Metagenomic and Predictive Functional Analysis of Microbial Communities in a Novel Wastewater Treatment System

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Alison Gomeiz
Bachelor of Science
George Mason University, 2019

Director: Benoit Van Aken, Associate Professor
Department of Chemistry and Biochemistry

Spring Semester 2021
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my father and my high school chemistry teacher, Stephen Fox. I owe my deep love for science and chemistry to both of you (respectively… No offense, Dad). I would not be writing this without the support you provided to me in my most influential years as a developing scholar.

To Mr. Fox, your faith in me has lasted with me through my higher education. My early mornings and late afternoons in AP Chem Club were escapes for me during some unstable times in my young adult life. Few things in my life were as concrete to me as chemistry. Your letters of encouragement remain some of my most important documents; they even take precedence to the degree on my wall. Without the passion for chemistry that you provided me with, I would not even have that degree. Thank you for your unending patience in answering my unending questions. It was undoubtedly annoying in the context of a high school classroom. Now, it is one of my best skills as a developing researcher. And now, I will never, *ever* stop asking questions.

To my Dad, I am most thankful for your nurturing persistence in my education. Thank you for believing in me before I knew how; it truly changed the trajectory of my life for the very best. In my earlier years when I was unsure if I would excel, you provided me with the assurance that I could not give myself. You taught me the deadly combination of intelligence with hard work, and I will carry that weapon with me for the rest of my life. It has yet to fail me in any battle I have come across. I am confident it will get me through every struggle I will face in the years to come.

ACKNOWLEDGEMENTS

I would like to thank my friends, family, and supporters who have collectively been my biggest supporters in the time I have pursued my dreams in academia. To my academic advisor, Dr. Benoit Van Aken, thank you for the guidance that you have provided me over the years as your student. Thank you to my committee for the advice provided in creating this thesis. To my sister, Megan, thank you for your constant, unconditional love. It has gotten me through the stressful periods in my work and life. To my grandparents, thank you for your unending generosity, support, and love. Thank you to my mother- and father-in-law who have provided me with the utmost hospitality and direction, which enabled me to complete my education and embark on my next phase of life. Thank you to my father-in-law specifically for your invaluable help in providing me with the bioinformatics knowledge needed for this thesis. To my partner, Benjamin, thank you for writing Python scripts that aided me immensely in the organization of the data for this study. And of course, thank you for your consistent patience and encouragement; your unwavering faith in my work has been more important than you know.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Sequence Batch Reactor ..... SBR
Plug Flow Reactor..... PFR
Aerobic Granular Sludge ..... AGS
Extracellular Polymeric Substance ..... EPS
Wastewater Treatment Plant ..... WWTP
Virginia Polytechnic Institute and State University..... Virginia Tech
Continuous Flow Reactor ..... CFR
Completely Stirred Tank Reactor ..... CSTR
Hydraulic Retention Time..... HRT
Primary Effluent..... PE
Settling velocity ..... V_s
Chemical Oxygen Demand ..... COD
Mixed Liquor Suspended Solids ..... MLSS
Mixed Liquor Volatile Suspended Solids ..... MLVSS
Polyphosphate-accumulating Organism ..... PAO
Potomac Science Center ..... PSC
Upper Occoquan Service Authority..... UOSA
Next-Generation Sequencing ..... NGS
Ribosomal Ribonucleic Acid ..... rRNA
RNA Integrity Number ..... RIN
Amazon Web Services ..... AWS
Amazon Machine Image ..... AMI
Elastic Compute Cloud EC2
Nucleotides ..... nts
Base Pair ..... bp
Actual Sequence Variant..... ASV
Operational Taxonomic Unit ..... OTU
Divisive Amplicon Denoising Algorithm, Version 2 ..... DADA2
Phylogenetic Investigation of Communities by Reconstruction of Unobserved States, Version 2 ..... PICRUSt2
Kyoto Encyclopedia of Genes and Genomes..... KEGG
ANOVA-like Differential Expression ALDEx2
Benjamini-Hochberg..... BH
Principle Coordinates Analysis..... PCoA
Protein to Polysaccharide..... PN/PS
Ribonuclease ..... RNase
Two-component System ..... TCS
Acyl-homoserine Lactone ..... AHL
Small regulatory RNA ..... sRNA
Internet Protocol version 4..... IPv4

ABSTRACT

METAGENOMIC AND PREDICTIVE FUNCTIONAL ANALYSIS OF MICROBIAL COMMUNITIES IN A NOVEL WASTEWATER TREATMENT SYSTEM

Alison Gomeiz, M.S.

George Mason University, 2021

Thesis Director: Dr. Benoit Van Aken

Aerobic granulation is an emerging microbial process in wastewater treatment that has shown to improve the efficiency of conventional activated sludge systems by accelerating sedimentation, improving organic waste, nitrogen, and phosphorus removal, and increasing microbial tolerance to toxic elements found in wastewater. Aerobic granulation results in large microbial aggregates that sediment faster contribute to higher cell tolerance than bacterial flocs found in conventional AS systems. To date, granulation can only be achieved in sequence batch reactors, which are largely incompatible with the continuous flow model used in modern wastewater treatment plants. Recent research conducted at Virginia Polytechnic Institute and State University has demonstrated for the first time that a proper ratio of feast to famine conditions is able to promote aerobic granulation in a simulated plug flow reactor, composed of a suite of completely stirred tank reactors in series. Feast and famine cycles are known to negatively select for filamentous microbes

that contribute to poor aggregate density in continuous flow wastewater reactors and positively select for microbes known to contribute to biofilm formation. The goal of the present study is to understand the mechanisms of aerobic granulation present in this novel reactor system by determining changes in the microbial community composition in aerobic granules compared to conventional microbial flocs. A metagenomic analysis was conducted using 16S rDNA sequencing on aerobic granules and conventional AS samples fed by the same wastewater. Taxonomic identification and predictive functional metagenomics were completed with bioinformatics software tools Bioconductor, PICRUSt2, and BURRITO. In contrast to previous metagenomic reports of AS in contemporary wastewater treatment facilities, the results of this thesis have revealed the microbial community changes and predicted functionality of bacteria in the new reactor system that efficiently facilitated the aerobic granulation process. These findings include increased prevalence of bacterial taxa such as Comamonadaceae, *Hydrogenophaga, Flavobacterium* and *Sphingopyxis* in the PFR system with aerobic granular sludge, which are known to produce extracellular polymeric substances and are commonly identified in SBRs. Additionally, drastic decreases were observed for taxa including Actinobacteria, Chloroflexi, *Accumulibacter, Microthrix,* and *Zoogloea*, which are filamentous groups largely associated with sludge bulking issues and poor sedimentation in traditional AS. Apparent variability in abundance between the chambers of the PFRs with failed granulation and successful granulation indicate higher compositional stability in AGS. Predictive functional analysis further indicates upregulation of proteins involved in biofilm formation pathways in AGS, such quorum sensing, secretion systems, and transporters.

Upregulation of nitrogen and sulfur metabolism in AGS also agree with the expected activity of granular sludge communities compared to flocs in AS. Somewhat unexpectedly, growth rates and other metabolic functions are upregulated in AS, which may be explained by the stationary phenotype of a mature AGS system. These results indicate the selection of bacteria that can contribute to biofilm formation and inhibition of filamentous microbes, suggesting that the feast/famine profile in the PFR systems provided similarly adequate conditions to that of other successful SBR systems.

1. INTRODUCTION

Background

Urban wastewater management in the US dates back as early as 1800[1]. While the technology of wastewater treatment has drastically changed since that time, one integral principle of water purification has persisted: the use of wastewater bacteria in organic waste decomposition. Contemporary wastewater treatment plants (WWTPs) employ a three-phase system in treatment. Primary treatment is characterized by multiple cycles of screening and physical separation of sediment and particles, secondary (or "conventional") treatment involves bacterial processing via the AS process, and the third treatment include nitrogen and phosphorous removal and/or disinfection. Even through engineering advancements, primary treatment alone is largely insufficient in providing high water quality[2,3]. It is estimated that 85% of organic material in sewage is processed during secondary treatment[2]. Thus, bacteria in WWTPs are critical to providing clean water to the environment. Gaining understanding of the bacterial mechanisms involved in the wastewater treatment process may provide insight in optimizing purification processes.

One important characteristic of these microbial cells is the ability to form aggregates. These aggregates can be characterized as either flocs or aerobic granules, depending on their settling characteristics. Aggregates allow for efficient separation of treated water from microbial biomass and recycling of the biomass in the head of treatment

system. Aerobic granules are superior to flocs in their settleability, waste removal ability, resistance to toxins and environmental changes, and biomass retention[4]. In recent years, aerobic granule architectures of bacterial communities have been identified in sequential batch reactors (SBRs). In the static, aerobic environment of the SBR system, bacterial communities adhere to one another to form aggregate structures. When these flocs meet sufficient size and density requirements, they are deemed aerobic granules[5]. These structures display phenotypes that are very similar to that of bacteria in traditional biofilm structures, such as increased metabolic function, increased production of extracellular polymeric substance (EPS) to reinforce compact structure, and promoted survival of the community in unfavorable conditions. Such qualities are incredibly desirable in wastewater treatment due to improved waste removal rates of organic material[6], nitrogen[7], phosphorus[8], and other contaminants[4], facilitated sludge-water separation, increased biomass retention, quicker settlement times, community resiliency to operational parameter alterations, and inhibition of sludge bulking[9]. Collectively, these characteristics lead to a more efficient purification system compared to the flock counterpart in conventional AS[10]. Unfortunately, the most common type of reactors used in WWTPs are continuous flow reactors (CFRs). For reasons that are not well understood, CFRs do not create the conditions necessary to promote aerobic granulation like their SBR counterpart. A civil engineering research group (led by Dr. Zhi-Wu Wang) at Virginia's Polytechnic Institute and University (Virginia Tech) has recently found that a simulated PFR comprised by a series of continuously stirred tank reactors (CSTRs), is able to replicate aerobic granulation of real domestic wastewater when biomass is subjected to feast and famine

cycles. Previous research has determined that feast and famine cycles are effective operational parameters employed to reduce the prevalence of filamentous microbes in aerobic granular sludge (AGS)[11]. As such, feast and famine cycles are commonly used as a condition to promote aerobic granulation in SBRs. Prior to this finding, aerobic granulation had been observed almost exclusively in SBR systems. However, it is still unclear why aerobic granulation occurred in such a system, while continuing to remain difficult to maintain in other continuous flow systems. This novel treatment system was not characterized with respect to either the microbial community or gene expression of the AS. To better elucidate the underlying mechanisms of the successful aerobic granulation, it is essential to study the microbial community growing in the PFR system.

