128,015
	PaP_EGSB2	14,031	55,956	3,811	111,884
	PaP_IC3	10,613	62,389	3,078	133,682
	PaP_IC4	9,616	60,493	2,584	132,449
	StarchC_Acid	2,458	9,323	62	18,023
	StarchC_Ferm	1,876	34,479	226	78,169
	StarchC_Ferm_O	1,584	38,535	196	97,228
	Ref_Agri	5,726	34,626	712	78,977
	Ref_Slaughter	7,035	43,171	1,294	102,688
Co-asm.	PaP_acids	9,874	33,175	1,211	67,702
	PaP_fermenters	36,030	60,552	9,661	144,455
	StarchC_fermenter	3,288	35,785	455	94,449
	Average	9,154	42,170	2,041	95,318
	Minimum	1,584	9,323	62	18,023
	Maximum	36,030	62,389	9,661	144,455

The number of contigs is approximately 1 % of the number of reads and the mean contig length is approximately 5x the mean read length in the unassembled data. 10-15 % of the contigs, and even almost 30 % for the methane reactors from the system ‘PaP’, are longer than 50,000 bp. Thus, the sequences were successfully assembled to more contiguous sequences. The sample StarchC_Acid has only 62 contigs longer than 50,000 bp which is only 3 % of all contigs. A wide range of N50 values is observed, with a minimum of 18,023 bp and a maximum of 144,455 bp. In general, a larger N50 value means more contiguous sequences. However,

the interpretation is difficult for metagenomic samples, since many genomes of different sizes are contributing to it.

A heatmap showing the fraction of reads of each sample that can be mapped back to each assembly is presented in Figure 14. The underlying data is available in the supplementary material (Table S15). Besides the described basic metrics, this is an important parameter to determine how well which sample is represented by which assembly.

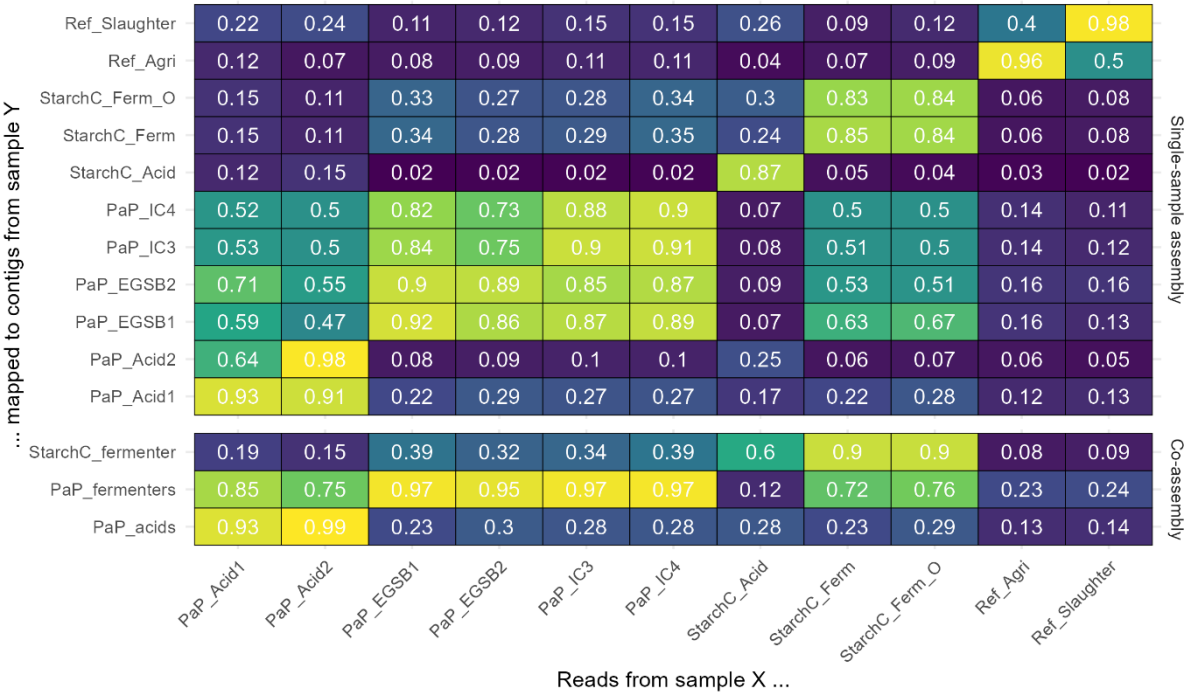

Figure 14: *Results of mapping reads back to assemblies. Numbers in the heatmap indicate the fraction of reads (x-axis) that can be aligned to the contigs in the assemblies (y-axis).*

As indicated by the values in the main diagonal in the upper part of the heatmap which are close to 1, the single-sample assemblies generally represent the sample they are based on well. At least 84 % of the reads can be mapped back. Furthermore, samples from equivalent reactors appear to be similar as well. For example, the reads of all methane reactor samples from the system ‘PaP’ can be mapped to all corresponding assemblies similarly well. The same is observed for the acidification reactors and for the samples Ferm and Ferm_O from the system ‘StarchC’. Moreover, similarities across the equivalent reactors can be observed. The assemblies obtained from the ‘PaP’ methane reactors explain approximately 50 % of the reads of the samples from the corresponding acidification reactors as well as the ‘StarchC’ methane reactor but only around 10 % of the reads in the remaining samples. On the other hand, the assemblies obtained from the ‘StarchC’ methane reactor samples only explain around 30 % of the reads in the ‘PaP’ methane reactor samples. This can be explained by the difference in sequencing yield. The results show again that there is little overlap between the ‘Ref’ samples and the ‘StarchC’ and ‘PaP’ samples. At the same time, they are similar to each other.

The crosstalk that is observed between equivalent reactors for the single-sample assemblies suggests that these samples are similar enough to combine them in a co-assembly, maximising the information that can be used for it. Indeed, the lower part of the plot shows that the co-assemblies represent even higher fractions of the reads in the respective samples. At least 90 % of the reads in the respective samples are represented by the co-assemblies.

Compared to single-sample assemblies, the number of contigs longer than 50,000 bp is increased for all co-assemblies. Moreover, co-assemblies reduce redundancy which is

reflected by the number of contigs that is lower than the sum of contigs in the corresponding single-sample assemblies.

Therefore, for the system ‘PaP’, the co-assembly PaP_acids is selected to retrieve bins from the acidification reactors and the co-assembly PaP_fermenters to retrieve bins from the methane reactors. The co-assembly StarchC_fermenter is chosen to retrieve bins for the methane reactor of the system ‘StarchC’. For all remaining reactors, the single-sample assemblies are used.

3.3.5 Recovering metagenome-assembled genomes (MAGs)

After assembling the metagenomic reads into contigs, MAGs were recovered by clustering the contigs into groups (bins) based on their abundance and sequence structure. This was done for the selected single-sample and co-assemblies using MetaBAT2.

Quality was evaluated using completeness and contamination scores obtained from CheckM2 and low-quality bins were removed from the dataset. The full table of results obtained from CheckM2 can be found in the supplementary material (Table S16). The number of obtained bins as well as their average size and coverage, before and after quality filtering, is given in Table 13.

Table 13: *Assembly-wise basic statistics of obtained bins before and after quality filtering.*

	All bins			After quality filtering		
Assembly	**Nr. of bins**	**Avg. bin size**	**Avg. coverage**	**Nr. of bins**	**Avg. bin size**	**Avg. coverage**
PaP_acids	219	1,077,669	32	51	2,587,260	60
PaP_fermenters	1,440	1,275,847	20	305	2,862,945	32
StarchC_Acid	16	735,131	53	2	1,707,173	99
StarchC_fermenter	52	1,913,929	71	16	3,039,193	156
Ref_Agri	200	551,805	54	11	1,175,345	230
Ref_Slaughter	286	676,332	43	34	1,662,615	105

A large fraction of bins is removed by quality filtering. After this, the average bin size is increased to 1,175,345 to 3,039,193 bp. The average coverage is higher after quality filtering which shows that abundant bins are recovered while low abundance bins are removed.

Table 14 shows the fraction of binned nucleotides, i.e. the summed length of all reads that are found in the bins divided by the summed length of all reads in the assembly, before and after removing low-quality bins.

A large fraction of the original data is not retained in the bins after quality filtering, especially for the assemblies StarchC_Acid, Ref_Agri, and Ref_Slaugher. While it is possible to improve the quality of the removed bins by manual refining them or by using short-read sequencing in a hybrid assembly, this is outside the scope of this thesis.

The distribution of quality scores, i.e. completeness and contamination, across the bins is visualised in histograms in Figure 15, again before and after quality filtering.

Table 14: *Fraction of binned nucleotides of each assembly before and after quality filtering.*

	Fraction of binned nucleotides [%]	
Assembly	**All bins**	**After quality filtering**
PaP_acids	72	40
PaP_fermenters	84	40
StarchC_Acid	51	15
StarchC_fermenter	85	41
Ref_Agri	56	7
Ref_Slaughter	64	19

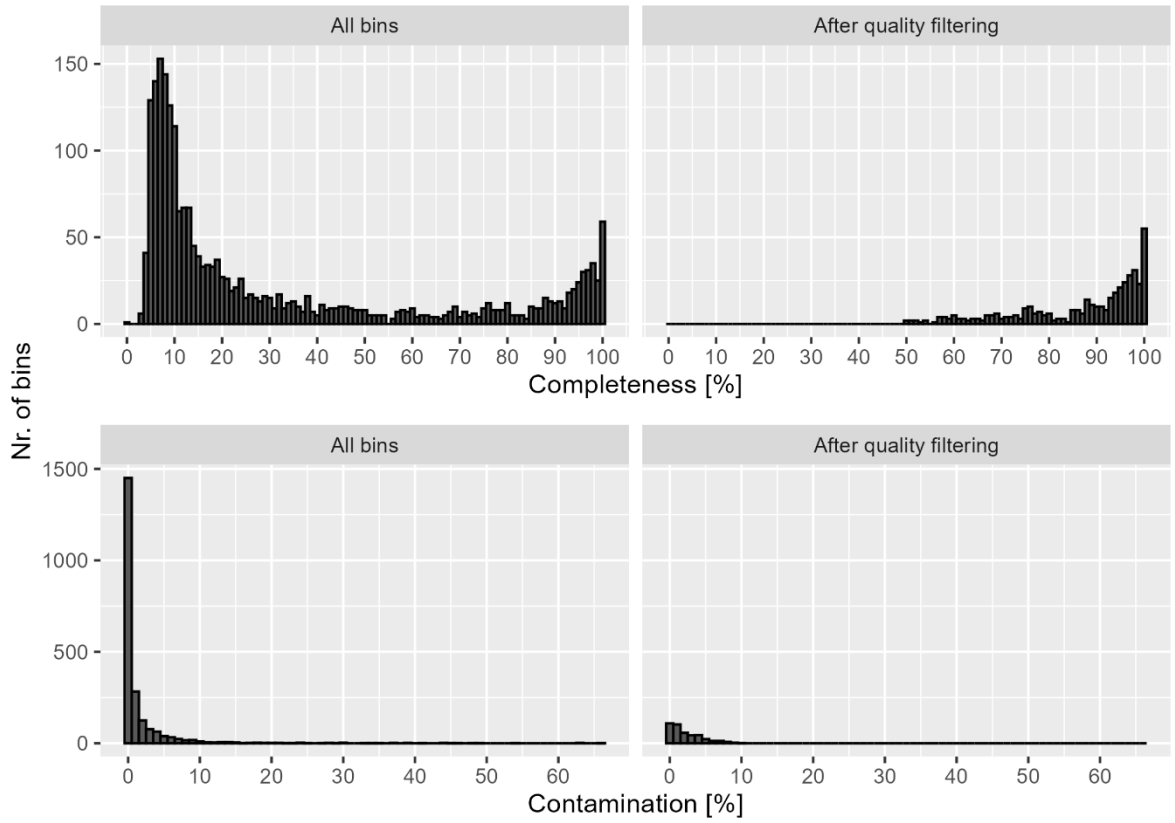

Figure 15: *Histogram of completeness (top) and contamination (bottom) of bins before (left) and after (right) quality filtering, as determined by CheckM2.*

It is clearly visible that a large number of bins in the original dataset has low completeness scores. These represent incomplete genomes, or extrachromosomal elements, and are successfully removed with the chosen threshold for quality filtering, while retaining the complete genomes. Contamination is similarly distributed before and after quality filtering. Thus, it appears that MetaBAT2 is minimizing contamination while tolerating incomplete genomes.

The quality filtered set of bins was dereplicated at ANI >95 % using dRep. The obtained clustering dendrograms can be found in the supplementary material (SF2 and SF3). 20 clusters of bins with average nucleotide identity (ANI) >95 % are identified by the tool. As one would

expect, most clusters contain one bin from the acidification reactor and one bin from the methane reactor assembly of the same system. This most likely results from microorganisms being transferred from the acidification to the methane reactor within the process. However, some bins with ANI >98 % originate from the assembly of the methane reactors of different systems. This shows that there is some redundancy across the different systems.

In summary, of the 2213 bins obtained from MetaBAT2, 419 bins are of high quality and the final dataset after dereplication comprises 398 metagenome-assembled genomes (MAGs).

The MAGs were taxonomically classified using GTDB-Tk and relative abundances were determined for all samples using CoverM (see supplementary Tables S16 and S17). The resulting taxonomic profiles are visualised in a bubble plot in Figure 16.

The final set of MAGs comprises 30 archaeal and 368 bacterial MAGs. In total, the MAGs are assigned to 46 different phyla, of which 7 are archaeal and 39 are bacterial phyla. Compared to the read-based taxonomic profiling (11 archaeal and 63 bacterial phyla), some phyla are not represented by a MAG. However, the most prevalent phyla are captured by the MAGs. Furthermore, the most diverse phyla, i.e. phyla with the largest number of different species, are the same for MAG representatives and read-based taxonomic profiling. They are Bacillota, Bacteroidota, Actinomycetota, Patescibacteriota, and Chloroflexota. Thus, there MAG representatives for the most diverse phyla were obtained.

Actinomycetota (especially *Bifidobacteria*) seem to be specific for the system ‘PaP’, and especially the acidification reactor. Bacillota are detected in all systems, but they are notably highly abundant in the acidification reactor (*Lactobacillus amylolyticus* and *Liquorilactobacillus vini*) and the methane reactor (genus *JAQWCN01*) of the system ‘StarchC’. The MAGs that are classified as Chloroflexota, Desulfobacterota, and Fermentibacterota are mainly found in the methane reactor of ‘PaP’. Possibly, they also occur in ‘StarchC’ but are not detected due to the lower sequencing depth. Overall, there are again on the one hand overlaps between ‘PaP’ and ‘StarchC’, and on the other hand between both ‘Ref’ samples.

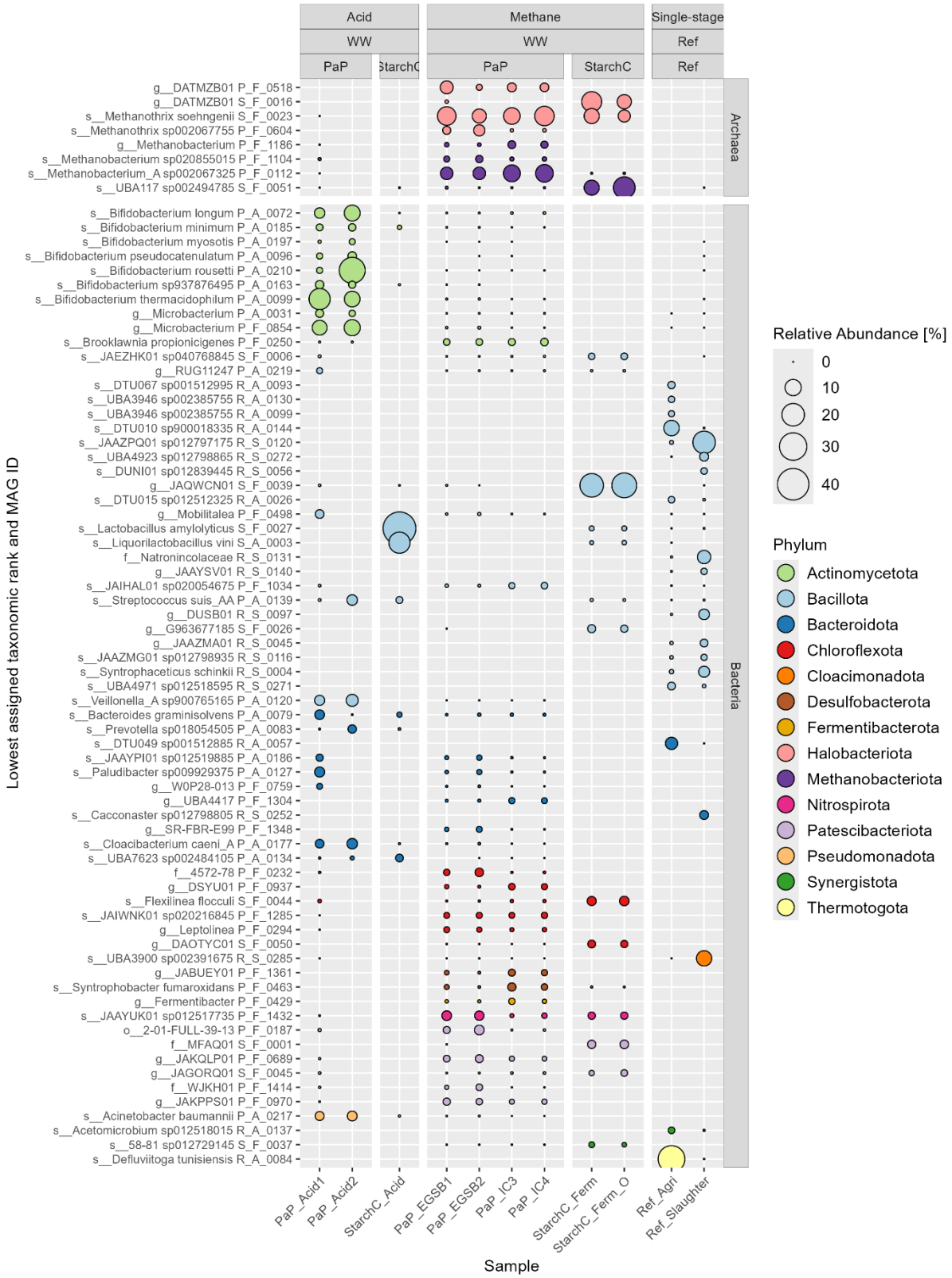

Figure 16: *Bubble plot for final set of MAGs. The size of the bubbles indicates the relative abundance of the MAG (y-axis) in the sample (x-axis). Only MAGs with minimum relative abundance of 1 % in any of the samples are shown. The bubbles are coloured by phylum.* ***Vertical layers*** *separate archaea and bacteria. The samples (x-axis) are organised by three horizontal layers.* ***Top layer:*** *Type of system they originate from. 'Single-stage': Typical single-stage systems. 'Acid': Acidification reactor in two-stage systems. 'Methane': Methane reactor in two-stage systems.* ***Middle layer:*** *Type of substrate that is used in the process. 'Ref': Reference samples; 'WW': Industrial wastewaters.* ***Bottom layer:*** *System names as used in Table 5 (section 2.3.1.1).*

3.3.6 Metadata analysis

Process parameters that were provided by the companies who are responsible for the AD systems, as well as physicochemical parameters that were determined for the provided samples, were analysed. The aim was to provide context for the sequencing results. The full results are available in the supplementary material (Table S2: physicochemical parameters; S10: process parameters).

The efficiency of the AD systems is monitored by the companies. The efficiency of removing organic matter from the wastewater, as relationship of organic matter in the influent and effluent, is shown in Figure 17A. The methane content in the produced biogas is shown in Figure 17B. Both parameters are shown for a time period of 12 days before to 12 days after the sampling date.

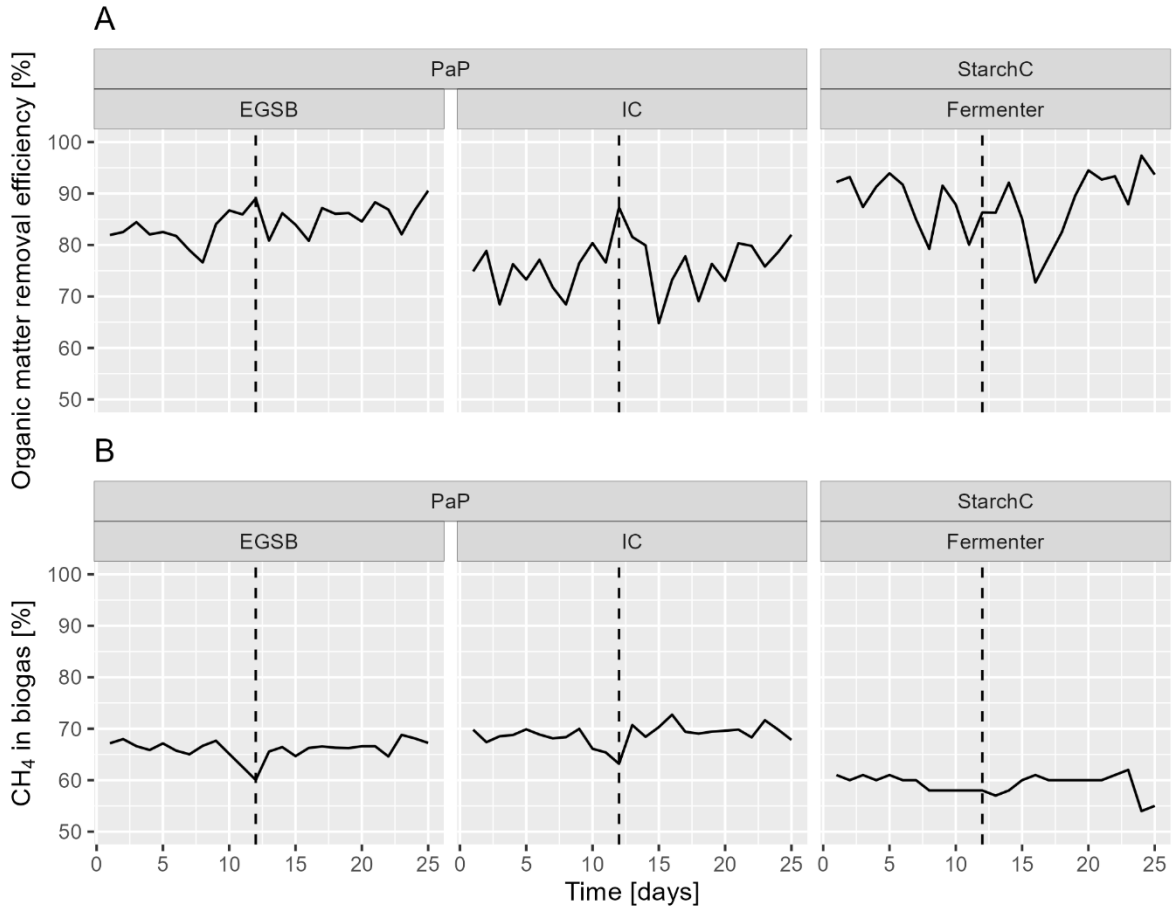

Figure 17: *Process parameters of anaerobic digestion systems in sampling period. The period of 12 days before to 12 days after the sampling day is shown. The sampling date is marked by a vertical dashed line. The graphs are organised by two horizontal layers.* ***Top layer:*** *System names as used in Table 5 (section 2.3.1.1).* ***Bottom layer:*** *Subsytem (described in section 2.4.1).* ***A:*** *Efficiency of organic matter removal, calculated as fraction of organic matter in the effluent with regards to organic matter in the effluent. Organic matter is approximated by chemical oxygen demand (COD, system 'PaP') or total organic matter (TOC, system 'StarchC').* ***B:*** *Fraction of methane in the biogas (offgas of the system). Note the range on the y-axis.*

The organic matter is removed from the wastewater efficiently in all systems. The system 'StarchC' has a generally higher removal efficiency than 'PaP', but it is fluctuating between 70 % and more than 95 %. In 'PaP', the wastewater stream is split into two parts and treated in two separate systems. One part is treated in two expanded granular sludge bed (EGSB) reactors, and the other one in two internal circulation (IC) reactors. The EGSB system is slightly more efficient and stable with a removal rate of 80-90 %, while the IC system shows more

fluctuation and has a removal rate of 65-85 %. At the time of sampling, it is 85-90 % of organic matter are removed in all systems.

Methane is the main energy carrier in biogas, and a high concentration is therefore desired. Typically, it makes up 50-75 % of the biogas (Braun 2007). The methane content in the biogas is stable in all systems. EGSB and IC in the ‘PaP’ system show a stable methane content of approximately 65 %. However, it drops to 60 % in both subsystems on the sampling day but quickly recovers. In the ‘StarchC’ system, the methane content is generally lower but also stable around 60 %.

The pH value is an important monitoring parameter because it provides a first estimate of the state of the process. Moreover, the dissociation of potentially inhibitory substances such as ammonia and hydrogen sulphide are dependent on the pH (Drosg 2013). The pH values that were measured in the samples from the systems ‘PaP’ and ‘StarchC’ are given in Table 15.

Table 15: *pH of samples in 'PaP' and 'StarchC' system. For description and schemes of the systems and sample names see 2.4.1.*

PaP						StarchC		
Acid1	**Acid2**	**EGSB1**	**EGSB2**	**IC3**	**IC4**	**Acid**	**Ferm**	**Ferm_O**
7.4	4.5	8.5	7.2	8.3	8.5	3.7	8.1	8.0

A two-stage AD process aims at the production of volatile fatty acids in the acidification reactor. The optimal pH range for this process has been reported to be 4-6.5 (Speece 1996) and is inhibited at extreme conditions, such as pH below 3 or above 12 (Feng et al. 2022). The pH in Acid2 in the ‘PaP’ system is within the optimal range while it is slightly below in Acid in the ‘StarchC’ system. However, the pH in Acid1 in the ‘PaP’ system is 7.4 which indicates a process disturbance. A neutral or slightly alkaline pH is typical for methane reactors (Drosg 2013). All methane reactors (EGSBs, ICs, Ferm, Ferm_O) have a pH between 7.2 and 8.5 which is in the optimal range for methanogens (Speece 1996).

A more detailed insight into the process can be obtained by the volatile fatty acid (VFA) profile. The VFA profiles that were determined for the analysed samples are presented in Figure 18.

VFAs and alcohols are produced in the acidogenesis step (Drosg 2013) and are typically found at high concentrations in the acidification reactor. They are consumed in the acetogenesis and methanogenesis steps (Drosg 2013) and should therefore be low in the methane reactor. This relationship can be observed in the investigated systems, except for the acidification reactor 1 in the system ‘PaP’ (PaP_Acid1). The VFA profile is similar to that in the other acidification reactor of the same system (PaP_Acid2), but the concentrations are lower. It appears that VFAs are either not produced at the same rate as in PaP_Acid2, or they are starting to be consumed in the same reactor. This further substantiates the suspicion of the process disturbance suggested by the pH.