Some of the results of this thesis are included in a publication by the author in collaboration with Virginia Tech whereby metagenomics analysis revealed taxonomic composition of a community which provided insight into the function of the population and the types of bacteria that thrive in aerobic granules for wastewater treatment. Using the taxonomic findings from metagenomics analysis, the predictive functionality of the microbial community was also determined to depict the functional response of bacteria in relationship with their environment.

Objective

The objective of this research is to apply a bioinformatic and molecular biology approach to solve civil engineering challenges in wastewater treatment efficiency. Because wastewater treatment is almost entirely dependent on the labor completed by bacteria in activated sludge, research focused solely on the phenotype of activated sludge is largely

insufficient in explaining the causality of complex mechanisms like aerobic granulation. This is demonstrated by the inability to achieve aerobic granulation in continuous flow systems, despite decades of research. Therefore, it is essential to study the microbial community of novel aerobic granulation systems such as the PFRs described in this work in order to better understand suspended biofilms.

Specific aims

This project was completed partially in collaboration with Virginia Tech researchers who have engineered wastewater reactor systems used to implement the aerobic granulation process using domestic wastewater. The reactors used for data processing combine elements from both CFRs and SBRs in order to facilitate the generation of aerobic granules. The specific aims of this study were to:

1) Determine microbial community compositional differences of different feast/famine ratios in order to identify trends between successful and failed aerobic granulation systems.
2) Determine microbial community compositional difference between the various chambers of the successful aerobic granulation system in order to identify trends between feast and famine conditions.
3) Predict the functional composition of aerobic granules compared to microbial flocs.
4) Predict the functional compositional shifts during induced feast and famine cycles in order to elucidate metabolic changes that may explain successful aerobic granulation.

Significance

Despite decades of research to successfully generate aerobic granules in continuous flow reactor systems, AGS is still only consistently produced in SBRs. There is a substantial demand for a continuous flow architecture that is able to facilitate aerobic granulation of AS. Our civil engineering collaborators have shown, possibly for the first time, successful aerobic granulation in a benchtop continuous flow system through the implementation of feast and famine conditions. However, the mechanisms behind successful aerobic granulation are poorly understood. Prior to this study, analysis on the successful aerobic granules had only been physically characterized by parameters such as settling times and relative densities. The composition and functions of bacteria in AS are clearly critical to the success of wastewater treatment. Thus, the results of this study are essential for improving understanding of the conditions of successful AGS, which is integral for future breakthroughs in wastewater treatment.2. Literature Review: Wastewater treatment and biomass

2.1 WWTP reactors and aerobic granulation

The microbial community structure in AS varies substantially depending on type of reactor system employed. While the operational parameters of one type of reactor system may facilitate aerobic granulation of bacteria, others may lead to poor bacterial aggregates that lead to poor settling of biomass and sludge bulking issues that reduce the efficiency of waste removal by the bacterial community. The stark phenotypic difference between these microbial communities in different reactors warrants analysis on what conditions either promote or downregulate aerobic granulation.

2.1.1 Continuous flow, sequential batch, and plug flow reactors

The main reactor employed by WWTPs in the United States are CFRs due to the simplicity of design and operation[4,12–14]. CFRs employ a constant flow model whereby wastewater influent can continuously be fed into and removed from the reactor system. By contrast, wastewater is introduced in batches, not continuously, in the alternative SBR system. In this system, the bacteria of a single batch will consume the available nutrients in the tank, eventually leading to nutritional depletion, or famine, by the end of the treatment cycle. The implementation of this feast-to-famine technique is thought to be a central variable for successful granulation in SBRs[4]. This is because famine conditions are known to inhibit the growth of filamentous bacteria which compromise aerobic granular formation[11]. Furthermore, nutritional depletion is one of the common environmental conditions known to upregulate biofilm formation in bacteria[15]. Aerobic granules are sometimes called "self-suspended biofilms" due to their phenotypic similarities to traditional biofilms. Unfortunately, SBRs are not compatible with continuous flow facilities[4,16]. Despite over twenty years of research on the aerobic granulation of bacteria in SBR models, CFRs are still unable to replicate aerobic granulation. These limitations have prevented larger WWTPs from taking advantage of the aerobic granulation process and its accompanying benefits.

To address the desire for a continuous flow aerobic granulation technique, a plug-flow system was designed to mimic the feast/famine profiles in SBR systems. The PFR is an ideal reactor model approximated in the application of many industries that provides a gradient of concentration within a tubular infrastructure. A limitation of continuous flow

models is that fluids continuously added to the reactor result in near-constant concentrations of nutrients. This leads to adverse effects on the resulting product, as it does not provide the concentration gradient needed for aerobic granulation to occur. Conversely, the PFR model provides a concentration gradient across the distance of a tubular reactor. This is achieved when fluid moves at a constant velocity within the reactor, creating thin, radial "plug" cross sections of fluid that run perpendicular to the direction of flow (Figure 1). Each cross section has a concentration, and the concentration decreases in the cross sections approaching the removal site (x =L).

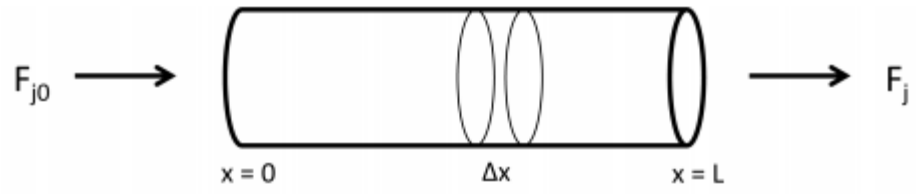


Figure 1. Schematic drawing of a plug flow reactor. The feeding site is represented by distance x = 0, and the removal site is represented by x = L. The starting and ending flow velocities are represented by Fj0 and Fj, respectively. Δx represents the "plug" cross section. In a perfect PFR system, the concentration of fluid contained in this Δx section will be uniform[17].

A common way to simulate the PFR system is to join multiple CSTRs in series such that each CSTR approximates a single "plug" cross section. Through this approximated plug flow system, the exit stream of the first CSTR becomes the feed stream of the next CSTR in the series. This way, nutrient availability decreases from one CSTR to the next, allowing for the implementation of feast/famine conditions. Concentration gradients are applied over the distance of x = L for a given period of time (often called the hydraulic retention time, HRT). This same concept is applied in SBRs, but instead of a distance vector in the PFR, the concentration gradient is applied as a time gradient[18]. Hypothetically, this same feast/famine cycle can be applied to the approximated PFR to yield successful aerobic granules[19]. Because each reactor or chamber is continually stirred, the model used in this study successfully created a continuous flow aerobic granulation reactor system[19].Thus, the PFR system attempts to reconcile the limitations of the CFR and SBR to produce AGS that can be implemented in large-scale WWTPs. In benchtop studies, PFRs can be simulated using a series of connected small CSTRs (Figure 2).

2.1.2 Operational parameters for aerobic granulation

Gravity selection pressure and feast and famine conditions

In 2019, it was discovered for the first time that a benchtop PFR fed with real domestic wastewater was able to induce aerobic granulation[19]. This was achieved through a combination of gravity selection pressure and replication of feast and famine cycles. For inducing feast and famine, the first chambers in the PFR are provided with a substrate-rich and oxygen-rich (aerobic) environment, followed by other chambers with a starvation

period (Figure 2)[20]. Due to the continuous operations of the PFR, this discovery may have important applications for AGS in large-scale WWTPs.

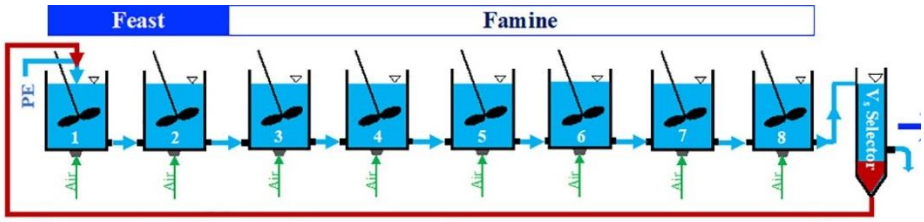


Figure 2. Schematic of an eight-chambered benchtop PFR. The domestic wastewater primary effluent (PE) is added to the first chamber and is processed sequentially to estimate plugged flow (direction of flow demonstrated by the blue arrows). Each chamber operates as a CSTR as represented by black stirring motors. All CSTRs are continuously aerated from the bottom of each chamber. The first two CSTRs represent a feast cycle; CSTRs 3-8 represent the famine cycle. After the final chamber, there is a settling velocity (V_s) selector where the densest biomass (red) is collected after a brief settling period (t = 4 min) and recycles to the first chamber[18].

It has been previously suggested that gravity selection pressure is a requisite for aerobic granulation[4,21], and recent work in civil engineering has focused on optimizing this process. Gravity selection functions by washing out suspended sludge with longer settling times while retaining the granular sludge that settles more quickly and has reportedly facilitated aerobic granulation[22]. It is notable that in CSTRs, this gravity selection process

must be coupled with a feast/famine cycle in order for aerobic granulation to take place[4]. Hence, feast/famine cycles appear to be an integral parameter for the promotion of AGS.