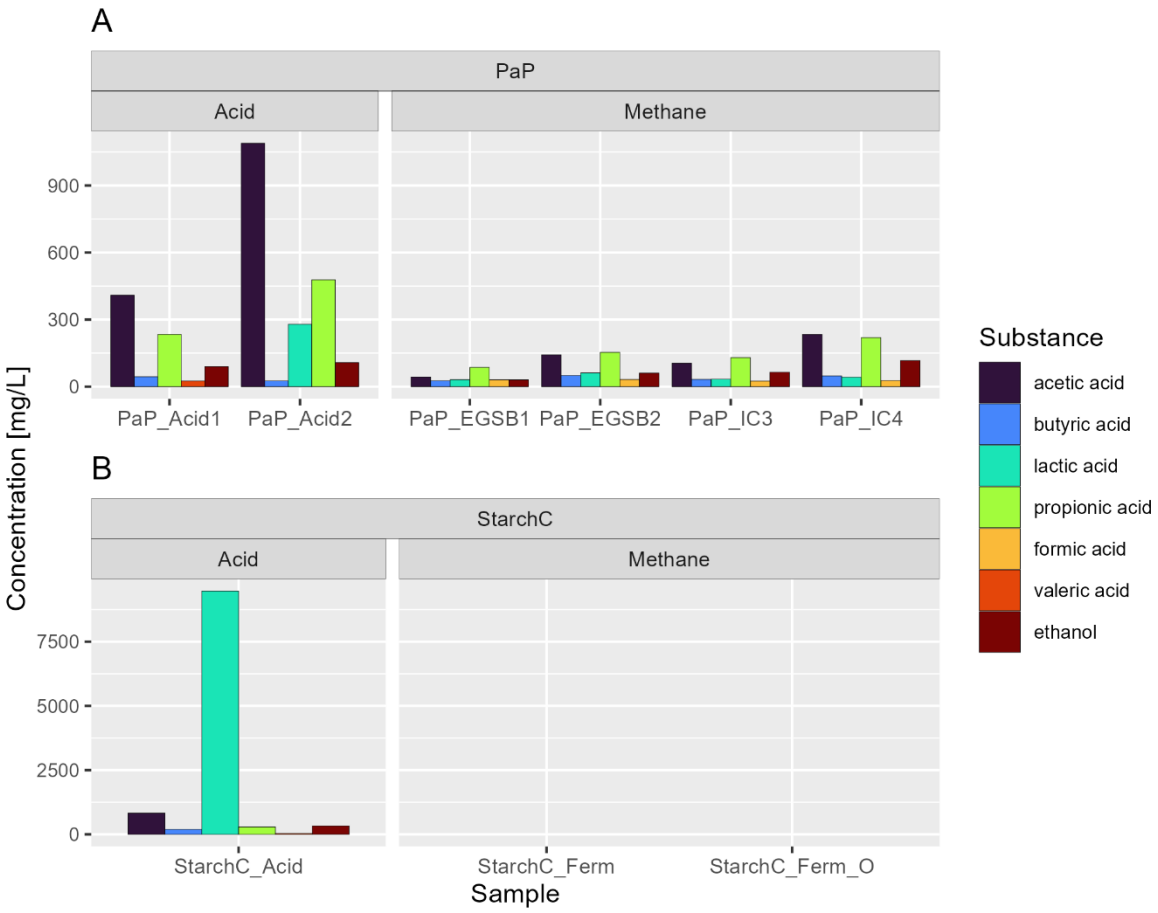

Figure 18: *Concentration of volatile fatty acids and alcohols in the liquid fraction of the samples.* ***A:*** *Samples from system 'PaP'.* ***B:*** *Samples from system 'StarchC'. Note that the y-axis scale is different in the upper and lower part of the plot due to the different range of values. Data was obtained by High Performance Liquid Chromatography (HPLC) analysis. The samples (x-axis) are organised by a horizontal layer according to the reactor type they originate from. 'Acid': Acidification reactor in two-stage systems. 'Methane': Methane reactor in two-stage systems.*

When comparing the VFA profiles in the acidification reactors of both systems (with focus on Acid2 for the 'PaP' system), major differences are observed. The main VFA in 'StarchC' is lactic acid. Its considerably high concentration (9.5 g/L) is presumably introduced with the wastewater and not produced inside the reactor. The wastewater is from a corn starch factory. Corn steeping is an important step in corn production and includes an active *Lactobacillus* sp. fermentation which yields lactic acid (Eckhoff and Watson 2009). Other VFAs are present at similar levels as in the system 'PaP', with acetic acid as the main acid (around 1 g/L). Besides these, butyric acid, propionic acid, valeric acid, and ethanol are detected in 'StarchC'. Interestingly, no VFAs were detected in the methane reactor which suggests complete conversion to AD products such as methane, carbon dioxide, and hydrogen. Besides acetic acid, propionic and lactic acid are detected at considerable concentrations in Acid2 of the 'PaP' system. Nevertheless, the lactic acid concentration (0.3 g/L) is far below that in 'StarchC'. Furthermore, ethanol and butyric acid are found. In contrast to 'StarchC', VFAs and alcohols are not completely consumed in the methane reactors. Ethanol concentration is approximately half of that in the acidification reactor and the majority of acetate, propionate, and lactate is consumed. However, the concentration of butyric acid is not considerably different in both subsystems. Moreover, formate is only detected in the methane reactors.

The presence of sulphate in AD systems can be an issue. Sulphate is reduced to sulphide by sulphate-reducing bacteria (SRB) (Stefanie et al. 1994). This leads to increased levels of dissolved sulphide in the reactor which inhibits anaerobic bacteria and methanogens (Stefanie

et al. 1994). Furthermore, it leads to increased levels of gaseous H_2S in the biogas, which is corrosive and toxic to humans (Stefanie et al. 1994). The concentrations of sulphate determined in the analysed samples is given in Table 16.

Table 16: *Sulphate concentrations [mg/L] in the liquid fraction of samples in 'PaP' and 'StarchC' system. Data was obtained using ion chromatography.*

PaP						StarchC		
Acid1	**Acid2**	**EGSB1**	**EGSB2**	**IC3**	**IC4**	**Acid**	**Ferm**	**Ferm_O**
0.0	50.4	0.0	0.0	0.0	0.0	139.2	0.0	0.0

Both investigated AD systems are associated with industries that are known to produce sulphate-rich wastewaters (Stefanie et al. 1994; Möbius 2010). It is therefore not surprising that sulphate was detected in one acidification reactor in the 'PaP' system (PaP_Acid2) and in that of the 'StarchC' system (StarchC_Acid). However, the concentration is below 200 mg/L in both systems which is not considered problematic (Stefanie et al. 1994). No sulphate is detected in the sample PaP_Acid1 or in the methane reactors. This suggests that the sulphate is reduced to sulphide, and thus the organic matter degraded, by SRB. This means that SRB might consume metabolites that could otherwise be used by methanogens for methane production. Relating the observed sulphate concentrations to the observed VFA concentrations suggests that this is only a small fraction of the VFAs and the competition is negligible. However, the resulting sulphide is inhibitory for methanogens (Madden et al. 2014).

3.3.7 Functional analysis

To obtain insights into the functional potential of the recovered MAGs, they were annotated using the KOfam database (Aramaki et al. 2020). Based on these KEGG orthologs annotations, their potential for key metabolic pathways in AD was investigated.

The functional analysis is based on KEGG (Kanehisa and Goto 2000) which has a system of orthologous groups of proteins identified by K numbers. Moreover, KEGG provides modules which define metabolic pathways by the K numbers of enzyme cascades. Thus, the annotations can be used to infer functions to MAGs, either on the individual enzyme or on the pathway (module) level.

The K number annotations of the obtained MAGs can be found in the supplementary material (Table S19). In total, there are hits for 5487 distinct orthologs in the MAGs. Of these, only 1309 K numbers are associated with KEGG modules. Thus, the majority of identified orthologous groups is not associated with modules. This suggests that the metabolic processes in the investigated systems are not well characterized and potentially novel information can be obtained from this analysis. However, it also means that a large fraction of data is omitted in a solely module-based analysis. Moreover, when inferring functions by evaluating module completeness alone, enzyme-level information, which provides insights into which reactions of a pathway are not encoded, is overlooked. For these reasons, manual module-like pathway definitions and enzyme-level information are used in the functional analysis.

3.3.7.1 Mapping MAGs to pathways and metabolites

In the following paragraphs, the results of investigating the genomic potential for the selected pathways will be presented. Only high abundance MAGs (>1 % in at least one sample) are considered.

Lactic acid fermentation

Four types of lactic acid fermentation are investigated. Apart from lactic acid, each yields different fermentation products. As seven MAGs are classified as *Bifidobacterium*, Bifidobacterium fermentation was selected for a detailed analysis. The schematic pathway map for this fermentation, which yields acetate and lactate, is shown in Figure 19. The KEGG hits for the associated enzymes are visualised in a heatmap in Figure 20. Both figures were used in combination to identify MAGs that are capable of Bifidobacterium fermentation.

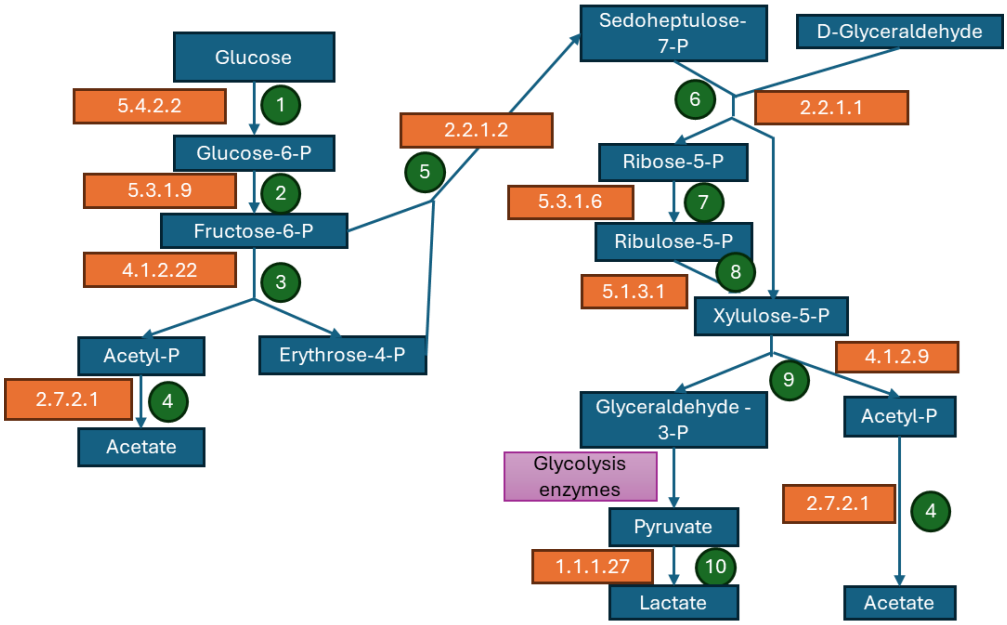

Figure 19: *Sketch of pathway map for Bifidobacterium fermentation. Blue squares indicate metabolites and are connected by arrows indicating enzymatic reactions. Orange squares indicate the EC numbers for enzymes that catalyse the reactions. The purple square is a placeholder for multiple steps of glycolysis that were not investigated. Green circles indicate the step number, as assigned in the pathway definition in supplementary Table S18.*

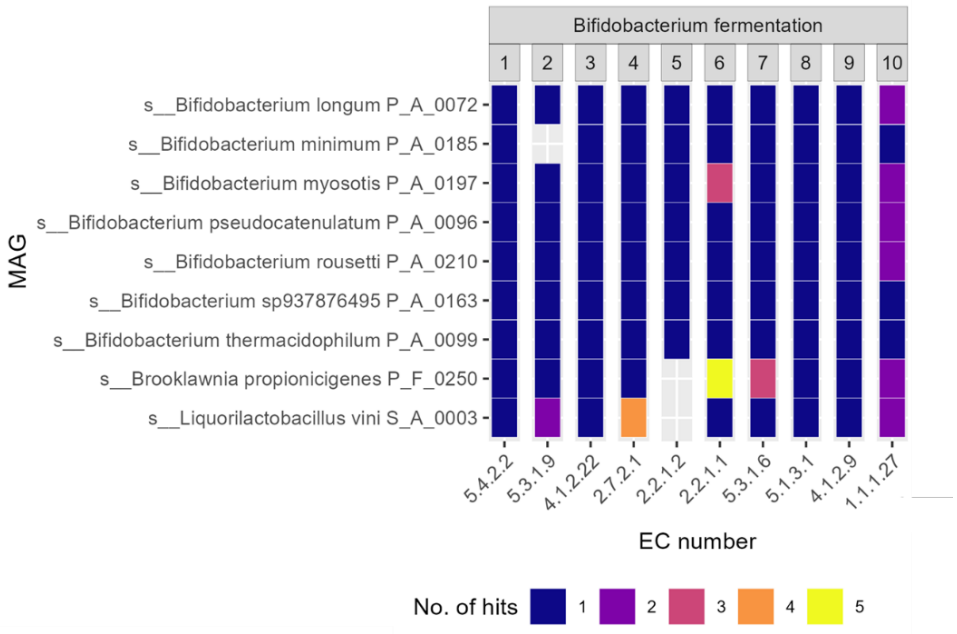

Figure 20: *Heatmap showing the KEGG hits for enzymes associated with Bifidobacterium fermentation in MAGs. Numbers on the top x-axis indicate the step number, as assigned in the pathway definition in supplementary Table S18 and depicted in Figure 19.*

As expected, *Bifidobacterium* is the main genus with hits for enzymes for this pathway. In addition, MAGs P_F_0250 (*Brooklawnia propionicigenes*) and S_A_0003 (*Liquorilactobacillus vini*) encode nine out of ten enzymes. However, they are missing a transaldolase (EC 2.2.1.2) which catalyses the important conversion of fructose-6-phosphate and erythrose-4-phosphate to sedoheptulose-7-phosphate. They are therefore not included in the set of MAGs assigned to Bifidobacterium fermentation. P_A_0185 (*Bifidobacterium minimum*) is missing the enzyme that converts glucose-6-phosphate to fructose-6-phosphate (EC 5.3.1.9). On the one hand, this could be attributed to the low completeness of this MAG of 62 %. On the other hand, the name suggests a reduced genome. The other MAGs encode all enzymes. Thus, all members of the genus *Bifidobacterium* that are shown in Figure 20 are assigned to Bifidobacterium fermentation and thus to the production of acetate and lactate.

A similar procedure was used to assign MAGs to the other pathways. In the following, the procedure will not be explained step by step as above, but the results of the analysis will be elaborated. The final assignments of MAGs to pathways and the associated metabolites are depicted in Figure 21 and listed in the supplementary material (Table S20).

Heterofermentative lactic acid fermentation yields, besides lactate, acetate if the substrates are pentoses and ethanol if the substrates are hexoses. Most, but not all, *Bifidobacteria* are capable of utilising both kinds of sugar for lactic acid fermentation. In addition, P_F_0250 (*Brooklawnia propionicigenes*) and S_A_0003 (*Liquorilactobacillus vini*) are capable of heterofermentative lactic acid fermentation. Thus, they are assigned to lactate, ethanol, and acetate formation (see Figure 21).

Besides the described pathways that yield a mix of acids and alcohols, lactic acid is formed by homofermentative lactic acid fermentation as the only product. Lactate dehydrogenase (EC 1.1.1.27) is used to convert pyruvate to lactate and is encoded by various MAGs of the phyla Actinomycetota, Bacillota, Bacteroidota, and Chloroflexota.

Acetic acid and ethanol are not exclusively produced in the described lactic acid fermentations but are side products of various metabolic pathways. The enzymes involved in the final reactions that yield acetate or ethanol usually work in both directions. Therefore, MAGs from multiple phyla are assigned to either utilising or producing them (see Figure 21).

Other volatile fatty acids

Acetic and lactic acid are the main acids that were detected in the acidification reactors. Their formation can be explained by the pathways described above. Besides, propionic acid is a major component of the liquid fraction in the acidification reactor 2 in the system ‘PaP’ (see Figure 18 in section 3.3.6). Moreover, butyric acid is detected in all acidification reactors, but at lower concentrations.