The effect of feast and famine cycles in wastewater treatment has been well-studied for many decades as a way to control the issue of sludge bulking in traditional AS WWTPs. Overgrowth of filamentous bacteria in WWTPs causes failure during the sedimentation step of wastewater treatment[23]. Thus, uncontrolled growth of these microbes is thoroughly detrimental to effective AS performance in wastewater treatment. Induction of sufficient famine is a very effective way to discourage filamentous bacteria from dominating the biomass, and the resulting microbial community provides peak substrate uptake, high oxygen utilization, and rapid settling times[11]. The feast and famine concept has been applied to many SBR systems in order to discourage the growth of filamentous bacteria that impede the aerobic granulation process and further improve organic removal capacity of wastewater by microorganisms[4].

Further, the ratio of feast to famine periods can drastically alter the ability of bacteria to aggregate into granules. For example, it has been reported that a feast to famine ratio greater than 0.5 inhibits the formation of aerobic granules[18]. Publications on successful feast/famine cycles employed in SBR systems report feast/famine ratios from as low as 0.16 to as high as 0.46[24–26]. Although the lower bound of the feast/famine ratio for successful aerobic granulation varies depending on operational parameters that vary between studies, values higher than 0.46 are universally attributed to aerobic granulation failure in these reports. The efficacy of feast/famine cycles have been studied mostly in SBR systems, which further highlights the novelty of the findings presented in this work.

A variety of feast to famine ratios are studied in this study: 1.0, 0.5, and 0.33. It is not expected that the two former ratios will result in successful granulation. The latter falls in the acceptable range reported in other studies, and therefore granulation is most expected in these conditions.

Additional conditions promoting aerobic granulation

While the emphasis of this project is on the feast/famine conditions (and gravity selection pressure, to a lesser extent) of a simulated PFR, there are other operational factors that must be considered in order to form aerobic granules from AS. These include applied shear stress, presence of an adequate carbon source[24], starting microbial composition of biomass, organic loading rate, volumetric exchange rate, HRT, and sedimentation time[27,28].

2.2 Advantages of aerobic granulation

Microbial communities in AGS enjoy many advantages over traditional AS that share striking similarities with traditional biofilm structures, such as stable physical structure, improved metabolic rates, higher nutrient retention, protection against toxins, environmental stress, and mechanical stress (such as shear stress)[29]. WWTPs that utilize AGS technology experience steep reductions in energy usage (up to 63%) due to these improved waste removal efficiencies and rapid sedimentation of aerobic granules[16]. AGS removal efficiencies of chemical oxygen demand (COD), nitrogen, and phosphorus increase by upwards of 67-96% compared to AS systems[8,30,31]. This is partially due to the compact structure of granules, which provide an anaerobic region for bacteria involved in phosphorus and denitrification. These anaerobic microbes support other functional groups such as nitrifiers and heterotrophic bacteria[27].

Furthermore, metabolism of these chemicals is much more stable and subject to fewer fluctuations than what is observed in AS[32,33]. Additionally, EPS in aerobic granules acts as a diffusion limitation barrier in adverse conditions to prevent toxins from penetrating biofilms such as aerobic granules[34]. This results in excellent tolerance to pH shock[35], salinity[36], and chemical shock (such as phenols[37] and pentachlorophenol[38]).

Rapid sedimentation in particular is due to the compact structure of these self-suspended biofilm aggregates, which simultaneously decrease settling times and increase the biomass that can be removed after treatment. Moreover, AGS serves to assuage a major bottleneck of modern WWTP design: reactor volume. The volume needed for a biological reactor reportedly decreased by 30% in a Polish WWTP that implemented AGS[27]. This occurs because the settled concentration of biomass, as measured by the mixed liquor (volatile) suspended solids (MLSS or MLVSS) content, is higher in AGS systems compared to their non-granulated counterparts, which increases the amount of biomass that can be removed during treatment. Simultaneously, AGS is less prone to sludge bulking issues that arise in AS that contribute poorly to sludge volume and efficiency. As a result, the overall volume of sludge generated during processing is significantly reduced[27].

2.3 Microbial community profiles in WWTPs

2.3.1 Community structure of activated sludge WWTPs

There is substantial literature describing the metagenomic community structure of bacteria in traditional AS WWTPs. These taxonomic levels have been identified down to the genus or species level in many instances.

Filamentous bacteria are commonly found in AS samples and are problematic for sludge bulking and foaming during wastewater treatment[39]. Sludge bulking reduces the efficiency of organic waste removal of biomass and is a common issue that plagues activated sludge. Some filamentous taxa that contribute to sludge bulking include the Chloroflexi phylum, Alphaproteobacteria, Gammaproteobacteria, and Actinobacteria classes, and the Saprospiraceae family[27,40–44]. Actinobacteria in particular were found to decrease from 33.7% to 14% abundance during the transition from AS to AGS treatment. It is important to note that filamentous groups are also important to the initial phases of aggregation in AGS systems. Although they may disappear from the biomass upon reaching steady state, filamentous groups are not exclusive to AS WWTPs[24]. However, the overgrowth of these groups is negatively associated with the proper function of wastewater treatment.

The microbes in AS are capable of degrading a variety of carbonaceous and nitrogenous materials. Additionally, a group of organisms called polyphosphate-accumulating organisms (PAO) are integral to the biological removal of phosphorus. Phosphorus is not as easily biodegradable as carbon and nitrogen material during conventional wastewater treatment, and thus it is poorly removed. PAOs uptake phosphate intracellularly, which makes them desirable for improving the quality of primary effluent. Common PAOs found in WWTP include Intrasporangiaceae, *Candidatus Accumulibacter*, *Dechloromonas*, *Tetrasphaera*, *Flavobacterium,* and *Sphingopyxis.* It is notable that a variety of other bacteria are capable of phosphorus accumulation, but PAOs are distinct due to their functional ability in various environmental conditions. To promote their

growth, many AS WWTPs apply an anaerobic selective pressure that promotes growth of PAOs and inhibits aerobic groups[45].

2.3.2 Community structure of AGS WWTPs

There are a limited number of studies discussing the differences observed in WWTPs that transition from traditional AS to AGS technology with the same infrastructure. Studies of this type would be ideal for illustrating microbial community shifts in aerobic granulation. These results could then be used as a benchmark for experimental designs like the PFR used in this study that possesses varying levels of granulation. To supplement these comparative studies, the community structure observed in SBRs compared to contemporary CFR WWTPs will also be included.

A WWTP in Poland that transitioned from AS to AGS technology revealed that a variety of taxa that were in higher abundance in the WWTP employing AGS, from phylum to genus level classification. EPS producers such as *Candidatus Competibacter, Flavobacterium, Sphingopxyis,* and *Dechloromonas* were differentially observed in AGS[27]. Because EPS biosynthesis is the basis of biofilm formation, it is expected that bacteria that are capable of EPS production would be more highly represented in AGS than in AS. Additional studies corroborate the presence of bacteria known to produce EPS. Some identified taxa in this functional category include *Hydrogenophaga*, *Acidovorax*[46], *Pseudomonas*, *Aeromonas*, *Arcobacter*, and *Acinetobacter*. Another integral characteristic of biofilm formation is the presence of cell surface hydrophobicity. A previous study on the relationship between cell hydrophobicity, EPS production, and community composition previously reported that increased cell hydrophobicity was associated with

higher EPS and a higher ratio of Flavobacteriales to Sphingobacteriales abundance. Conversely, high cell hydrophilicity was associated with lower EPS production and a higher abundance of Sphingobacteriales. Because EPS is directly responsible for cell hydrophobicity, it is of great interest to understand the cell hydrophobicity of specific microbes to assess their potential aggregation abilities.

3. MATERIALS AND METHODS

3.1 Sample acquisition

Concentrated AS and aerobic granule samples are obtained from the Virginia Tech Department of Civil and Environmental Engineering laboratory in Manassas, VA. Samples were taken from the Occoquan Watershed Monitoring Lab in Manassas, VA. Samples were allowed to settle after acquisition (no set time provided). Once settled, the concentrate that sedimented at the bottom was collected into Nasco Whirl-Pak™ 100 mL bags and transported to a -80°C freezer at the Virginia Tech laboratory in Manassas. These samples were then transported from Manassas to the George Mason University Potomac Science Center (PSC) campus on dry ice and stored at -80°C.

3.1.1 Data set

Activated sludge samples were collected from three different benchtop designs using 4-, 6-, and 8-chambered PFR systems (n = 4, 6, and 8, respectively) from Virginia Tech's Civil and Environmental Engineering lab in Manassas, VA. As a control, AS was acquired from real domestic wastewater (n = 2) from the Upper Occoquan Watershed Service Authority (UOSA) wastewater resource recovery facility in Centreville, VA. This same wastewater was used for seeding the PFR systems.

Samples acquired from the chambered reactors were labelled X-Yz whereby X represents the total number of chambers in the PFR system, Y represents the specific chamber number that was sampled, and z signifies triplicates taken from the chamber where applicable (labelled as A, B, and C).

3.1.2 Feast/famine ratios of the PFR systems

The feast to famine ratio of the 4-, 6-, and 8-chambered reactors were 1:1, 1:2, and 1:3 (1, 0.5, 0.33), respectively. In this study, feast condition was defined by sufficient substrate for microbial growth, whereas famine condition was defined by insufficient substrate. The end of the feast period is determined by the last chamber with a growth rate ≥ decay rate. The determination of these rates is based on COD removal and described in further detail in Sun et al. (2021).

For all three reactors, the first two chambers receiving PE were determined to be in feasted conditions, and the following chambers in famine, as shown in Figure 3. The PE fed in these PFR systems comes from UOSA domestic wastewater effluent.