Propionic acid can be produced or oxidised by the methylmalonyl or the acryloyl pathway. Both pathways generally work in both directions. MAG P_F_0250 (*Brooklawnia propionicigenes*) is the only MAG for which all enzymes of the methylmalonyl pathway are identified. It is not clear from this analysis alone whether it produces or oxidises propionic acid.

The MAG S_F_0039 (class Clostridia, family Eubacteriaceae) encodes all enzymes except one (EC 4.2.1.54) for the acryloyl pathway (propionate utilisation or oxidation) and all enzymes for butyrate oxidation. It is the most abundant MAG in the methane reactor of ‘StarchC’ and only traces are detected in ‘PaP’. Butyrate appears not to be metabolised in the system ‘PaP’ because the concentrations are uniform across all reactors, while it is completely removed in the methane reactor for ‘StarchC’. Thus, it is not surprising that a butyrate oxidiser is only detected in ‘StarchC’. A similar pattern is noticed for propionic acid which is completely removed in the methane reactor of ‘StarchC’ and, in the methane reactors of ‘PaP’, only reduced by approximately half compared to the acidification reactors.

Acetogenesis

Three MAGs that have the genomic potential for acetogenesis have been identified. Two of them (P_F_1361 and P_F_0463) use sulphate as electron acceptor and are described in more detail below. They are identified in the methane reactors of 'PaP', with higher abundance in the IC reactors than in the EGSB reactors. S_F_0050 is identified in the methane reactor of 'StarchC' and it is assumed that it represents the main responsible organism for the conversion of H_2 and CO_2 to acetate in this reactor, thus providing substrate for acetoclastic methanogens.

Methanogenesis

They key enzyme that is shared by all methanogenesis pathways is methyl-coenzyme M reductase (MCR; EC:2.8.4.1) (Harirchi et al. 2022). All archaeal MAGs encode this enzyme and are thus capable of the core part of methanogenesis (reduction of coenzyme M to methane).

In accordance with previous findings suggesting that the main methanogenesis pathways in AD are acetoclastic and hydrogenotrophic (Harirchi et al. 2022; Niya et al. 2024; Kim et al. 2022), there are no hits for the genes encoding the enzymes specific for methylotrophic methanogenesis.

Interestingly, all archaeal MAGs encode the enzymes for the reactions specific to acetoclastic as well as those specific for hydrogenotrophic methanogenesis. This indicates that all are capable of both pathways which is contrary to what has been described in literature. Four of the MAGs are classified as Methanotrichaceae (phylum Halobacteriota, class Methanosarcinia) who have previously been described as acetoclastic (Ren et al. 2025; Kim et al. 2022; Niya et al. 2024). It has been suggested that they are obligate acetoclastic (Theuerl et al. 2015) but there has been evidence that they are additionally capable of using the hydrogenotrophic pathway (Niya et al. 2024; Thauer et al. 2008; Lyu et al. 2018). The remaining four MAGs are classified as Methanobacteriaceae (phylum Methanobacteriota, class Methanobacteria) who have previously been described as hydrogenotrophic methanogens, without the ability to utilise acetate (Niya et al. 2024; Dueholm et al. 2024; Kim et al. 2022).

Interestingly, all except one (P_F_0518; family Methanotrichaceae) archaeal MAGs encode a coenzyme F_{420}-associated formate dehydrogenase (K00125; Fdh). This enzyme reduces coenzyme F_{420} which is consecutively used to produce H_2. The produced H_2 is used for hydrogenotrophic methanogenesis (Lupa et al. 2008). Thus, they can presumably utilise formate as initial substrate in hydrogenotrophic methanogenesis.

Sulphate reduction

The process of sulphate reduction is in direct competition for substrates with methanogenesis (Harirchi et al. 2022). It is of special interest in this analysis because the wastestreams that are treated in the investigated systems have high sulphate content and it was observed that sulphate is completely removed during AD, presumably by sulphate reduction (see section 3.3.6).

The enzymes for dissimilatory sulphate reduction are encoded in three of the MAGs. They are detected in the methane reactors of both systems, 'PaP' and 'StarchC'. They are not detected in the 'Ref' samples which suggests that this pathway is specific for these systems. Two of the MAGs are classified as Syntrophobacteraceae (phylum Desulfobacterota, class Syntrophobacteria; P_F_0463 and P_F_1361) and one is classified as Dissulfurispiraceae (phylum Nitrospirota; P_F_1432). Interestingly, the Syntrophobacteraceae MAGs are more abundant in IC reactors in 'PaP' while the Dissulfurispiraceae MAG is more abundant in EGSB reactors of 'PaP' and in the methane reactor of 'StarchC'. None are detected in the acidification reactors.

Syntrophobacteraceae, in particular *Syntrophobacter fumaroxidans*, has been characterized as syntrophic propionate-oxidising bacterium which oxidises propionate and provides H_2 and CO_2 to its syntrophic partners, hydrogenotrophic methanogens (Harmsen et al. 1998; Westerholm et al. 2021; Doloman et al. 2024; Cai et al. 2016). For the oxidation of propionate, it uses sulphate as electron acceptor (Harmsen et al. 1998). The enzymes necessary for propionate oxidation are encoded in both Syntrophobacteraceae MAGs which indicates that both are taking this route. In addition, the enzymes for acetogenesis are found in both MAGs. This indicates that they further oxidise the produced acetate to CO_2. Thus, they live in syntrophy with hydrogenotrophic methanogens by providing H_2 but are competing with acetoclastic methanogens or acetate oxidising bacteria for acetate (Westerholm et al. 2021). *Syntrophobacter fumaroxidans* has been associated with the formation of granular biomass (Doloman et al. 2024; Westerholm et al. 2021; Madden et al. 2014) which is in accordance with mainly observing them in the IC reactors, where the biomass is immobilised in granules.

In contrast, MAG P_F_1432 (phylum Nitrospirota, order Thermodesulfovibrionales, family Dissulfurispiraceae) does not encode the propionate oxidation or acetogenesis pathway, but it encodes formate hydrogenases and the enzymes for carbon fixation from CO_2. It is hypothesized that it is a hydrogenotrophic sulphate reducer which competes with hydrogenotrophic methanogens.

Figure 21: *Heatmap that indicates the assignment of MAGs to the investigated pathways. Only MAGs > 1 % abundance are shown. Vertical layers separate archaea and bacteria. Horizontal layers organise the pathways into the most relevant metabolites.*

3.3.7.2 System-level analysis

After analysing each pathway on its own, they are put into relation with each other. Moreover, the occurrence of MAGs and their associated pathways is mapped to the reactors of both systems, ‘PaP’ and ‘StarchC’. This is supposed to provide additional evidence for potential roles of the identified MAGs in the AD process.

The relationship of the investigated metabolites and pathways and their mapping to the acidification or methane reactors are depicted in Figure 22. In addition, taxonomic bar plots specific for each of the reactors are shown to visualise which phyla are associated with which steps.

In the acidification reactor of the system ‘PaP’, Actinomycetota appear to be involved in all pathways to produce acetate, lactate, propionate, and ethanol. They are the most abundant phylum in this reactor (see Figure 16). The main genus is *Bifidobacterium* who is responsible for acetate, lactate, and ethanol production while propionate is produced by *Microbacterium* and MAGs from the phylum Bacillota. In contrast, Bacillota is the main phylum responsible for acetate, lactate, propionate, butyrate, and ethanol production in ‘StarchC’. Two MAGs of this phylum are highly abundant in the acidification reactor and therefore it is expected that they are responsible for the production of the main metabolites. This shows that especially the first steps of AD, where the differences in the substrate components are most important, can be accomplished by different taxa. However, it has to be noted that the sequencing depth is not sufficient to be confident about the absence of other taxa.

Butyrate metabolism is only detected in the system ‘StarchC’. In ‘PaP’ butyrate does not seem to be metabolised at all while Bacillota are associated with it in both ‘StarchC’ reactors.

Comparing the association of taxa with metabolites between the acidification and methane reactor within ‘PaP’, large differences are observed. For example, besides Actinomycetota, Chloroflexota, Bacteroidota and Bacillota are associated with lactate in the methane reactors. Presumably, Actinomycetota produce it while the members of the other phyla rather metabolise it to acetate. In contrast, Bacillota is the only phylum associated with lactate, propionate, and butyrate in both reactors of ‘StarchC’. This can again be attributed to their high abundance and low sequencing depth for the associated samples.

As described above, acetate and ethanol are metabolites of various pathways, and it is therefore not surprising that multiple phyla are associated with both compounds.

Formate seems to be associated with the last steps of AD. The phyla that encode formate dehydrogenases are mainly methanogenic archaea, besides sulphate reducers and acetogens. This is similar in both systems.

Acetogenesis and sulphate reduction are only detected in bacteria in methane reactors. Interestingly, acetogenesis is almost exclusively associated with the sulphate reducing Desulfobacterota in ‘PaP’, while almost exclusively Chloroflexota are capable of it in ‘StarchC’. Methanogenesis is exclusively detected in archaea occurring in the methane reactors, as expected.

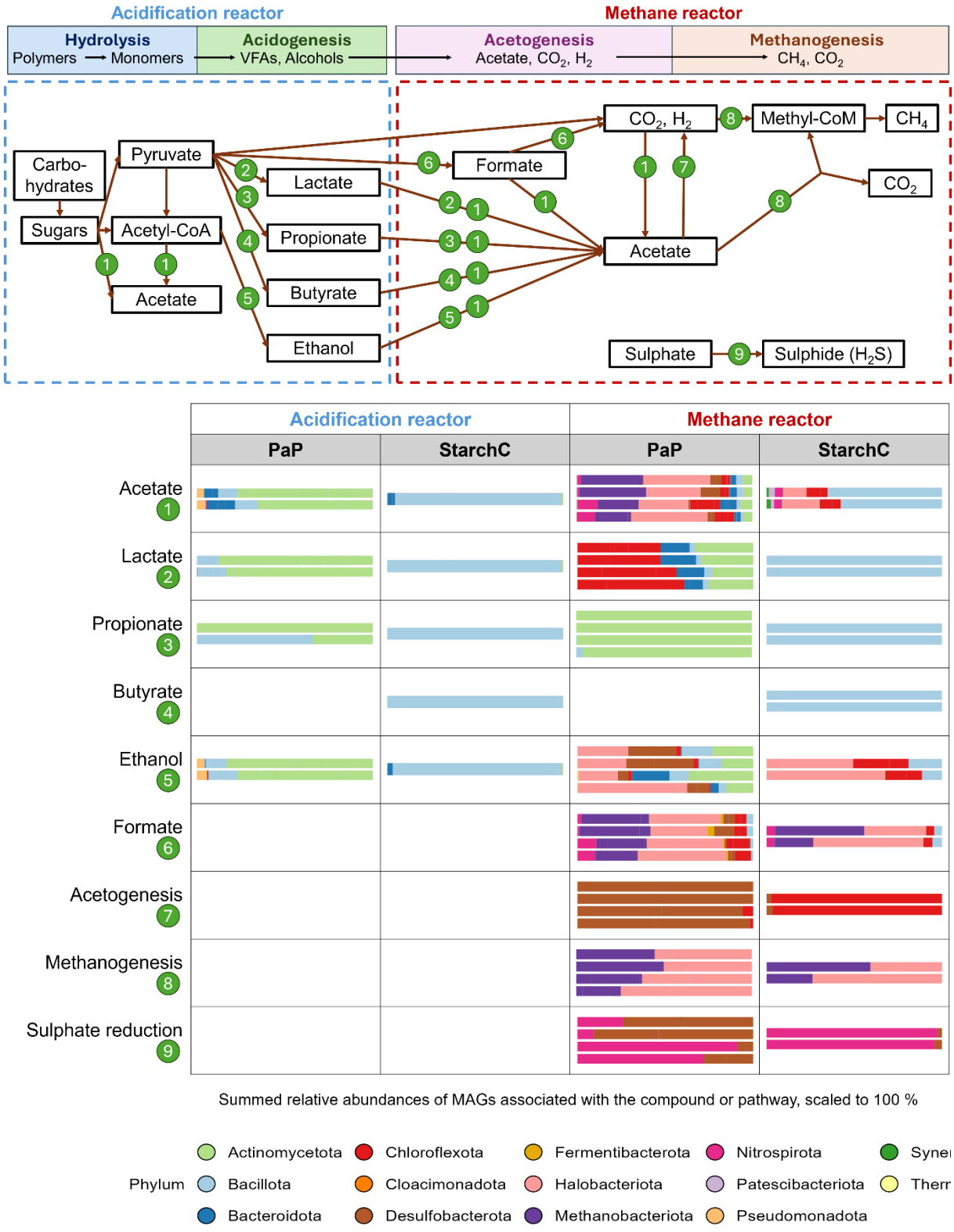

Figure 22: *Key anaerobic digestion pathways and phylum level taxonomy of the MAGs linked to each process. Numbers in green circles indicate the reactions in the pathway map that are associated with the metabolites in the table. Bars include all MAGs > 1 % abundance. The bars show the summed relative abundances of MAGs that are associated with the metabolites or pathways on the y-axis. The figure is designed to emphasize the contribution of each phylum to the corresponding metabolic pathways. Due to scaling the bars to 100 %, the figure does not reflect their contribution to the overall process.*

4 Discussion and future perspectives

In this thesis, the microbial communities in various anaerobic digestion (AD) systems were investigated. Previous studies have mostly focused on systems that use common substrates such as agricultural feedstock, manure, or activated sludge (Campanaro et al. 2018; Dueholm et al. 2024; Jun et al. 2025). They have identified key phyla in the AD microbial community. However, it has been shown that the used substrate strongly influences the microbial community composition (Dueholm et al. 2024; Harirchi et al. 2022) and thus, findings might not be generalisable across systems that use different substrates. An aim of this thesis was therefore to provide insights into yet understudied systems. AD systems which use uncommon substrates such as agro-industrial residues or industrial wastewaters were thus included in the amplicon-sequencing dataset that was analysed in the first part of this thesis.

A phylum-level analysis of the taxonomic profiles revealed some core phyla that match with previous studies (Theuerl et al. 2015; Campanaro et al. 2016; Cai et al. 2016; Treu et al. 2016). Bacillota and Bacteroidota, which are prevalent in all of the analysed samples, have previously been described as key phyla in AD. Their distinct abundances in the samples suggest that their members are involved in the hydrolysis and acidification step of AD. Other phyla were detected only in specific samples which hints at their specific tasks or preferred environments. Despite the similarities that were observed at the phylum level, an ASV level analysis revealed that most taxa are indeed system-specific or even reactor-specific. This shows that an analysis at the phylum-level is not sufficient to taxonomically describe the microbial community of a system and can only provide a broad overview. It was therefore beneficial to include a more detailed ASV level analysis.

To provide a broader context and find more global patterns regarding the microbial community composition, a publicly available dataset from the MiDAS project (Dueholm et al. 2024) was included in the analysis. The sample similarity of the merged dataset was investigated in an ordination. One limitation of this analysis is that the MiDAS 5 database consists to a large extent of samples from anaerobic digesters that treat activated sludge. Thus, the dataset is biased to this substrate type while other substrates are underrepresented. However, substrate types that are similar to those used in the herein investigated systems as well as other, more common substrate types, are also represented in the dataset. Furthermore, the dataset includes thermophilic systems. Moreover, it was the largest uniformly created dataset that could be found. The analysis revealed that the substrate used in the AD shape the microbial community to some extent and that the samples analysed in this thesis are indeed distinct from the majority of the samples in the MiDAS 5 database. This emphasizes the importance of investigating AD systems that use uncommon substrates and that this thesis provides insights into the specifics of the microbial community in an understudied system type.

By comparing the taxonomic classification of the ASVs obtained from using three databases, it was shown that the AD microbial community is not represented well by the broadly used SILVA database (versions 138.1 and 138.2). It became evident that using a database which is specific for a similar environment, specifically MiDAS 5 (Dueholm et al. 2024), results in the classification of a much larger fraction of ASVs at genus and species rank, despite being smaller with less overall entries. These results show that taxonomic classification largely depends on the database that is used and that broad databases are not always the best choice. Especially in AD, the taxonomy of process-specific microorganisms is still poorly characterised (Dueholm et al. 2024; Treu et al. 2016) and thus novel taxa are expected. It was shown in this thesis that using a specialised database can improve the results in this context substantially.

The specific functions of taxa in AD and their complex interactions have not yet been studied in detail and their metabolic potential is largely unknown. Therefore, a set of selected samples

was analysed using metagenomic methods in the second part of this thesis. Based on the results obtained from the amplicon-based analysis and on sample availability, the systems ‘PaP’ and ‘StarchC’ were selected as they represent understudied systems which was observed in the ordination with the samples from the MiDAS database. These systems were expected to provide new insights as their microbial communities were shown to be most distinct from systems that use more common substrates. Furthermore, they are two-stage systems which facilitates functional analysis.

After long-read sequencing of the samples, quality control of the results revealed that the sequencing yield for the samples from the system ‘StarchC’ was lower than for the ‘PaP’ samples. Therefore, the coverage of both systems was not uniform. In ‘StarchC’, only the high-abundance MAGs could be recovered while more of the diversity was captured in ‘PaP’. To obtain valid results, it was therefore necessary to adapt the original aims of directly comparing the systems. The number of MAGs that could be recovered from ‘PaP’ is large and allowed a detailed analysis. Therefore, the focus was laid on the system ‘PaP’, while the system ‘StarchC’ was analysed separately, but still used to provide additional context. In a future study, the reasons for this underperformance could be determined to gain more information about ‘StarchC’.

A functional analysis was performed to obtain insights into the potential roles of the MAGs recovered from the sequencing data. This analysis is based on KEGG, using the KOfam database. This database is very broad and not specialised in the AD environment. As it was shown earlier for taxonomic classification, genes that are specific for AD microbial communities might be underrepresented in KEGG which can bias the results. Important functions might not have been detected. This is reflected by the few AD pathways for which KEGG modules are available. Moreover, structural homology of functional distinct enzymes can lead to wrong conclusions. Nevertheless, KEGG proved to be useful for the conducted analysis. Replacing KEGG modules with manually curated pathway definitions based on hits for K numbers allowed the workflow to overcome some of these limitations and better adapt the analysis to the research questions. Moreover, the combination of analysing pathways by enzyme hits, relative abundances of MAGs, and metadata such as measured concentrations in the reactors, proved to be successful in finding potential roles of MAGs. As the current analysis is only based on genomic information, the hypothesised roles should be validated by transcriptomic and proteomic methods in future studies. Moreover, characterising selected MAGs by a more complete pathway reconstruction would be insightful, as some of the MAGs represent novel genomes. The functional analysis could further be expanded to track the carbon flow through the fermenter in a more complete way. The hydrolysis stage was largely omitted in the analysis because the substrates in the systems are expected to mainly consist of already broken-down biomass such as sugar oligomers and monomers. However, the identification of CAZymes (i.e. carbohydrate-active enzymes), or enzymes related to protein or lipid metabolism, could provide valuable additional information about the systems. In general, the analysis could be expanded by other enzyme databases.

A funnel-like specialisation structure, as described in previous studies (Kim et al. 2022; Campanaro et al. 2016), was observed, as a large number of MAGs was assigned to acidogenesis pathways while only few MAGs could be assigned to acetogenesis and methanogenesis. The acidogenesis-related MAGs have been shown to potentially be responsible for the production of acids that have been found at high concentrations in the acidification reactors, i.e. mainly lactic acid and acetic acid, as well as ethanol. The analysis further revealed that all of the three MAGs that were assigned to acetogenesis are exclusively found in methane reactors. One of them (class Anaerolineae) is more abundant in ‘StarchC’. The other two (family Syntrophobacteraceae) are more abundant in ‘PaP’ and, interestingly, at the same time sulphate reducers. Therefore, both are presumably reducing sulphate using

acetate as electron donor, and are thus competing with acetoclastic methanogens. The third sulphate reducer (genus *Fermentibacter*) is the only one which appears to be competing with hydrogenotrophic methanogens. The analysis surprisingly suggests that all MAGs that are members of both detected archaeal families (Methanotrichaceae and Methanobacteriaceae) are capable of hydrogenotrophic and acetoclastic methanogenesis. They might be able to switch between both pathways, depending on their competitors and available substrates. The potential for hydrogen production or utilisation can be investigated by identifying genes that encode hydrogenases. However, the interpretation of such pathways is challenging as they operate in both directions, and their role is thus versatile. Future studies could focus on this topic.

Overall, the results provide promising insights into the microbial community of the selected systems which should be investigated further in the future. It would be insightful to characterise the structure of the granular biomass in the system ‘PaP’ by separately sequencing the distinct layers of the granules. Furthermore, a time series could provide a better understanding of the stability of the systems and, if combined with other parameters, potentially reveal taxa that are associated with stable or unstable process periods. A detailed characterisation of the MAGs in a system could be used to set up a database and, in the future, map ASVs to MAGs. This way, the taxonomic profiles obtained by amplicon sequencing could be connected with the inferred functions attributed to the MAGs. This could provide a basis for biogas plant operators to determine the causes for process disturbances, and to optimise the process.

5 Conclusion

In this thesis, the microbial communities in various anaerobic digestion systems were analysed, before a selection of systems was investigated in more detail. An amplicon-sequencing based analysis revealed patterns in the community composition that appeared to be mainly driven by substrate types. Moreover, two-stage systems provided first insights into the potential roles of taxa. In spite of similarities at the phylum rank, it could be shown that ASVs are mainly specific for certain systems. A multivariate analysis showed the samples are underrepresented with regards to a more global dataset and thus, this thesis provides insights into understudied system types. It was shown that samples from systems treating industrial wastewater are distinct from the majority of the samples in the global dataset. This led to the decision to further investigate the two-stage systems that treat industrial wastewaters using metagenomics. After long-read sequencing and assembly of the reads into contigs, mapping the reads back to the assemblies revealed that the ‘Ref’ samples which represent common AD systems are similar to each other but highly dissimilar from the ‘PaP’ and ‘StarchC’ samples which represent systems that treat industrial wastewaters. This emphasizes that the investigated systems have distinct features from systems that have been investigated previously. 2213 bins could be recovered from the assemblies. After removing low quality bins, 419 MAGs were finally obtained. MAG recovery was most successful for the system ‘PaP’ and thus the focus of further analyses was on this system. Taxonomic profiling showed that there is only little overlap of MAGs between the acidification and the methane reactor of the systems. In a functional analysis, pathways and metabolites could be assigned to MAGs. The results suggest that lactate and acetate are mainly produced by Bifidobacterium fermentation in the acidification reactor of ‘PaP’. Moreover, three sulphate reducing MAGs were identified, two of which are also capable of acetogenesis. Furthermore, it was shown that all archaeal MAGs encode enzymes for both methanogenic pathways, the hydrogenotrophic and acetoclastic route.

Overall, this thesis provides valuable insights into the microbial community and their potential functions in an understudied system type. A number of MAGs responsible for the production of acidogenesis products could be identified and the results reveal interesting potential interactions between acetogenic and sulphate reducing bacteria and methanogens which can serve as basis for future studies. This thesis contributes to a better understanding of AD systems which ultimately might lead to a more efficient biogas production process.

6 Supplementary material

All supplementary files are available at https://github.com/lisa94fc/ADMicrobes_supp. The directory `data` contains an xlsx-file with all supplementary tables as well as two pdf-files. All scripts used in this thesis can be found in the directory `scripts`. Table S1 lists all supplementary files and their description.

Table S1: *Description of supplementary files.*

DIRECTORY: `data`	
Filename	**Description**
SF1_supplementary-tables.xlsx	Xlsx-file with all supplementary tables providing detailed results that were used for creating the figures in this thesis and additional data supplementing the results shown in the thesis.
SF2_Prim-clust-dend.pdf	Primary clustering results obtained during dereplication of bins using dRep. See section 2.4.4.2.
SF3_Sec-clust-dend.pdf	Secondary clustering results obtained during dereplication of bins using dRep. See section 2.4.4.2.
DIRECTORY: `scripts/amplicon-dataset`	
Filename	**Description**
01-amplicon-primer-trimming.Rmd	R-Markdown file for amplicon dataset pipeline, step 1 (primer trimming using cutadapt). To be used with config.yaml. See section 2.3.2.1.
02-amplicon-filter-trim.Rmd	R-Markdown file for amplicon dataset pipeline, step 2 (quality filtering and trimming using DADA2). To be used with config.yaml. See section 2.3.2.2.
03-amplicon-inference.Rmd	R-Markdown file for amplicon dataset pipeline, step 3 (inferring ASVs using DADA2). To be used with config.yaml. See section 2.3.2.3.
04-amplicon-bimeras.Rmd	R-Markdown file for amplicon dataset pipeline, step 4 (removing bimeras and merging paired-end reads using DADA2). To be used with config.yaml. See section 2.3.2.4.
05-amplicon-taxonomy.Rmd	R-Markdown file for amplicon dataset pipeline, step 5 (taxonomic classification using DADA2). To be used with config.yaml. See section 2.3.2.5.
06-amplicon-compare-classifiers.Rmd	R-Markdown file for amplicon dataset pipline, step 6 (evaluating the performance of the databases used for taxonomic classification). See section 2.3.2.5.
07-amplicon-table-formatting.Rmd	R-Markdown file for amplicon dataset pipeline, step 7 (formatting results and merging datasets). To be used with config.yaml. See section 2.3.3.1.
08-amplicon-plots.Rmd	R-Markdown file for amplicon dataset pipeline, step 8 (visualisation of results using ggplot2). To be used with config.yaml. See section 2.3.3.1.
Multivariate-analysis-w-MiDAS.Rmd	R-Markdown file for computing and visualising ordinations of the merged dataset that includes samples from MiDAS and this thesis. See section 2.3.3.3.

Table S1 (continued): *Description of supplementary files.*

DIRECTORY: `scripts/amplicon-dataset`	
Filename	**Description**
08-amplicon-plots.Rmd	R-Markdown file for amplicon dataset pipeline, step 8 (visualisation of results using ggplot2). To be used with config.yaml. See section 2.3.3.1.
Multivariate-analysis-w-MiDAS.Rmd	R-Markdown file for computing and visualising ordinations of the merged dataset that includes samples from MiDAS and this thesis. See section 2.3.3.3.
config.yaml	Config-file for setting parameters for the amplicon dataset pipeline. See section 2.3.2.
fastqc_multiqc.sh	Bash script for obtaining quality profiles of sequencing reads (per sample and per run) using FastQC and MultiQC. See section 2.3.2.2.
merge-thesis-midas.sh	Bash script for merging the MiDAS and thesis dataset. See section 2.3.3.2.
DIRECTORY: `scripts/ont-dataset`	
Filename	**Description**
01_filtlong.sh	Bash script for quality filtering of ONT sequencing results. See section 2.4.4.2.
02_QC_ont.sh	Bash script for quality control of ONT sequencing results. See section 2.4.4.2.
03_metaphlan_batch.sh	Bash script for read-based taxonomic profiling of ONT sequencing results using MetaPhlAn 4. See section 2.4.4.2.
04_singlem.sh	Bash script for read-based taxonomic profiling of ONT sequencing results using SingleM. See section 2.4.4.2.
05_flye.sh	Bash script for assembly of ONT sequencing reads. See section 2.4.4.2.
06_map_reads_back.sh	Bash script for mapping ONT sequencing reads to obtained assemblies using minimap2 and SAMtools. See section 2.4.4.2.
07_coverm_contig.sh	Bash script for determining coverages of contigs using CoverM. See section 2.4.4.2.
08_medaka.sh	Bash script for polishing assemblies using medaka. See section 2.4.4.2.
09_jgi_summarize.sh	Bash script for obtaining coverages of contigs to be used for binning with MetaBAT2. See section 2.4.4.2.
10_metabat.sh	Bash script for binning of ONT sequencing results using MetaBAT2. See section 2.4.4.2.
11_quast.sh	Bash script for quality control of assemblies using Quast. See section 2.4.4.2.
12_checkm2.sh	Bash script for quality control of bins using CheckM2. See section 2.4.4.2.

Table S1 (continued): *Description of supplementary files.*

DIRECTORY: `scripts/ont-dataset`	
Filename	**Description**
13_gtdbtk.sh	Bash script for taxonomic classification of bins using GTDB-Tk. See section 2.4.4.2.
14_coverm_genome.sh	Bash script for estimating relative abundances of MAGs using CoverM. See section 2.4.4.2.
15_anvio_kegg.sh	Bash script for functional annotation of MAGs using anvi'o and KEGG. See section 2.4.4.2.
Pathway_mapping.Rmd	R-Markdown file for assigning MAGs to pathways and metabolites by analysing the functional annotation results. See section 2.4.4.2.
Pathways_plots.Rmd	R-Markdown file for visualising the results of the functional analysis. See section 2.4.4.2.