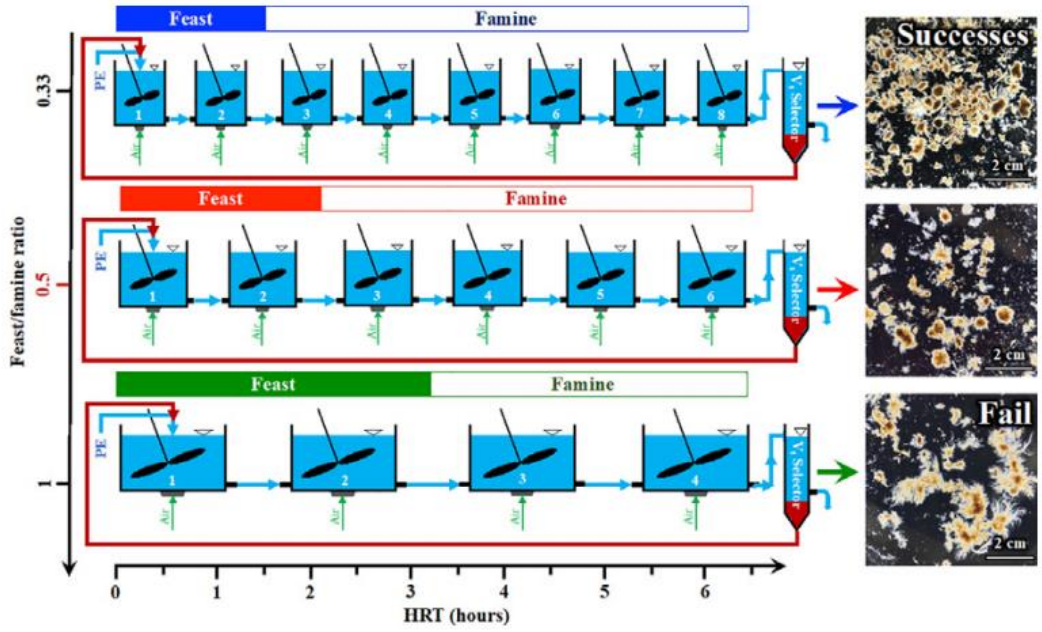


Figure 3. Diagram of the feast to famine cycles employed in the 8-, 6-, and 4-chambered reactors. Feast/famine ratios are shown vertically, and the hydraulic retention time (HRT) is shown horizontally. On the right, the morphology of the mature granules for each reactor is shown on petri dishes[18].

3.2 NGS-based amplicon sequencing

3.2.1 DNA Extraction

Samples were thawed in a 25°C water bath and removed promptly upon achieving a homogenous liquid consistency, as heating to room temperature may contribute to nucleic acid degradation. Twelve mL of thawed sample were centrifuged for 10 min at 10,000 rpm. Approximately 250 mg of the pelleted material was then extracted using the DNeasy Powersoil Kit (Qiagen) following the manufacturer's protocol[47]. The homogenization step

was completed with a BeadBug™ Homogenizer[48] for 30 seconds at 3000 rpm. The homogenization step was repeated up to three times (total time 90 seconds) to achieve full homogeneity and placed on ice in between homogenizations. Unless otherwise specified, samples were maintained on ice in between extraction steps to further minimize degradation.

3.2.2 PCR amplification

DNA samples were assessed for purity and quantity via spectrophotometry using a NanoDrop (Life Technologies). Purity is determined through A260/A280 and A260/A230 ratios. Next, 16S rRNA gene amplification was conducted via qPCR using primers that target the V3-V4 hypervariable region of prokaryotic small ribosomal subunit rRNA gene. The primers selected for the purpose of this project have been frequently reported in human microbiome 16S rRNA amplicon studies and provide the highest coverage of the Bacteria domain without detectable bias towards specific taxa within the Bacteria domain[49]. These primers are purchased from Integrated DNA technologies (IDT, Coralville, Iowa) and contain partial adapter sequences for Illumina sequencing:

341F: 5’ CCTACGGGNGGCWGCAG

785R: 5’ GACTACHVGGGTATCTAATCC

Amplification of DNA samples was conducted in triplicates and with a control (Milli-Q water). The volume of reagents per 20 µL PCR well were: 0.8 µL elution buffer (1/5 diluted), 0.1 µL 341F primer, 0.1 µL 785R primer, 10 µL ABsolute Blue QPCR, and 9 µL DNA (or Milli-Q water for control). The PCR conditions were: (a)b 2 min at 96°C, (b) 45 cycles of 25 sec at 95°C, 1 min at 50°C, and 50 sec at 72°C, and (c) 1 min at 72°C.

3.2.3 Illumina® sequencing

Following amplification, samples were shipped on ice packs to Genewiz® for Illumina®-based amplicon next generation sequencing (NGS). Sequencing results were provided as paired-end FASTQ files that are demultiplexed and have the adapters removed. Files from forward reads are labeled with the suffix "_R1", while the reverse reads are labeled "_R2". These results were obtained from the GeneWiz website and were then subject to bioinformatics analysis.

3.2.4 RNA extraction

Samples were thawed in RNAlater at 4°C for 24-48 hr prior to extraction to allow the solution to thoroughly penetrate bacterial cells. RNA extraction was then completed with both the PureLink™ RNA Mini Kit and RNeasy PowerSoil Total RNA Kit, following manufacturing protocols. Unless otherwise specified, samples were maintained on ice in between extraction steps to further minimize degradation. Quantity and preliminary quality were evaluated using a NanoDrop. Quality was then assessed with the Agilent 4150 TapeStation electrophoresis system. RNA quality is assessed by RNA integrity numbers (RIN) from 1-11 where 1 represents highly degraded RNA, and 11 represents the highest quality RNA. The results below were completed in February 2021 (Figure 4).

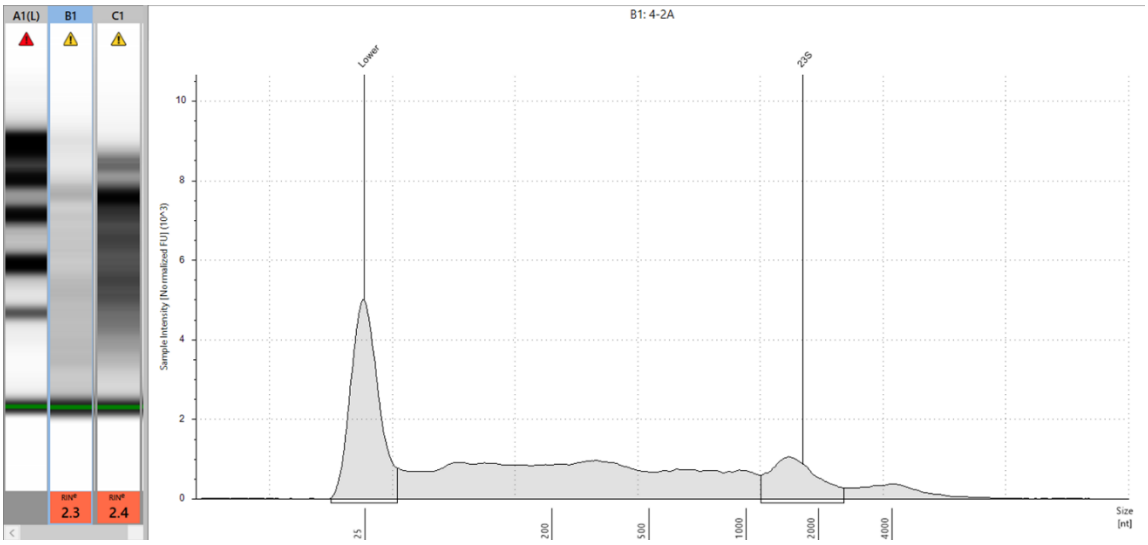


Figure 4. RNA TapeStation analysis from the 4-2 (B1) and 6-4 (C1) PFR chambers, along with the RNA ladder (A1). The gels for the three samples, their respective RIN values, and electropherogram for the 4-2 chamber are shown.

3.3 Metagenomics analysis

Metagenomic analysis was completed almost exclusively with Bioconductor. Bioconductor is an open-source software project written in the R programming language with over 3000 packages designed to analyze and visualize biological data. DADA2 is one of the most popular software packages in Bioconductor, and as discussed in *Appendix* A.3: Sample inference: Denoising or clustering, it provides a denoising algorithm superior to traditional clustering methods. An additional benefit of denoising is that the outputs, ASVs, provide better input data for identifying chimeric sequences that arise during PCR amplification. The R script for this section is provided in Appendix for the reader's convenience.

3.3.1. Amazon Web Services (AWS)

While Bioconductor can be used on any computer that can run R, some packages require extensive computer power for their algorithms. The denoising and *de novo* chimera removal steps can take weeks to complete on a conventional computer due to the complexity of the algorithms and size of the FASTQ input files. Through Amazon Web Services (AWS), bioinformatics scripts can be run on virtual machines in under an hour at $0.68 per hour[50]. A 16-core EC2 (elastic computer cloud) virtual machine accessed from AWS is used to run R script, Bioconductor, and its packages for metagenomic analysis.

3.3.2 Primer trimming

As discussed in *Appendix* A.1.1 Primer trimming, specific trimming methods are preferred for samples sequenced by brands such as Illumina due to the propensity of the types of sequencing errors that arise from Illumina-specific sequencing technology. 16S rDNA amplicons were sequenced using Illumina NGS (see *Section 4.2.2*). Thus, it was decided that specific trimming techniques would yield the highest quality output. PANDAseq and `Cutadapt` are two common primer trimming tools with specific and non-specific trimming capabilities. Both tools were tested using a set of forward and reverse FASTQ files from the 4-chambered reactor. Ultimately, `Cutadapt` was chosen to process FASTQ reads due to inconsistent selectivity observed in the PANDAseq tool, which is discussed in more detail in Appendix Specific trimming: PANDAseq vs. Cutadapt.

3.3.3 Quality filtering and trimming

Quality was evaluated and visualized using DADA2's `plotQualityProfile` function and QIIME1's VSEARCH function `fastq_stats`. Most forward FASTQ files maintained Q scores above 35, even at position 250 in the read, and thus they were not subject to quality trimming. Reverse reads were found to be lower quality on average, as expected. Up to six bases were trimmed due to poor quality across all reverse files in order to ensure a 12 bp overlap requirement for merging was fulfilled. Many reverse reads were found to have terminal Q scores of up to 30, so truncating the ends even in reverse reads was very minimal in some datasets. Quality trimming was executed by calling the `filterAndTrim` function in DADA2.

3.3.4 Learn errors

Sequence error analysis was completed by calling the `learnErrors` function in DADA2. This function contains a fully automated algorithm and has no parameters to assign. The learned error rates are then used as an input for the `dada` denoising function.

3.3.5 Dereplication

Dereplicating reads was executed using the `derep` function in DADA2. In order for reads to be dereplicated together into a single representative sequence, they must be a 100% match, or a 1.00 tolerance.

3.3.6 Inferring sample composition

Denoising methods were employed for sample inference. The advantages of denoising over traditional clustering methods are discussed in Appendix A.3: Sample inference: Denoising or clustering. The DADA algorithm was the denoising method

selected for inferring sample composition and was executed in R using the DADA2 package.

3.3.7 Merging

Complementary forward and reverse FASTQ files ("_R1" and "_R2" matches) are merged using the `mergePairs` function in DADA2. The defaults for the error threshold (zero) and the minimum overlap length requirement (12 bp) were used for merging.

3.3.8 Generate sequence table

The outputs of the merged function are single-stranded ASVs and their abundances. This data is presented as a list. In order to present this data as a matrix of samples and sequences (ASVs), the `makeSequenceTable` function is executed with the merged object as the input. The output is called a sequence table, which is a similar format to an OTU (operational taxonomic units) table. Instead of representative OTUs, the true sequences are provided in the table. This sequence table is still subject to further downstream filtering (chimera removal) and is not the final object. chimera removal are properly vetted ASVs and their abundances per sample. The final, actual sequences, and can now be saved as a text file in FASTA format.

3.3.9 Remove chimeras (*de novo* method)

The DADA2 package contains a de novo tool for chimera identification that studies the individual dataset for the presence of chimeric sequences. This tool, called `removeBimeraDenovo`, was executed to perform chimera removal. There are no parameters to assign with this function, as the only input data to identify chimeras are from

the relationships found within the sequences analyzed. The ASVs can be provided from the sequence table generated in *Section 4.3.9*.

If the input file is provided as a sequence table, the output will also be a sequence table object. This chimera-free sequence table is the culmination of the metagenomic workflow.

3.3.10 Assign taxonomy

The final step is to test the fully vetted ASVs against a 16S rRNA database. The function `assignTaxonomy` in DADA2 is employed for this step. The input files for this function are the ASVs and desired database. The chimera-free sequence table object generated from *Section 4.3.9* fulfills the former input object. The ASVs were tested using the SILVA v138 rRNA database. The output object from `assignTaxonomy` is a text file that relates the ASV number to a taxonomy. SILVA provides taxonomies up to the genus level, although some species and strains may be available for some ASVs.

Only 0.32% of reads were unable to be classified to at least the Kingdom level.

3.3.11 Reformatting output files

The two main objects generated from the DADA2 pipeline are the sequence table matrix and the taxonomy text file. These files must be reformatted for some software packages such as PICRUSt2 and BURRITO. There are three main files expected in pipelines: an ASV FASTA text file, a count table, and the taxonomy text file. The former two objects both come from the chimera-free sequence table, and the latter merges the `assignTaxonomy` object with headers from the sequence table.

The sequence table organizes ASVs into columns and samples into rows. There are often thousands of ASVs that make the size of this table unmanageably large. In order to make the data more manageable, this matrix is split into two files: the ASV FASTA file and the count table. The ASV FASTA file provides the full sequence for each ASV and a shortened header that represents each sequence. To reformat this table, the column headers (the sequences) are extracted from the sequence table. Then, a header is provided for each sequence starting with ">ASV_1", then ">ASV_2, >ASV_3, … >ASV_n" for the entire number of sequences 'n'. To fulfill proper format for FASTA files, headers must contain the ">" prefix, followed by a sequence that is tab-delimited.

To generate the count table, the ASVs are substituted in as the new column headers for the sequence table, completely eliminating the sequences from the matrix. This modified sequence table with simplified column headers is the count table.

In the taxonomy object, the row headers are changed from the default to match the shortened "ASV_n" headers in the count table and FASTA file. Column headers represent the various taxonomic levels and are not reformatted. This text file is the taxonomy file. These objects are imported into Excel for ordinance plot visualizations.

3.4 Predictive functional analysis

3.4.1 PICRUSt2

PICRUSt2 is a Python script that analyzes the metagenomic outputs generated by packages like DADA2. There are minimal parameters to assign when executing the script because it is testing ASVs (also accepts OTUs) against functional databases such as KEGG. The input files for PICRUSt2 are the count table and the ASV FASTA file (which provides

the “key” for the count table). The `--stratified` flag must be specified in this pipeline in order to generate stratified outputs. The python script for this section is provided in Appendix for the reader’s convenience

After the main pipeline `picrust2_pipeline.py` has been executed, the `convert_table.py` function is executed with the stratified outputs in order to generate the legacy format from the `pred_metagenome_contrib.tsv` object. The output of this conversion is the `pred_metagenome_contrib.legacy.tsv` object, which is needed for visualization web servers such as BURRITO.

Other outputs like `KO_predicted.tsv` are provided by the PICRUSt2 script. `KO_predicted.tsv` can be used as an input file for the KEGG Mapper tool called “Search Pathways”. The Mapper function allows the user to visualize the components of different functional pathways that are predicted to be present.

3.4.2 BURRITO and KEGG scores

The web server visualization tool BURRITO is used to find relationships between the taxonomy and predicted function of a microbiome. A limitation of metatranscriptomics analysis is that it fails to assign function to the microbial taxa contributing to the function. On the other hand, metagenomics fails to determine the true function contributed by the microbial community. BURRITO attempts to reconcile these issues by estimating the contribution of each function that is attributed to each taxon[51]. Functional annotations are based on KEGG (Kyoto Encyclopedia of Genes and Genomes) orthology groupings. The KEGG database is updated daily and contains genomic information as well as high order functional pathways.

The inputs for BURRITO are the count table, the matching taxonomy text file, and the `pred_metagenome_contrib.legacy.tsv` object from PICRUSt2. The outputs are a functional attribution table and a percent contribution table of each taxa to a specific function. These exports are then visualized in Excel.

3.5 Determination of Significance

To identify significant differences between samples, the ANOVA-like Differential Expression (ALDEx2) tool from Bioconductor was used[52]. ALDEx2 uses a centered-log ratio in order to determine differences between high throughput sequencing data. ALDEx2 was used for both determining significant differences in taxonomic classifications and functional prediction analysis. ALDEx2 employs both the Welchs test and Wilcoxon test for determining P-values between samples. If $P < 0.05$ by either test, the taxonomic classification or function was considered significant and included in downstream visualizations. Both tests employ a Benjamini-Hochberg (BH) correction of raw P-values. Hence, the output P-values are represented as `we.eBH` or `wi.EBH` objects for Welchs (we) or Wilcoxon (wi) P values with BH corrections.

The R script for the ALDEx2 package is available in Appendix .

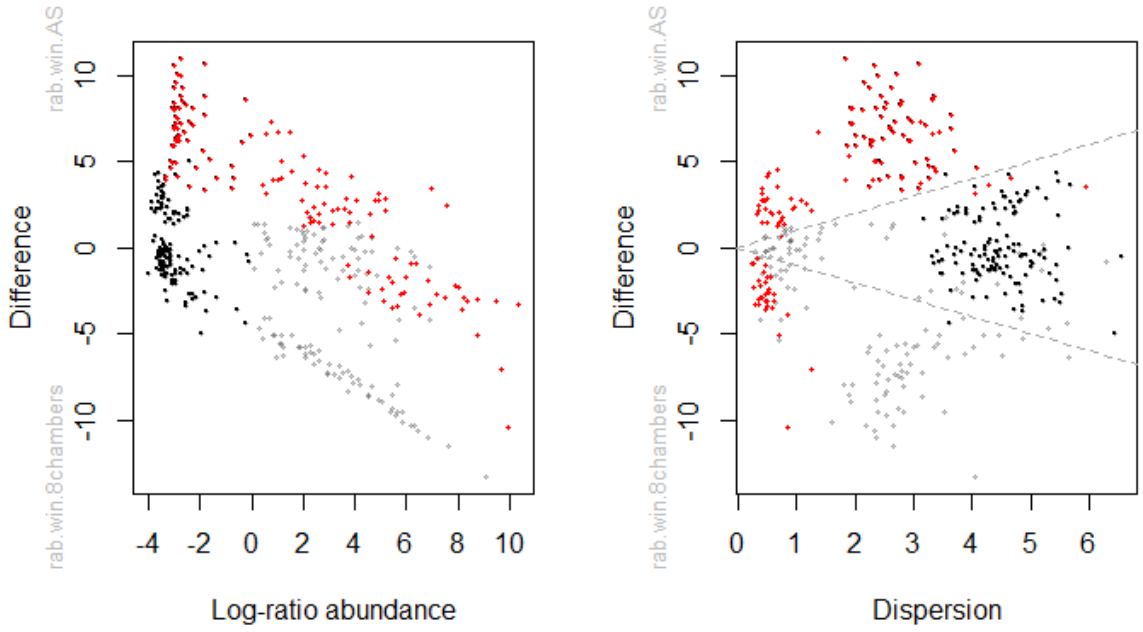


Figure 5. Schematic ALDEx2 plots between genera of the 8-chambered PFR (n= 8) and AS samples (n=2). The Bland-Altman plot, or MA plot (a), shows the relationship between the abundance (centered-log ratio) and difference. The effect plot, or MW plot (b), shows relationship between dispersion and difference. Statistically significant values in both plots are shown in red. Abundant but non-significant values are shown in grey; rare but non-significant values are shown in black.

3.6 Visualizations

3.6.1 Ordinance plots

To generate ordinance plots, total counts for each taxonomic classification per sample need to be determined. ASV count tables are imported into Excel. If needed, this table is transposed such that the ASV IDs are row headers. The taxonomic identifiers in

the taxonomy file correspond to the ASV IDs in the count table; these classifications (from phylum to genus, where applicable) are then inserted to replace these ASV IDs.