7 References

Agostini, Sara, Moriconi, Francesco, Zampirolli, Mauro, Padoan, Diego, Treu, Laura, Campanaro, Stefano, and Favaro, Lorenzo. 2023. 'Monitoring the Microbiomes of Agricultural and Food Waste Treating Biogas Plants over a One-Year Period'. *Applied Sciences* 13 (17): 9959. https://doi.org/10.3390/app13179959.

Al Seadi, Teodorita, Rutz, Dominik, Janssen, Rainer, and Drosg, Bernhard. 2013. 'Biomass Resources for Biogas Production'. In *The Biogas Handbook*, edited by Wellinger, Arthur, Murphy, Jerry, and Baxter, David. 19–51. Woodhead Publishing. https://doi.org/10.1533/9780857097415.1.19.

Andrews, Simon (FastQC). 2010. 'FastQC: A Quality Control Tool for High Throughput Sequence Data'. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Apprill, A., McNally, S., Parsons, R., and Weber, L. 2015. 'Minor Revision to V4 Region SSU rRNA 806R Gene Primer Greatly Increases Detection of SAR11 Bacterioplankton'. *Aquatic Microbial Ecology* 75 (2): 129–37. https://doi.org/10.3354/ame01753.

Aramaki, Takuya, Blanc-Mathieu, Romain, Endo, Hisashi, Ohkubo, Koichi, Kanehisa, Minoru, Goto, Susumu, and Ogata, Hiroyuki. 2020. 'KofamKOALA: KEGG Ortholog Assignment Based on Profile HMM and Adaptive Score Threshold'. *Bioinformatics* 36 (7): 2251–52. https://doi.org/10.1093/bioinformatics/btz859.

Aroney, Samuel T. N., Newell, Rhys J. P., Nissen, Jakob N., Camargo, Antonio Pedro, Tyson, Gene W., and Woodcroft, Ben J. 2025. 'CoverM: Read Alignment Statistics for Metagenomics'. *Bioinformatics* 41 (4): btaf147. https://doi.org/10.1093/bioinformatics/btaf147.

Barnett, David J. m, Arts, Ilja C. w, and Penders, John. 2021. 'microViz: An R Package for Microbiome Data Visualization and Statistics'. *Journal of Open Source Software* 6 (63): 3201. https://doi.org/10.21105/joss.03201.

Beraud-Martínez, Liov Karel, Betancourt-Lozano, Miguel, Gómez-Gil, Bruno, Asaff-Torres, Ali, Monroy-Hermosillo, Oscar Armando, and Franco-Nava, Miguel Ángel. 2024. 'Methylotrophic Methanogenesis Induced by Ammonia Nitrogen in an Anaerobic Digestion System'. *Anaerobe* 88: 102877. https://doi.org/10.1016/j.anaerobe.2024.102877.

Blanco-Míguez, Aitor, Beghini, Francesco, Cumbo, Fabio, et al. 2023. 'Extending and Improving Metagenomic Taxonomic Profiling with Uncharacterized Species Using MetaPhlAn 4'. *Nature Biotechnology* 41 (11): 1633–44. https://doi.org/10.1038/s41587-023-01688-w.

Braak, Cajo J. F. ter. 1986. 'Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis'. *Ecology* (Durham) 67 (5): 1167–79.

Brand, Teun van den. 2025. *Ggh4x: Hacks for 'Ggplot2'*. Manual. https://doi.org/10.32614/CRAN.package.ggh4x.

Braun, Rudolf. 2007. 'Anaerobic Digestion: A Multi-Faceted Process for Energy, Environmental Management and Rural Development'. In *Improvement of Crop Plants for Industrial End Uses*, edited by Ranalli, P. 335–416. Springer Netherlands. https://doi.org/10.1007/978-1-4020-5486-0_13.

Breton-Deval, Luz, Salinas-Peralta, Ilse, Alarcón Aguirre, Jaime Santiago, Sulbarán-Rangel, Belkis, and Gurubel Tun, Kelly Joel. 2021. 'Taxonomic Binning Approaches and Functional Characteristics of the Microbial Community during the Anaerobic Digestion of Hydrolyzed Corncob'. *Energies* 14 (1): 1. https://doi.org/10.3390/en14010066.

Buchfink, Benjamin, Reuter, Klaus, and Drost, Hajk-Georg. 2021. 'Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND'. *Nature Methods* 18 (4): 366–68. https://doi.org/10.1038/s41592-021-01101-x.

Cai, Mingwei, Wilkins, David, Chen, Jiapeng, Ng, Siu-Kin, Lu, Hongyuan, Jia, Yangyang, and Lee, Patrick K. H. 2016. 'Metagenomic Reconstruction of Key Anaerobic Digestion Pathways in Municipal Sludge and Industrial Wastewater Biogas-Producing Systems'. *Frontiers in Microbiology* 7: 778. https://doi.org/10.3389/fmicb.2016.00778.

Callahan, Benjamin J., McMurdie, Paul J., Rosen, Michael J., Han, Andrew W., Johnson, Amy Jo A., and Holmes, Susan P. 2016. 'DADA2: High-Resolution Sample Inference from Illumina Amplicon Data'. *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

Callahan, Benjamin J., Sankaran, Kris, Fukuyama, Julia A., McMurdie, Paul J., and Holmes, Susan P. 2016. 'Bioconductor Workflow for Microbiome Data Analysis: From Raw Reads to Community Analyses'. Preprint, F1000Research. https://doi.org/10.12688/f1000research.8986.2.

Camargo, Antonio Pedro, Roux, Simon, Schulz, Frederik, Babinski, Michal, Xu, Yan, Hu, Bin, Chain, Patrick S. G., Nayfach, Stephen, and Kyrpides, Nikos C. 2024. 'Identification of Mobile Genetic Elements with geNomad'. *Nature Biotechnology* 42 (8): 1303–12. https://doi.org/10.1038/s41587-023-01953-y.

Camargo, Franciele Pereira, Sakamoto, Isabel Kimiko, Delforno, Tiago Palladino, Midoux, Cédric, Duarte, Iolanda Cristina Silveira, Silva, Edson Luiz, Bize, Ariane, and Varesche, Maria Bernadete Amâncio. 2023. 'Microbial and Functional Characterization of Granulated Sludge from Full-Scale UASB Thermophilic Reactor Applied to Sugarcane Vinasse Treatment'. *Environmental Technology* 44 (21): 3141–60. https://doi.org/10.1080/09593330.2022.2052361.

Campanaro, Stefano, Treu, Laura, Kougias, Panagiotis G., De Francisci, Davide, Valle, Giorgio, and Angelidaki, Irini. 2016. 'Metagenomic Analysis and Functional Characterization of the Biogas Microbiome Using High Throughput Shotgun Sequencing and a Novel Binning Strategy'. *Biotechnology for Biofuels* 9 (1): 26. https://doi.org/10.1186/s13068-016-0441-1.

Campanaro, Stefano, Treu, Laura, Kougias, Panagiotis G., Luo, Gang, and Angelidaki, Irini. 2018. 'Metagenomic Binning Reveals the Functional Roles of Core Abundant Microorganisms in Twelve Full-Scale Biogas Plants'. *Water Research* 140: 123–34. https://doi.org/10.1016/j.watres.2018.04.043.

Campanaro, Stefano, Treu, Laura, Rodriguez-R, Luis M., Kovalovszki, Adam, Ziels, Ryan M., Maus, Irena, Zhu, Xinyu, Kougias, Panagiotis G., Basile, Arianna, Luo, Gang, Schlüter, Andreas, Konstantinidis, Konstantinos T., and Angelidaki, Irini. 2020. 'New Insights from the Biogas Microbiome by Comprehensive Genome-Resolved Metagenomics of Nearly 1600 Species Originating from Multiple Anaerobic Digesters'. *Biotechnology for Biofuels* 13 (1): 25. https://doi.org/10.1186/s13068-020-01679-y.

Cao, Jiachang, Zhang, Chen, Li, Xiang, Wang, Xueye, Dai, Xiaohu, and Xu, Ying. 2025. 'Microbial Community and Metabolic Pathways in Anaerobic Digestion of Organic Solid Wastes: Progress, Challenges and Prospects'. *Fermentation* 11 (8): 457. https://doi.org/10.3390/fermentation11080457.

Carrez, MC. 1908. 'Le Ferrocyanure de Potassium et l'acétate de Zinc Comme Agents de Défécation Des Urines'. *Annales de Chimie Analytique* 13: 97–101.

Chaumeil, Pierre-Alain, Mussig, Aaron J., Hugenholtz, Philip, and Parks, Donovan H. 2020. 'GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database'. *Bioinformatics* 36 (6): 1925–27. https://doi.org/10.1093/bioinformatics/btz848.

Chaumeil, Pierre-Alain, Mussig, Aaron J., Hugenholtz, Philip, and Parks, Donovan H. 2022. 'GTDB-Tk v2: Memory Friendly Classification with the Genome Taxonomy Database'. *Bioinformatics* 38 (23): 5315–16. https://doi.org/10.1093/bioinformatics/btac672.

Chklovski, Alex, Parks, Donovan H., Woodcroft, Ben J., and Tyson, Gene W. 2023. 'CheckM2: A Rapid, Scalable and Accurate Tool for Assessing Microbial Genome Quality Using Machine Learning'. *Nature Methods* 20 (8): 1203–12. https://doi.org/10.1038/s41592-023-01940-w.

Da Costa Gomez, Claudius. 2013. 'Biogas as an Energy Option: An Overview'. In *The Biogas Handbook*, edited by Wellinger, Arthur, Murphy, Jerry, and Baxter, David. 1–16. Woodhead Publishing. https://doi.org/10.1533/9780857097415.1.

Danecek, Petr, Bonfield, James K., Liddle, Jennifer, Marshall, John, Ohan, Valeriu, Pollard, Martin O., Whitwham, Andrew, Keane, Thomas, McCarthy, Shane A., Davies, Robert M., and Li, Heng. 2021. 'Twelve Years of SAMtools and BCFtools'. *GigaScience* 10 (2): giab008. https://doi.org/10.1093/gigascience/giab008.

De Coster, Wouter, and Rademakers, Rosa. 2023. 'NanoPack2: Population-Scale Evaluation of Long-Read Sequencing Data'. *Bioinformatics* 39 (5): btad311. https://doi.org/10.1093/bioinformatics/btad311.

De Lemos Chernicharo, C. A. 2007. *Anaerobic Reactors*. Vol. 4. Biological Wastewater Treatment. https://iwaponline.com/ebooks/book/79/.

Deamer, David, Akeson, Mark, and Branton, Daniel. 2016. 'Three Decades of Nanopore Sequencing'. *Nature Biotechnology* 34 (5): 518–24. https://doi.org/10.1038/nbt.3423.

Doloman, Anna, Besteman, Maaike S., Sanders, Mark G., and Sousa, Diana Z. 2024. 'Methanogenic Partner Influences Cell Aggregation and Signalling of Syntrophobacterium Fumaroxidans'. *Applied Microbiology and Biotechnology* 108 (1): 127. https://doi.org/10.1007/s00253-023-12955-w.

Domínguez-Espinosa, María Emperatriz, Cruz-Salomón, Abumalé, Ramírez de León, José Alberto, Hernández-Méndez, Jesús Mauricio Ernesto, and Santiago-Martínez, Michel Geovanni. 2023. 'Syntrophy between Bacteria and Archaea Enhances Methane Production in an EGSB Bioreactor Fed by Cheese Whey Wastewater'. *Frontiers in Sustainable Food Systems* 7. https://doi.org/10.3389/fsufs.2023.1244691.

Drosg, Bernhard. 2013. 'Process Monitoring in Biogas Plants'. IEA Bioenergy Task 37. https://task37.ieabioenergy.com/technical-reports/process-monitoring-in-biogas-plants/.

Drosg, Bernhard, Neubauer, Matthias, Marzynski, Marceli, and Meixner, Katharina. 2021. 'Valorisation of Starch Wastewater by Anaerobic Fermentation'. *Applied Sciences* 11 (21): 10482. https://doi.org/10.3390/app112110482.

Dueholm, Morten K. D., Andersen, Kasper S., Korntved, Anne-Kirstine C., et al. 2024. 'MiDAS 5: Global Diversity of Bacteria and Archaea in Anaerobic Digesters'. *Nature Communications* 15 (1): 5361. https://doi.org/10.1038/s41467-024-49641-y.

Eckhoff, Steven R., and Watson, Stanley A. 2009. 'Corn and Sorghum Starches: Production'. In *Starch*, 3rd edn, edited by BeMiller, James and Whistler, Roy. 373–439. Academic Press. https://doi.org/10.1016/B978-0-12-746275-2.00009-4.

Eddy, Sean R. 2011. 'Accelerated Profile HMM Searches'. *PLOS Computational Biology* 7 (10): e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Edgar, Robert C. 2010. 'Search and Clustering Orders of Magnitude Faster than BLAST'. *Bioinformatics* 26 (19): 2460–61. https://doi.org/10.1093/bioinformatics/btq461.

Edgar, Robert C. 2017. 'Updating the 97% Identity Threshold for 16S Ribosomal RNA OTUs'. Preprint, bioRxiv. https://doi.org/10.1101/192211.

Eren, A. Murat, Kiefl, Evan, Shaiber, Alon, et al. 2020. 'Community-Led, Integrated, Reproducible Multi-Omics with Anvi'o'. *Nature Microbiology* 6 (1): 3–6. https://doi.org/10.1038/s41564-020-00834-3.

Ewels, Philip, Magnusson, Måns, Lundin, Sverker, and Käller, Max. 2016. 'MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report'. *Bioinformatics* 32 (19): 3047–48. https://doi.org/10.1093/bioinformatics/btw354.

Feng, Siran, Ngo, Huu Hao, Guo, Wenshan, Chang, Soon Woong, Nguyen, Dinh Duc, Liu, Yi, Zhang, Shicheng, Phong Vo, Hoang Nhat, Bui, Xuan Thanh, and Ngoc Hoang, Bich. 2022. 'Volatile Fatty Acids Production from Waste Streams by Anaerobic Digestion: A Critical Review of the Roles and Application of Enzymes'. *Bioresource Technology* 359: 127420. https://doi.org/10.1016/j.biortech.2022.127420.