This newly merged file with the taxonomies and count table has a unique string of taxonomic classification per row (since each ASV was unique). However, two different rows may share classifications. For example, two adjacent rows may have a different genus classification but are contained in the same Flavobacteriaceae family. These two rows will need to be merged at the level up to the point where the names diverge. In this example, this would be to the family level. The python script for this merge step can be found in Appendix E. The resulting merged files can then be inserted in Excel to generate normalized ordinance bar plots.

3.6.2 Principal coordinate analysis (PCoA)

Principal coordinate analysis (PCoA) was conducted on the ASVs of the full dataset. This was completed using the combined sequence table with the full sequence in the column headers. This input file (along with a sample data table detailing the different feast/famine ratios of the samples) was analyzed using the `get_pcoa` tool from the Bioconductor package MicrobiotaProcess[53]. The R script for this step can be found in.

3.6.3 Heatmaps

Heatmaps are generated using the base R function `heatmap`. Input files used for this visualization included taxonomic count tables and functional attribution tables from BURRITO. The green/red color palette is from the gplots package from R. Color keys and histograms are generated using the `heatmap.2` tool. The R script for this visualization can be found in Appendix .

4. RESULTS AND DISCUSSION

4.1 Reactor results

Aerobic granule formation was readily observed in the 8-chambered PFR system, compromised in the 6-chambered PFR, and completely failed in the 4-chambered PFR. All of the results described in this section were obtained by our Virginia Tech collaborators, Dr. Z. Wang and Dr. Y. Sun. These results are described to provide necessary background for the metagenomic findings.

4.1.1 Physical characteristics of biomass

Sludge volumetric indices (SVI) are used to evaluate the settling efficiency of the different PFR systems. SVIs are presented as functions of time such as SVI_5 and SVI_{30} (volumes recorded at t = 5, 30 minutes). $SVI_{30} \leq 60$ mL/g and SVI_5:$SVI_{30} \leq 1$ are typical benchmarks for good settleability of AGS[4]. SVI values obtained at steady state (Table 1) show successful aerobic granulation in the 8-chambered PFR, compromised granulation in the 6-chambered PFR, and complete failure of granular aggregation in the 4-chambered PFR. Similarly, the specific densities of the 8-chambered PFR align with values previously reported for AGS (1.13 to 1.20[54–57]), whereas the 4- and 6-chambered PFRs resemble values associated with AS (1.001 to 1.01[54,55,58]). Sludge morphology results (Figure 6a) reveal irregular, fluffy, large, poorly circular flocs in the 4-chambered PFR system. The 6-chambered reactor has a more regular shape, although the particle size is still large and loose. In the 8-chambered reactor, dense and spherical granules are visually dominant. These observations are confirmed through the particle size distribution in Figure 6*b*. Most

flocs in the 4-chambered PFR are over 8 mm in diameter, compared to 4-5 mm in the 6-chambered PFR, and 1-2 mm in the 8-chambered PFR. The circularity distribution of the particles (Figure 6*c)* show strongest circularity in the 8-chambered PFR that then declines in the 6- and 4-chambered PFRs. Interestingly, the evolution of the granular morphology between chambers of the 8-chambered PFR (Figure 6d) reveal a mixture of flocs and spherical aggregates. Their size and circularity are initially comparable to the 4-chambered PFR values, but there is a drastic improvement observed in the third chamber, coinciding with the famine cycle.

These results show a relationship between the feast/famine ratio and physical characteristics. This correlation is confirmed in Figure 7, which shows strong, monotonic relationships ($R^2 > 0.85$) between the feast/famine ratio and various parameters. This includes a negative relationship between feast/famine ratio and circularity, zone settling velocity (V_{zs}, indicates the speed by which biomass settles[59]), and specific gravity (R^2 = 0.95, 0.96, and 0.95, respectively) and positive relationships between feast/famine ratio and median granule diameter, SVI_{30}, and SVI_5:SVI_{30} (R^2 = 0.99, 0.96, and 0.86, respectively). These correlations illustrate that longer famine periods are associated with improved settleability, while longer feeding periods are associated with poor settleability and morphology. The SVI_{30} in the 6-chambered PFR reveals an interesting deviation from these trends; instead of increasing, the SVI_{30} is nearly identical to the SVI_{30} of 8-chambered PFR (Figure 7c). This reveals that the 0.5 feast/famine ratio possesses some physical characteristics of AGS, suggesting a transitional phenotype that exists between activated granular sludge and traditional AS. This phenotype is observed in the physical appearance

of the aggregates of the 6-chambered PFR (Figure 6a) whereby large filamentous flocs are observed alongside granules observed in the 8-chambered PFR. This aligns with literature that find the upper boundary of feast/famine ratios for successful aerobic granulation to be 0.5.

The changes in these physical characteristics at different feast/famine ratios may be explained by compositional shifts in the microbial community. Because sludge bulking is due to an overgrowth of filamentous bacteria, poor settleability parameters may be related to high abundances of filamentous bacteria. Conversely, AGS with good settleability should correspond to a a high abundance of EPS-producing bacteria.

4.1.2 Biochemical characteristics of biomass

The reduction in total COD from the biomass strongly agrees with the first order rate law for all three PFRs ($R^2 > 0.93$). The rate is fastest in the 8-chambered PFR, followed by the 6- and 4-chambered PFR (rate constants = 1.4, 2.0, and 2.2 hr^{-1}, respectively). Mixed liquor suspended solids (MLSS) are generally composed of bacteria and other suspended compounds. Mixed liquor volatile suspended solids (MLVSS) are slightly different in that they only consider "volatile" solids, which are materials that burn in 550°C temperatures (the overwhelming majority of which is bacteria)[60]. MLSS and MLVSS values are higher in AGS compared to AS, and theses higher concentrations are associated with improved COD removal rates. This relationship demonstrates excellent fitness (Figure 7d) for MLVSS and the rate constant versus the feast/famine ratio ($R^2 > 0.98$ for all three parameters). Improved waste-clearing capabilities is an integral advantage of AGS systems

and may be explained through compositional and phenotypic differences of bacteria in granules.

Across all PFR systems, the highest EPS concentration is observed the 8-chambered PFR (Table 1), in both major components: polysaccharides (PS) and proteins (PN). Both PS and PN are integral to promoting the granular aggregation of bacteria in wastewater treatment[21]. An increase in these values is indicative of increased cohesive activity in the 8-chambered PFR, which indicates increased prevalence of EPS-producing microbes. When evaluating EPS production in this PFR across its chambers, the PN/PS ratio decreases with time (HRT). It is reported that the consumption of PN during famine is favorable for cell hydrophobicity of AGS because it acts as an energy source and increases the representation of hydrophobic PS[61,62]. Generally, PN content in EPS is positively correlated with cell hydrophobicity[63,64] due to hydrophobic amino acid side chains[65]. If PN/PS is too high, increased prevalence of hydrophilic amino acid side chains may decrease overall hydrophobicity, which is highly destabilizing to a microbial community's ability to develop granules[66]. While it is not always the case[63,64,67–70], high PN/PS ratios are usually attributable to fluffier flocs. Thus, the reduction in PN/PS over time may suggest that famine plays a vital role in maintaining a stable granule structure.

Table 1. Physical and biochemical parameters of the 4-, 6-, and 8-chambered PFR systems at steady state. Published by Sun et al. (2020).

Parameters		Units	Chamber number		
			4	6	8
SVI_{30}		mL g^{-1}	68 ± 9	53 ± 7	52 ± 3
SVI_5:SVI_{30}		N/A[a]	1.56 ± 0.06	1.42 ± 0.04	1.18 ± 0.03
V_{zs}		m h^{-1}	7.9 ± 1.2	11.8 ± 1.5	15.0 ± 0.7
d_{50}		mm	12.4	4.1	2.1
Circularity (median value)		N/A	0.09	0.34	0.55
Specific gravity		N/A	1.03 ± 0.09	1.12 ± 0.07	1.19 ± 0.07
EPS	PS	mg g^{-1}	9.5 ± 0.5	9.9 ± 0.7	12.0 ± 0.8
	PN		20.4 ± 1.0	22.6 ± 1.5	38.2 ± 2.0
MLSS		mg L^{-1}	942 ± 122	2169 ± 145	2299 ± 177
MLVSS		mg L^{-1}	798 ± 102	1857 ± 129	2012 ± 172
SRT		days	2.0	5.0	6.4

[a] N/A: not applicable.

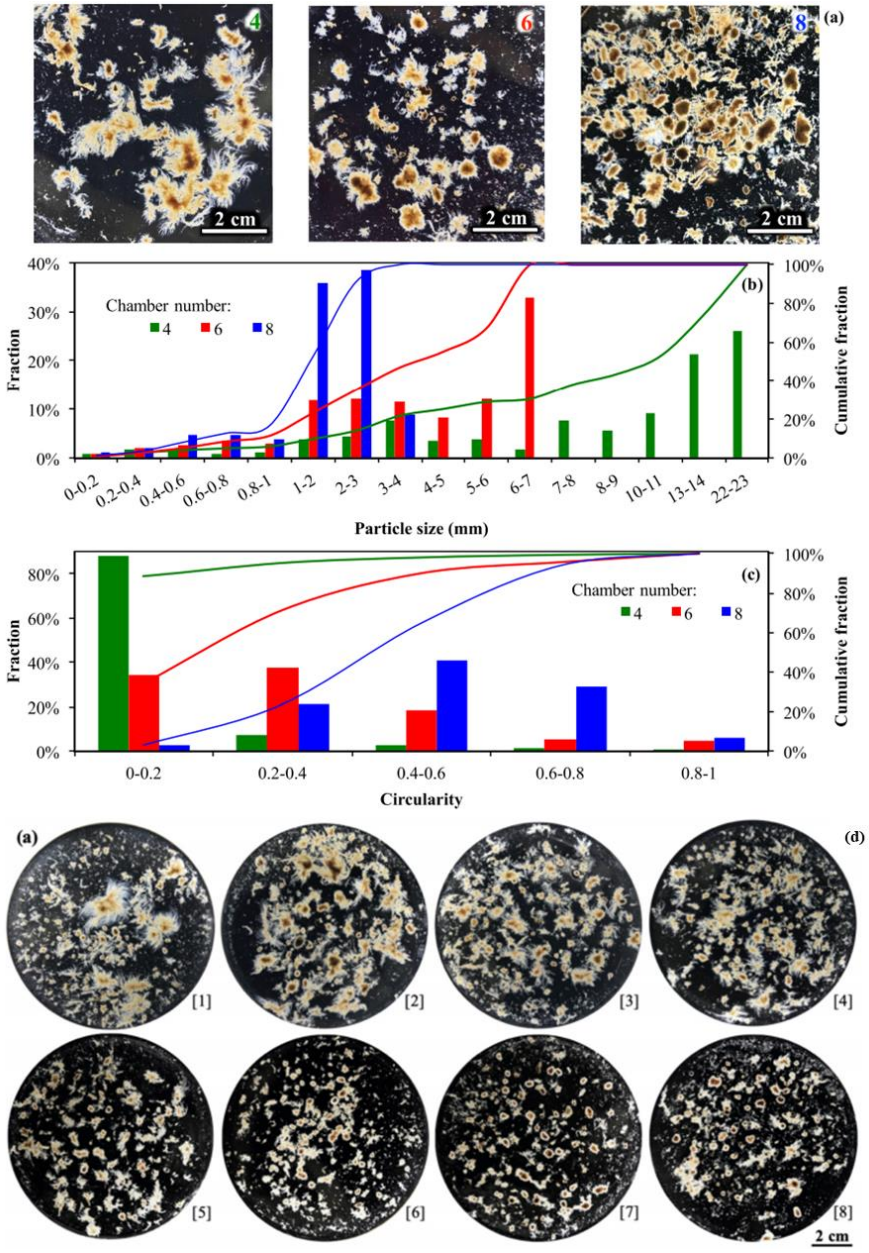


Figure 6. Sludge morphologies at steady state. Petri dish photos (a), particle size distributions (b), and circularity distribution of particles (c) are shown for the different PFR systems. Petri dish photos across the chambers of the 8-chambered PFR are also shown (d). Published by Sun et al. (2020).

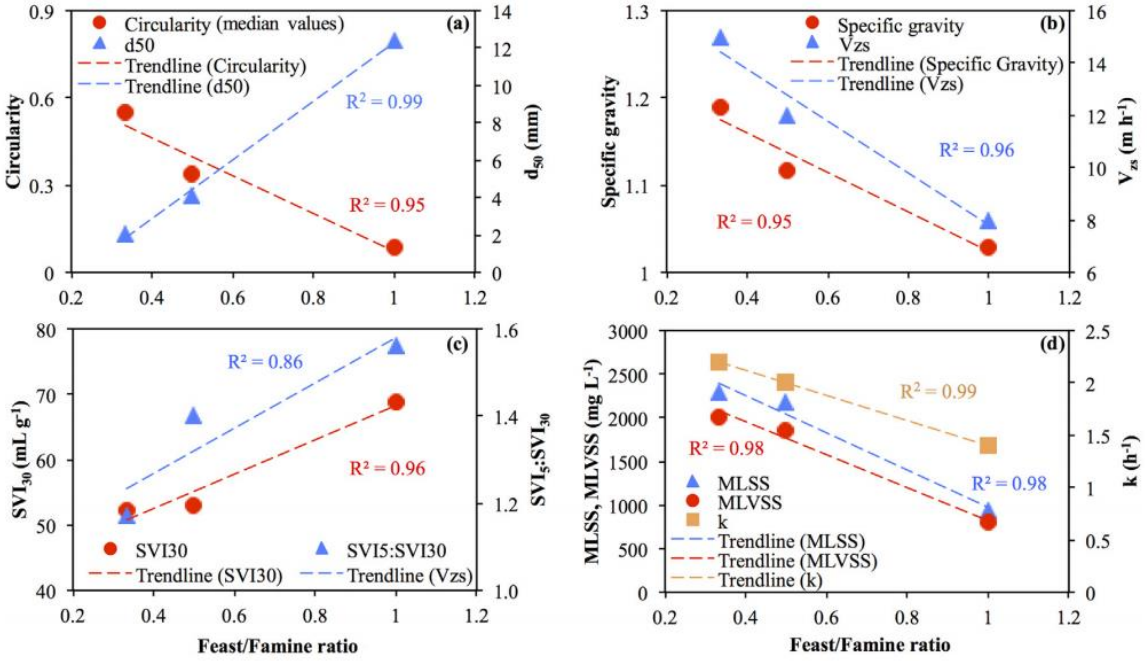
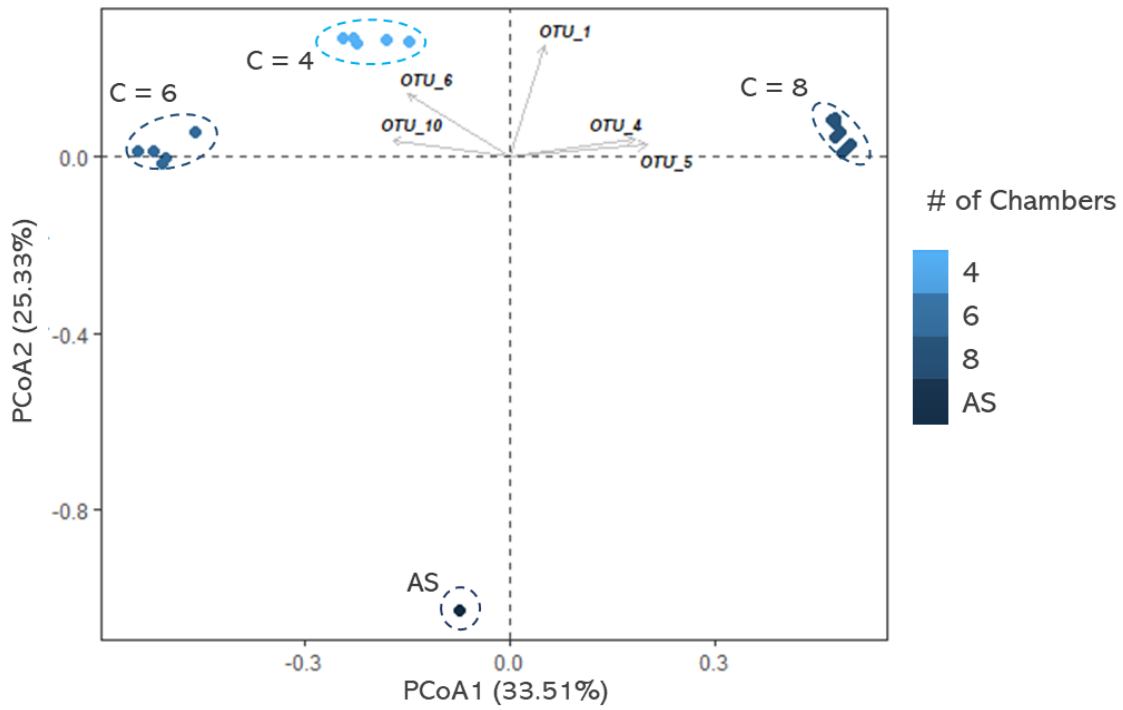


Figure 7. Dependence of sludge characteristics such as median particle diameter size (d_{50}) and circularity (a), SVI_{30} and SVI_5:SVI_{30} on feast/famine ratios (b), the zone settling velocity (V_{zs}) and specific gravity (c), and the rate constant (k), mixed liquor suspended solid concentration, and mixed liquor volatile suspended solid concentration (d). Published by Sun et al. (2020).

4.2 Community profile of AS and AGS

To visualize large-scale trends within the data, a principal coordinates analysis (PCoA) was conducted on the entire dataset. PCoA is a useful analytical tool for complex data such as metagenomics results through simplification of input parameters. As is shown in Figure 8, distinct separation between the four sample sets is observed. The strongest similarity between groups is observed between the 4- and 6-chambered PFRs. Clustering is weakest in these two groups, indicated by the larger ellipses around the individual samples. Clustering is improved in the 8-chambered PFR samples and is strongest in the AS samples (n = 2). This may suggest more compositional heterogeneity in the 4- and 6-chambered PFRs (and the 8-chambered PFR, to a lesser extent) between different chambers, which is not observed between the AS samples. The isolated clustering of the 8-chambered PFR and AS from other sample sets suggest distinct taxonomic abundances. The relatedness between the 4- and 6-chambered PFRs indicates more similar phenotypes. The "OTU" vectors provided correlate with the five ASVs that are most influential in the clustering process. From these vectors, it is shown that the taxa clustered most strongly with the 8-chambered PFR include *Sphaerotilus*, *Arcobacter*, and Bacteroidia; the taxa clustered most strongly with the 4- and 6-chambered PFRs are *Hydrogenophaga* and *Haliscomenobacter*.



OTU_1: Gammaproteobacteria; Burkholderiales; Comamonadaceae; *Sphaerotilus*
OTU_4: Campilobacterota; Campylobacteria; Campylobacterales; Arcobacteraceae; *Arcobacter*
OTU_5: Bacteroidota; Bacteroidia; NA; NA; NA
OTU_6: Proteobacteria; Gammaproteobacteria; Burkholderiales; Comamonadaceae; *Hydrogenophaga*
OTU_10: Bacteroidota; Bacteroidia Chitinophagales; Saprospiraceae; *Haliscomenobacter*

Figure 8. Principle coordinates analysis (PCoA) of all ASVs for all datasets. The control (AS) and three different PFR systems are clustered in dashed ellipses. The vector of five main OTUs are shown.