Granger, Marthe, Marnane, Ian, and Martin-Montalvo Alvarez, Daniel. 2019. *Industrial Waste Water Treatment – Pressures on Europe's Environment*. No. 23/2018. European Environment Agency. https://www.eea.europa.eu/en/analysis/publications/industrial-waste-water-treatment-pressures.

Gurevich, Alexey, Saveliev, Vladislav, Vyahhi, Nikolay, and Tesler, Glenn. 2013. 'QUAST: Quality Assessment Tool for Genome Assemblies'. *Bioinformatics* 29 (8): 1072–75. https://doi.org/10.1093/bioinformatics/btt086.

Hackmann, Timothy J. 2024. 'The Vast Landscape of Carbohydrate Fermentation in Prokaryotes'. *FEMS Microbiology Reviews* 48 (4): fuae016. https://doi.org/10.1093/femsre/fuae016.

Hao, Liping, Michaelsen, Thomas Yssing, Singleton, Caitlin Margaret, Dottorini, Giulia, Kirkegaard, Rasmus Hansen, Albertsen, Mads, Nielsen, Per Halkjær, and Dueholm, Morten Simonsen. 2020. 'Novel Syntrophic Bacteria in Full-Scale Anaerobic Digesters Revealed by Genome-Centric Metatranscriptomics'. *The ISME Journal* 14 (4): 906–18. https://doi.org/10.1038/s41396-019-0571-0.

Harirchi, Sharareh, Wainaina, Steven, Sar, Taner, Nojoumi, Seyed Ali, Parchami, Milad, Parchami, Mohsen, Varjani, Sunita, Khanal, Samir Kumar, Wong, Jonathan, Awasthi, Mukesh Kumar, and Taherzadeh, Mohammad J. 2022. 'Microbiological Insights into

Anaerobic Digestion for Biogas, Hydrogen or Volatile Fatty Acids (VFAs): A Review'. *Bioengineered* 13 (3): 6521–57. https://doi.org/10.1080/21655979.2022.2035986.

Harmsen, Hermie J. M., Van Kuijk, Bernardina L. M., Plugge, Caroline M., AKKERMANS, Antoon D. L., De Vos, Willem M., and Stams, Alfons J. M. 1998. 'Syntrophobacter Fumaroxidans Sp. Nov., a Syntrophic Propionate-Degrading Sulfate-Reducing Bacterium'. *International Journal of Systematic and Evolutionary Microbiology* 48 (4): 1383–87. https://doi.org/10.1099/00207713-48-4-1383.

Heider, Johann. 2014. 'Mikrobielle Gärungen (Mirobial fermentations)'. In *Allgemeine Mikrobiologie (General Microbiology)*, 9th edn, edited by Fuchs, Georg. 410–441. Georg Thieme Verlag.

Holl, Elena, Steinbrenner, Jörg, Merkle, Wolfgang, Krümpel, Johannes, Lansing, Stephanie, Baier, Urs, Oechsner, Hans, and Lemmer, Andreas. 2022. 'Two-Stage Anaerobic Digestion: State of Technology and Perspective Roles in Future Energy Systems'. *Bioresource Technology* 360: 127633. https://doi.org/10.1016/j.biortech.2022.127633.

Hu, Huifeng, Karwautz, Clemens, Duszka, Kalina, Karner, Thomas, Wagner, Isabella C., Grander, Christoph, Grander, Wilhelm, Steinwidder, Laura, Boito, Lucilla, Velde, Viktor Van de, Bauters, Marijn, Boeckx, Pascal, Seki, David, Glasl, Bettina, Thiele, Stefan, Schmidt, Hannes, Séneca, Joana, Wagner, Michael, and Pjevac, Petra. 2025. 'Revised 16S rRNA V4 Hypervariable Region Targeting Primers Enhance Detection of Patescibacteria and Other Lineages across Diverse Environments'. Preprint, bioRxiv. https://doi.org/10.1101/2025.11.26.690684.

Hyatt, Doug, Chen, Gwo-Liang, LoCascio, Philip F., Land, Miriam L., Larimer, Frank W., and Hauser, Loren J. 2010. 'Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification'. *BMC Bioinformatics* 11 (1): 119. https://doi.org/10.1186/1471-2105-11-119.

Illumina. 2017. *Bcl2fastq*. V. 2.20.0.422. Illumina, released. https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html.

'Introduction_to_how_ONT's_medaka_works'. n.d. Accessed 15 May 2026. https://epi2me.nanoporetech.com/notebooks/Introduction_to_how_ONT%27s_medaka_works.html.

Jonge, Nadieh de, Davidsson, Åsa, Cour Jansen, Jes la, and Nielsen, Jeppe Lund. 2020. 'Characterisation of Microbial Communities for Improved Management of Anaerobic Digestion of Food Waste'. *Waste Management* 117: 124–35. https://doi.org/10.1016/j.wasman.2020.07.047.

Jun, Hangbae, Kadam, Rahul, Jo, Sangyeol, and Park, Jungyu. 2025. 'Unravelling Metabolic Pathways and Evaluating Process Performances in Anaerobic Digestion of Livestock Manures'. *Water* 17 (24): 3464. https://doi.org/10.3390/w17243464.

Juneja, Ankita, Sharma, Navneet, Cusick, Roland, and Singh, Vijay. 2019. 'Techno-Economic Feasibility of Phosphorus Recovery as a Coproduct from Corn Wet Milling Plants'. *Cereal Chemistry* 96 (2): 380–90. https://doi.org/10.1002/cche.10139.

Jünemann, Sebastian, Kleinbölting, Nils, Jaenicke, Sebastian, Henke, Christian, Hassa, Julia, Nelkner, Johanna, Stolze, Yvonne, Albaum, Stefan P., Schlüter, Andreas, Goesmann, Alexander, Sczyrba, Alexander, and Stoye, Jens. 2017. 'Bioinformatics for NGS-Based Metagenomics and the Application to Biogas Research'. *Journal of Biotechnology*, Bioinformatics Solutions for Big Data Analysis in Life Sciences presented by the

German Network for Bioinformatics Infrastructure, vol. 261: 10–23. https://doi.org/10.1016/j.jbiotec.2017.08.012.

Kanehisa, Minoru, Furumichi, Miho, Tanabe, Mao, Sato, Yoko, and Morishima, Kanae. 2017. 'KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs'. *Nucleic Acids Research* 45 (D1): D353–61. https://doi.org/10.1093/nar/gkw1092.

Kanehisa, Minoru, and Goto, Susumu. 2000. 'KEGG: Kyoto Encyclopedia of Genes and Genomes'. *Nucleic Acids Research* 28 (1): 27–30. https://doi.org/10.1093/nar/28.1.27.

Kanehisa, Minoru, Goto, Susumu, Sato, Yoko, Kawashima, Masayuki, Furumichi, Miho, and Tanabe, Mao. 2014. 'Data, Information, Knowledge and Principle: Back to Metabolism in KEGG'. *Nucleic Acids Research* 42 (D1): D199–205. https://doi.org/10.1093/nar/gkt1076.

Kang, Dongwan D., Li, Feng, Kirton, Edward, Thomas, Ashleigh, Egan, Rob, An, Hong, and Wang, Zhong. 2019. 'MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies'. *PeerJ* 7: e7359. https://doi.org/10.7717/peerj.7359.

Kim, Na-Kyung, Lee, Sang-Hoon, Kim, Yonghoon, and Park, Hee-Deung. 2022. 'Current Understanding and Perspectives in Anaerobic Digestion Based on Genome-Resolved Metagenomic Approaches'. *Bioresource Technology* 344: 126350. https://doi.org/10.1016/j.biortech.2021.126350.

Klindworth, Anna, Pruesse, Elmar, Schweer, Timmy, Peplies, Jörg, Quast, Christian, Horn, Matthias, and Glöckner, Frank Oliver. 2013. 'Evaluation of General 16S Ribosomal RNA Gene PCR Primers for Classical and Next-Generation Sequencing-Based Diversity Studies'. *Nucleic Acids Research* 41 (1): e1. https://doi.org/10.1093/nar/gks808.

Kolmogorov, Mikhail, Bickhart, Derek M., Behsaz, Bahar, Gurevich, Alexey, Rayko, Mikhail, Shin, Sung Bong, Kuhn, Kristen, Yuan, Jeffrey, Polevikov, Evgeny, Smith, Timothy P. L., and Pevzner, Pavel A. 2020. 'metaFlye: Scalable Long-Read Metagenome Assembly Using Repeat Graphs'. *Nature Methods* 17 (11): 1103–10. https://doi.org/10.1038/s41592-020-00971-x.

Kolmogorov, Mikhail, Yuan, Jeffrey, Lin, Yu, and Pevzner, Pavel A. 2019. 'Assembly of Long, Error-Prone Reads Using Repeat Graphs'. *Nature Biotechnology* 37 (5): 540–46. https://doi.org/10.1038/s41587-019-0072-8.

Kroiss, Helmut, and Svardal, Karl. 1988. 'Aufwärtsdurchströmter Schlammbettreaktor mit Drehverteiler (EKJ-Reaktor)'. In *Anaerobe Abwasserreinigung: Grundlagen und großtechnische Erfahrung*, edited by Helmut Kroiss. 59–78. Technische Universität Wien – Institut für Wassergüte und Landschaftswasserbau.

Larralde, Martin. 2022. 'Pyrodigal: Python Bindings and Interface to Prodigal, an Efficient Method for Gene Prediction in Prokaryotes'. *Journal of Open Source Software* 7 (72): 4296. https://doi.org/10.21105/joss.04296.

Latorre-Pérez, Adriel, Villalba-Bermell, Pascual, Pascual, Javier, and Vilanova, Cristina. 2020. 'Assembly Methods for Nanopore-Based Metagenomic Sequencing: A Comparative Study'. *Scientific Reports* 10 (1): 13588. https://doi.org/10.1038/s41598-020-70491-3.

Legendre, Pierre, and Legendre, Louis. 2012. 'Canonical Analysis'. In *Numerical Ecology*, edited by Legendre, Pierre, and Legendre, Louis. 625–710. Elsevier. https://doi.org/10.1016/B978-0-444-53868-0.50011-3.

Lettinga, G., Velsen, A. F. M. van, Hobma, S. W., Zeeuw, W. de, and Klapwijk, A. 1980. 'Use of the Upflow Sludge Blanket (USB) Reactor Concept for Biological Wastewater Treatment, Especially for Anaerobic Treatment'. *Biotechnology and Bioengineering* 22 (4): 699–734. https://doi.org/10.1002/bit.260220402.

Li, Dan, Kim, Minsoo, Kim, Hyunjin, Choi, Okkyoung, Sang, Byoung-In, Chiang, Pen Chi, and Kim, Hyunook. 2018. 'Evaluation of Relationship between Biogas Production and Microbial Communities in Anaerobic Co-Digestion'. *Korean Journal of Chemical Engineering* 35 (1): 179–86. https://doi.org/10.1007/s11814-017-0246-3.

Li, Heng. 2018. 'Minimap2: Pairwise Alignment for Nucleotide Sequences'. *Bioinformatics* 34 (18): 3094–4100. https://doi.org/10.1093/bioinformatics/bty191.

Liang, Xiaoli, Xu, Yanpeng, Yin, Liang, Wang, Ruiming, Li, Piwu, Wang, Junqing, and Liu, Kaiquan. 2023. 'Sustainable Utilization of Pulp and Paper Wastewater'. *Water* 15 (23): 4135. https://doi.org/10.3390/w15234135.

Lim, Jun Wei, Park, Tansol, Tong, Yen Wah, and Yu, Zhongtang. 2020. 'The Microbiome Driving Anaerobic Digestion and Microbial Analysis'. *Advances in Bioenergy* 5: 1–61. https://doi.org/10.1016/bs.aibe.2020.04.001.

Lindholm-Lehto, Petra C., Knuutinen, Juha S., Ahkola, Heidi S. J., and Herve, Sirpa H. 2015. 'Refractory Organic Pollutants and Toxicity in Pulp and Paper Mill Wastewaters'. *Environmental Science and Pollution Research* 22 (9): 6473–99. https://doi.org/10.1007/s11356-015-4163-x.

Lupa, Boguslaw, Hendrickson, Erik L., Leigh, John A., and Whitman, William B. 2008. 'Formate-Dependent H2 Production by the Mesophilic Methanogen Methanococcus Maripaludis'. *Applied and Environmental Microbiology* 74 (21): 6584–90. https://doi.org/10.1128/AEM.01455-08.

Lyu, Zhe, Shao, Nana, Akinyemi, Taiwo, and Whitman, William B. 2018. 'Methanogenesis'. *Current Biology* 28 (13): R727–32. https://doi.org/10.1016/j.cub.2018.05.021.

MacKenzie, Morgan, and Argyropoulos, Christos. 2023. 'An Introduction to Nanopore Sequencing: Past, Present, and Future Considerations'. *Micromachines* 14 (2): 459. https://doi.org/10.3390/mi14020459.

Madden, Pádhraig, Al-Raei, Abdul M., Enright, Anne M., Chinalia, Fabio A., Beer, Dirk de, O'Flaherty, Vincent, and Collins, Gavin. 2014. 'Effect of Sulfate on Low-Temperature Anaerobic Digestion'. *Frontiers in Microbiology* 5. https://doi.org/10.3389/fmicb.2014.00376.

Mao, Chunlan, Feng, Yongzhong, Wang, Xiaojiao, and Ren, Guangxin. 2015. 'Review on Research Achievements of Biogas from Anaerobic Digestion'. *Renewable and Sustainable Energy Reviews* 45: 540–55. https://doi.org/10.1016/j.rser.2015.02.032.

Martin, Marcel. 2011. 'Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads'. *EMBnet.Journal* 17 (1): 10–12. https://doi.org/10.14806/ej.17.1.200.

Matsen, Frederick A., Kodner, Robin B., and Armbrust, E. Virginia. 2010. 'Pplacer: Linear Time Maximum-Likelihood and Bayesian Phylogenetic Placement of Sequences onto a Fixed Reference Tree'. *BMC Bioinformatics* 11 (1): 538. https://doi.org/10.1186/1471-2105-11-538.

McIlroy, Simon Jon, Saunders, Aaron Marc, Albertsen, Mads, Nierychlo, Marta, McIlroy, Bianca, Hansen, Aviaja Anna, Karst, Søren Michael, Nielsen, Jeppe Lund, and Nielsen, Per Halkjær. 2015. 'MiDAS: The Field Guide to the Microbes of Activated Sludge'. *Database* 2015. https://doi.org/10.1093/database/bav062.

Meyer, Torsten, and Edwards, Elizabeth A. 2014. 'Anaerobic Digestion of Pulp and Paper Mill Wastewater and Sludge'. *Water Research* 65: 321–49. https://doi.org/10.1016/j.watres.2014.07.022.

Möbius, Christian H. 2010. *Abwasser der Papier- und Zellstoffindustrie (Wastewater of pulp and paper industry)*. 4th edn. CM Consult.

MP Biomedicals. 2021. 'FastDNA SPIN Kit for Soil: Detailed Protocol'. MP Biomedicals. https://www.mpbio.com/media/document/file/protocol/dest/f/a/s/t/d/FastDNA_SPIN_Kit_for_Soil_Detailed_Protocol_Insert_2021_WEB.pdf.

Neogi, Shubhaneel. 2020. *Studies on the Biological Treatment of Wastewater From Starch Industry for Pollution Control*. 1. Auflage. GRIN Verlag. 2523045. https://research.ebsco.com/linkprocessor/plink?id=85d7b43d-c0d0-368b-b85d-32f0d9d8af4c.

Niya, Btissam, Yaakoubi, Kaoutar, Beraich, Fatima Zahra, Arouch, Moha, and Kadmiri, Issam Meftah. 2024. 'Current Status and Future Developments of Assessing Microbiome Composition and Dynamics in Anaerobic Digestion Systems Using Metagenomic Approaches'. *Heliyon* 10 (6). https://doi.org/10.1016/j.heliyon.2024.e28221.

Ogbu, Cosmas Chikezie, and Okey, Stephen Nnaemeka. 2023. 'Agro-Industrial Waste Management: The Circular and Bioeconomic Perspective'. In *Agricultural Waste - New Insights*, edited by Ahmad, Fiaz, Sultan, Muhammad. 1–37. IntechOpen. https://doi.org/10.5772/intechopen.109181.

Oksanen, Jari, Simpson, Gavin L., Blanchet, F. Guillaume, et al. 2025. *Vegan: Community Ecology Package*. Manual. https://doi.org/10.32614/CRAN.package.vegan.

Olm, Matthew R., Brown, Christopher T., Brooks, Brandon, and Banfield, Jillian F. 2017. 'dRep: A Tool for Fast and Accurate Genomic Comparisons That Enables Improved Genome Recovery from Metagenomes through de-Replication'. *The ISME Journal* 11 (12): 2864–68. https://doi.org/10.1038/ismej.2017.126.

Oren, Aharon, Arahal, David R., Göker, Markus, Moore, Edward R. B., Rossello-Mora, Ramon, and Sutcliffe, Iain C. 2023. 'International Code of Nomenclature of Prokaryotes. Prokaryotic Code (2022 Revision)'. *International Journal of Systematic and Evolutionary Microbiology* 73 (5a): 005585. https://doi.org/10.1099/ijsem.0.005585.

Oxford Nanopore Technologies. 2024. *Secondary Analysis Update*. 11:25. https://www.youtube.com/watch?v=IB6DmU40NIU.

Oxford Nanopore Technologies. 2025a. *Dorado*. V. 7.11.2. Released. https://nanoporetech.com/es/software/other/dorado-basecall-server/history.

Oxford Nanopore Technologies. 2025b. *Medaka*. Python. V. 2.1.1. Released. https://github.com/nanoporetech/medaka.

Oxford Nanopore Technologies. 2025c. *MinKNOW*. V. 25.06.16. Released. https://nanoporetech.com/software/devices/p2i/software/linux/history.

Paliy, O., and Shankar, V. 2016. ‘Application of Multivariate Statistical Techniques in Microbial Ecology’. *Molecular Ecology* 25 (5): 1032–57. https://doi.org/10.1111/mec.13536.

Parks, Donovan H., Chuvochina, Maria, Rinke, Christian, Mussig, Aaron J., Chaumeil, Pierre-Alain, and Hugenholtz, Philip. 2022. ‘GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy’. *Nucleic Acids Research* 50 (D1): D785–94. https://doi.org/10.1093/nar/gkab776.

Pjevac, Petra, Hausmann, Bela, Schwarz, Jasmin, Kohl, Gudrun, Herbold, Craig W., Loy, Alexander, and Berry, David. 2021. ‘An Economical and Flexible Dual Barcoding, Two-Step PCR Approach for Highly Multiplexed Amplicon Sequencing’. *Frontiers in Microbiology* 12. https://doi.org/10.3389/fmicb.2021.669776.

Pollock, Jolinda, Glendinning, Laura, Wisedchanwet, Trong, and Watson, Mick. 2018. ‘The Madness of Microbiome: Attempting To Find Consensus “Best Practice” for 16S Microbiome Studies’. *Applied and Environmental Microbiology* 84 (7): e02627-17. https://doi.org/10.1128/AEM.02627-17.

Posit team. 2025. *RStudio: Integrated Development Environment for R*. Manual. Posit Software, PBC. http://www.posit.co/.

Quast, Christian, Pruesse, Elmar, Yilmaz, Pelin, Gerken, Jan, Schweer, Timmy, Yarza, Pablo, Peplies, Jörg, and Glöckner, Frank Oliver. 2013. ‘The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools’. *Nucleic Acids Research* 41 (Database issue): D590-596. https://doi.org/10.1093/nar/gks1219.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, released. https://www.R-project.org/.

Ramette, Alban. 2007. ‘Multivariate Analyses in Microbial Ecology’. *FEMS Microbiology Ecology* 62 (2): 142–60. https://doi.org/10.1111/j.1574-6941.2007.00375.x.

Ren, Ze, Wang, Mei, Yu, Jinlei, Zhang, Lixiang, Lin, Zhenmei, Li, Xia, and Zhang, Yunlin. 2025. ‘Unearthing Vertical Stratified Archaeal Community and Associated Methane Metabolism in Thermokarst Sediments’. *Environmental Microbiology* 27 (5): e70110. https://doi.org/10.1111/1462-2920.70110.

Shen, Wei, Sipos, Botond, and Zhao, Liuyang. 2024. ‘SeqKit2: A Swiss Army Knife for Sequence and Alignment Processing’. *iMeta* 3 (3): e191. https://doi.org/10.1002/imt2.191.

Show, Kuan-Yeow, Yan, Yuegen, Yao, Haiyong, Guo, Hui, Li, Ting, Show, De-Yang, Chang, Jo-Shu, and Lee, Duu-Jong. 2020. ‘Anaerobic Granulation: A Review of Granulation Hypotheses, Bioreactor Designs and Emerging Green Applications’. *Bioresource Technology* 300: 122751. https://doi.org/10.1016/j.biortech.2020.122751.

Shubhaneel, Neogi, Apurba, Dey, and Kumar, Chaterjee Pradip. 2018. ‘Corn Starch Industry Wastewater Pollution and Treatment Processes- A Review’. *Journal of Biodiversity and Environmental Sciences* 12 (3): 283–93.

Singh, Bikram Jit, Chakraborty, Ayon, and Sehgal, Rippin. 2023. ‘A Systematic Review of Industrial Wastewater Management: Evaluating Challenges and Enablers’. *Journal of Environmental Management* 348: 119230. https://doi.org/10.1016/j.jenvman.2023.119230.

Speece, Richard E. 1996. *Anaerobic Biotechnology for Industrial Wastewaters*. Archae Press.

Speth, Daan R., Pullen, Nick, Aroney, Samuel T. N., Coltman, Benjamin L., Osvatic, Jay, Woodcroft, Ben J., Rattei, Thomas, and Wagner, Michael. 2025. 'GlobDB: A Comprehensive Species-Dereplicated Microbial Genome Resource'. *Bioinformatics Advances* 5 (1): vbaf280. https://doi.org/10.1093/bioadv/vbaf280.

Stefanie, J. W. H. Oude Elferink, Visser, André, Pol, Look W. Hulshoff, and Stams, Alfons J. M. 1994. 'Sulfate Reduction in Methanogenic Bioreactors'. *FEMS Microbiology Reviews* 15 (2–3): 119–36. https://doi.org/10.1111/j.1574-6976.1994.tb00130.x.

Stoyancheva, Galina, Kabaivanova, Lyudmila, Hubenov, Venelin, and Chorukova, Elena. 2023. 'Metagenomic Analysis of Bacterial, Archaeal and Fungal Diversity in Two-Stage Anaerobic Biodegradation for Production of Hydrogen and Methane from Corn Steep Liquor'. *Microorganisms* 11 (5): 5. https://doi.org/10.3390/microorganisms11051263.

Swift, Dionne, Cresswell, Kellen, Johnson, Robert, Stilianoudakis, Spiro, and Wei, Xingtao. 2023. 'A Review of Normalization and Differential Abundance Methods for Microbiome Counts Data'. *WIREs Computational Statistics* 15 (1): e1586. https://doi.org/10.1002/wics.1586.

Tauseef, S. M., Abbasi, Tasneem, and Abbasi, S. A. 2013. 'Energy Recovery from Wastewaters with High-Rate Anaerobic Digesters'. *Renewable and Sustainable Energy Reviews* 19: 704–41. https://doi.org/10.1016/j.rser.2012.11.056.

Thauer, Rudolf K., Kaster, Anne-Kristin, Seedorf, Henning, Buckel, Wolfgang, and Hedderich, Reiner. 2008. 'Methanogenic Archaea: Ecologically Relevant Differences in Energy Conservation'. *Nature Reviews Microbiology* 6 (8): 579–91. https://doi.org/10.1038/nrmicro1931.

Theuerl, Susanne, Klang, Johanna, and Prochnow, Annette. 2019. 'Process Disturbances in Agricultural Biogas Production—Causes, Mechanisms and Effects on the Biogas Microbiome: A Review'. *Energies* 12 (3): 365. https://doi.org/10.3390/en12030365.

Theuerl, Susanne, Kohrs, Fabian, Benndorf, Dirk, Maus, Irena, Wibberg, Daniel, Schlüter, Andreas, Kausmann, Robert, Heiermann, Monika, Rapp, Erdmann, Reichl, Udo, Pühler, Alfred, and Klocke, Michael. 2015. 'Community Shifts in a Well-Operating Agricultural Biogas Plant: How Process Variations Are Handled by the Microbiome'. *Applied Microbiology and Biotechnology* 99 (18): 7791–803. https://doi.org/10.1007/s00253-015-6627-9.

Trego, Anna Christine, Mills, Simon, and Collins, Gavin. 2021. 'Granular Biofilms: Function, Application, and New Trends as Model Microbial Communities'. *Critical Reviews in Environmental Science and Technology* 51 (15): 1702–25. https://doi.org/10.1080/10643389.2020.1769433.

Treu, Laura, Kougias, Panagiotis G., Campanaro, Stefano, Bassani, Ilaria, and Angelidaki, Irini. 2016. 'Deeper Insight into the Structure of the Anaerobic Digestion Microbial Community; the Biogas Microbiome Database Is Expanded with 157 New Genomes'. *Bioresource Technology* 216: 260–66. https://doi.org/10.1016/j.biortech.2016.05.081.

Trigodet, Florian, Sachdeva, Rohan, Banfield, Jillian F., and Eren, A. Murat. 2026. 'Troubleshooting Common Errors in Assemblies of Long-Read Metagenomes'. *Nature Biotechnology*, 1–10. https://doi.org/10.1038/s41587-025-02971-8.

UN-Water. 2024. *Progress on Wastewater Treatment*. https://www.unwater.org/publications/progress-wastewater-treatment-2024-update.

Venkiteshwaran, Kaushik, Bocher, Benjamin, Maki, James, and Zitomer, Daniel. 2015. 'Relating Anaerobic Digestion Microbial Community and Process Function : Supplementary Issue: Water Microbiology'. *Microbiology Insights* 8s2: MBI.S33593. https://doi.org/10.4137/MBI.S33593.

Veseli, Iva, Chen, Yiqun T., Schechter, Matthew S., Vanni, Chiara, Fogarty, Emily C., Watson, Andrea R., Jabri, Bana, Blekhman, Ran, Willis, Amy D., Yu, Michael K., Fernàndez-Guerra, Antonio, Füssel, Jessika, and Eren, A. Murat. 2025. 'Microbes with Higher Metabolic Independence Are Enriched in Human Gut Microbiomes under Stress'. *eLife* 12: RP89862. https://doi.org/10.7554/eLife.89862.

Wall, David M., and O'Shea, Richard. 2025. 'Biogas Systems in Industry: An Analysis of Sectoral Usage, Sustainability, Logistics and Technology Development'. IEA Bioenergy Task 37. https://task37.ieabioenergy.com/technical-reports/biogas-systems-in-industry/.

Wang, Qiong, Garrity, George M., Tiedje, James M., and Cole, James R. 2007. 'Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy'. *Applied and Environmental Microbiology* 73 (16): 5261–67. https://doi.org/10.1128/AEM.00062-07.

Weinrich, Sören, and Nelles, Michael. 2021. *Basics of Anaerobic Digestion: Biochemical Conversion and Process Modelling*. DBFZ-Report 40. DBFZ Deutsches Biomasseforschungszentrum.

Westerholm, Maria, Calusinska, Magdalena, and Dolfing, Jan. 2021. 'Syntrophic Propionate-Oxidizing Bacteria in Methanogenic Systems'. *FEMS Microbiology Reviews* 46 (2): fuab057. https://doi.org/10.1093/femsre/fuab057.

Wick, Ryan. 2025. *Filtlong*. V. 0.3.1. Released. https://github.com/rrwick/Filtlong.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Averick, Mara, Bryan, Jennifer, et al. 2019. 'Welcome to the tidyverse'. *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Woodcroft, Ben J., Aroney, Samuel T. N., Zhao, Rossen, Cunningham, Mitchell, Mitchell, Joshua A. M., Nurdiansyah, Rizky, Blackall, Linda, and Tyson, Gene W. 2025. 'Comprehensive Taxonomic Identification of Microbial Species in Metagenomic Data Using SingleM and Sandpiper'. *Nature Biotechnology*, 1–6. https://doi.org/10.1038/s41587-025-02738-1.

Wu, Jinyan, He, Yuan, Zhou, Guangrong, Wei, Fuyao, Chen, Tingting, Wang, Xiaoyuan, Chen, Shenglong, Deng, Xue, and Su, Chengyuan. 2025. 'Understanding of the Effect of COD/SO42− Ratios and Hydraulic Retention Times on an MFC-EGSB Coupling System for Treatment Sulfate Wastewater: Performance and Potential Mechanisms'. *Journal of Water Process Engineering* 71: 107244. https://doi.org/10.1016/j.jwpe.2025.107244.