Next, heatmaps were generated at the phylum, class, order, family, and genus levels of the dataset to ascertain additional macroscale trends at each taxonomic rank. Similar trends in microbial abundance between the sample sets is observed across different taxonomic ranks. Hence, it can be concluded that the phylum rank (Figure 9a) largely

determines the clustering for the lower ranks (Figure 9b-e). In agreement with the PCoA plot, there is strong similarity between the 4- and 6-chambered PFRs. Some variations in abundance are observed in the 4-2 chamber and the 6-4 chambers at the phylum level, which may explain the poorer clustering of these chambers in Figure 8. In the 8-chambered PFR, some variation in abundance becomes clear at the order level (Figure 9c) in the 8-4 and 8-5 chambers. Because variation is not observed at higher taxonomic ranks, this may contribute to the stronger clustering shown in the PCoA for the 8-chambered system. Interestingly, there seems to be similarity in the highly abundant taxa of the 8-chambered PFR and AS samples. Conversely, the PCoA plot indicates distinct separation between these sample sets. To better determine the extent of similarity between these two groups, ALDEx2 was used to find taxa with abundances that are significantly different between the 8-chambered PFR and AS. The percentage of these significantly different taxa are shown for these two groups and other sample sets in Table 2 at each taxonomic rank. In accordance with both PCoA and heatmap trends, the strongest similarity is observed between the 4- and 6-chambered PFRs. Interestingly, high similarity is also observed between the 6- and 8-chambered PFR. These similarities were not visible in either the PCoA or heatmaps. This finding provides the first indication of an intermediate microbial compositional state in the 6-chambered PFR (feast/famine = 0.5). This aligns with the morphological and settling results of aggregates in the 6-chambered PFR that revealed compromised granulation (Figure 6) and the literature which indicates a successful granulation threshold at a feast/famine ratio ≤ 0.5. Thus, it is of great interest to determine the abundance of specific groups that explain the intermediate physical characteristics of AGS in the 6-chambered

PFR. The strongest dissimilarity observed between AS and the 8-chambered PFR reveals that the apparent trend of similarity in the heatmaps is misrepresented. There are clearly large red areas in the heatmaps for both AS and 8-chambered PFR samples, but the results from Table 2 indicate that the specific taxa that are overrepresented vary substantially between the sample sets. The high dissimilarity between the 8-chambered PFR and AS indicate that the largest metagenomic differences of the dataset exist between the PFR system with AGS and the continuous flow system with traditional activated sludge.

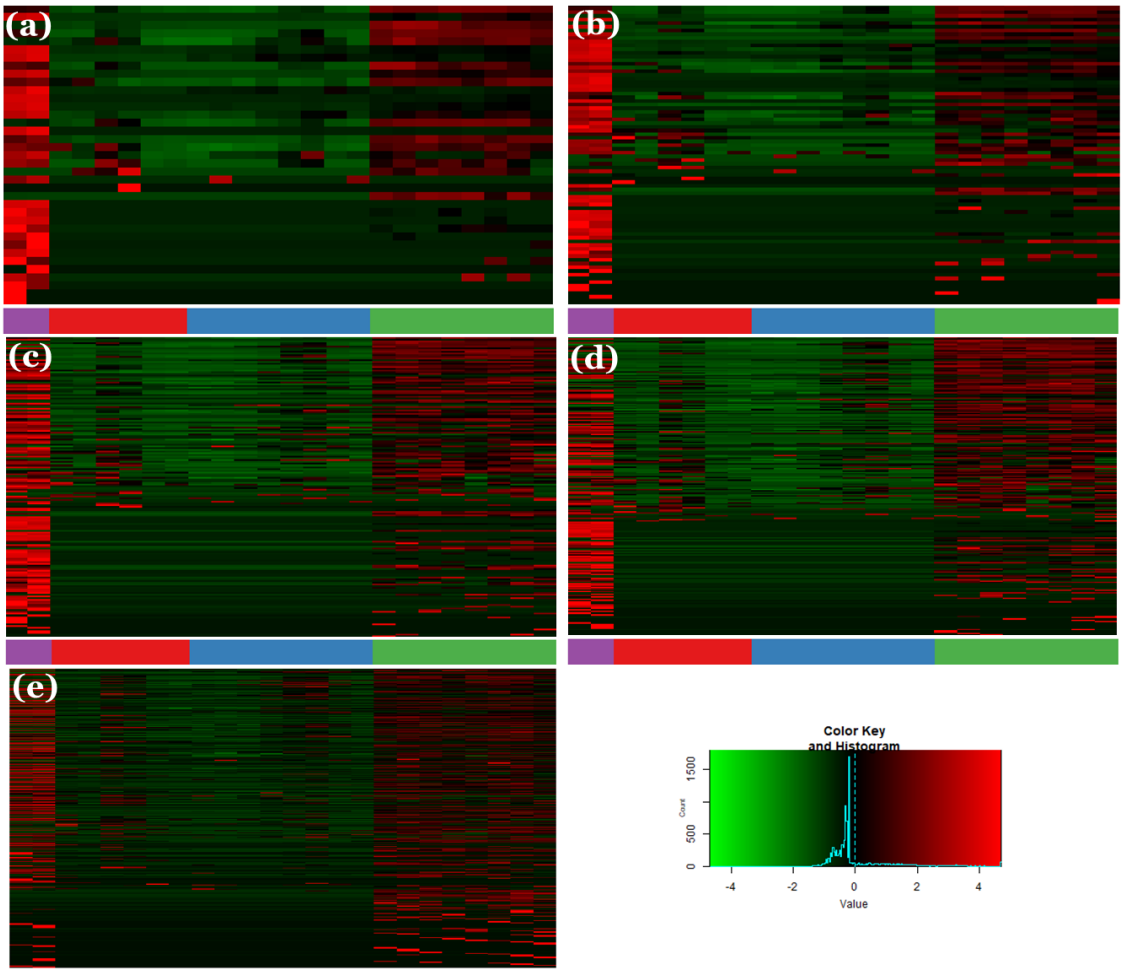


Figure 9. Heatmap across all samples across all major taxonomic classifications: phyla (a), class (b), order (c), family (d), and genera (e). Rows represent the various taxonomic identifiers at the given level. Columns represent samples. AS samples are shown in purple, the 1.0 feast/famine ratio PFR in red, the 0.5 feast/famine ratio PFR in blue, and the 0.3 feast/famine ratio PFR in green.

Table 2. Statistical outputs from ALDEx2 showing the percentage of taxa identified as statistically significant ($P < 0.05$) between two sets of data. Percentages are provided for each taxonomic level.

	Samples compared					
Level	**4, 6**	**6, 8**	**4, 8**	**6, AS**	**4, AS**	**8, AS**
Phylum	20%	33%	47%	57%	57%	40%
Class	15%	30%	35%	41%	47%	61%
Order	19%	36%	37%	35%	40%	60%
Family	18%	31%	32%	36%	34%	42%
Genus	14%	5%	26%	31%	28%	31%
	→ Increasing dissimilarity					

4.2.1 Phylum, class, and order distributions

The phyla that are differentially abundant in AS (Figure 10a) largely agree with trends that have been previously described in continuous flow AS systems[27,46,71,72]. Some of these taxa, such as Chloroflexi and Planctomycetota, contain filamentous groups that are associated with sludge bulking and foaming issues in WWTPs[44]. There is a distinct difference in abundance between the AS and PFR systems. The higher abundance of filamentous groups in AS compared to the PFR systems suggests a successful environmental pressure in the feast/famine systems against the selection of filamentous bacteria. Trends in abundance between the different PFRs are not well enough defined at this taxonomic rank to explain differences in their morphologies.

Myxococcota is not a phylum commonly identified in AS or AGS, likely due to a recent reclassification of Deltaproteobacteria into four novel phyla (including Myxococcota)[73]. [27,74]. Unexpectedly, *Haliangium* accounts for over half of Myxococcota in AS samples (Table 3). Discussion on this halophilic genus in coastal marine environments in WWTPs is limited. Studies surveying community compositions in different AS WWTPs have reported that *Haliangium* is one of the most significantly differential abundant taxa[75]. One paper assessing the microbial composition during winter operation of an AS WWTP in winter identified *Haliangium* as a core genus[76]. The UOSA facility and PFRs in this study were both sampled during the winter, but a high abundance is not observed in the PFRs, indicating temperature is an unrelated factor. It is most likely that feast and famine conditions discourage proliferation of *Haliangium*, though abundance of Myxococcota (*Haliangium*) in AS lacks a clear explanation.

The phyla most abundant in the 8-chambered PFR (Figure 10b) agree with trends previously described in the literature on aerobic granules collected from SBR systems[27]. The most abundant phylum across all PFR chambers and AS (AS) samples is Proteobacteria (Figure 10), which agrees with the available literature for both AGS and AS systems. The percent abundance ranges from 32% to 73% across all samples. These results may suggest that high levels of feast and famine (1.0 and 0.5) select for particularly high abundances of Proteobacteria. Bacteroidota is the second-most abundant phylum across all samples. The low abundance of AS is comparable to the 4- and 6-chambered PFRs. The significantly high abundance observed in the 8-chambered PFR may be explained by the relationship between Bacteroidota, cell hydrophobicity, and EPS production of AGS. A study on variable cell hydrophobicity of AS and AGS found that Bacteroidota was present in only trace amounts (1%) in hydrophilic microbial communities, but in hydrophobic communities, its abundance increased to nearly 50% of the AGS community[66]. This in turn was associated with higher production of EPS and stronger settleability of. This is further affirmed by the fact that a majority (>60%) of Bacteroidota in the 8-chambered reactor belong to the Flavobacteriales order, which are well-known to produce EPS in biofilms[46]. Thus, the high abundance of Bacteroidota in the aggregates of the 8-chambered PFR suggests increased EPS production and granule maturation. The similar abundances of Bacteroidota in the 4- and 6-chambered PFRs to the control (AS) suggest that the higher feast/famine ratios do not properly select for EPS producers needed for successful granulation. In the 8-chambered PFR, about 74% of the microbes in Campilobacterota belong to *Arcobacter*, which was identified as a major OTU in Figure 8. *Arcobacter* has

previously been identified as the most prevalent genus in mature AGS[74]. Thus, the high abundance of Bacteroidota and Campilobacterota in the 8-chambered PFR suggests improved EPS production of these aggregates. This is supported by the morphological results of the aggregates in the 8-chambered PFR, which show mature granule structures.

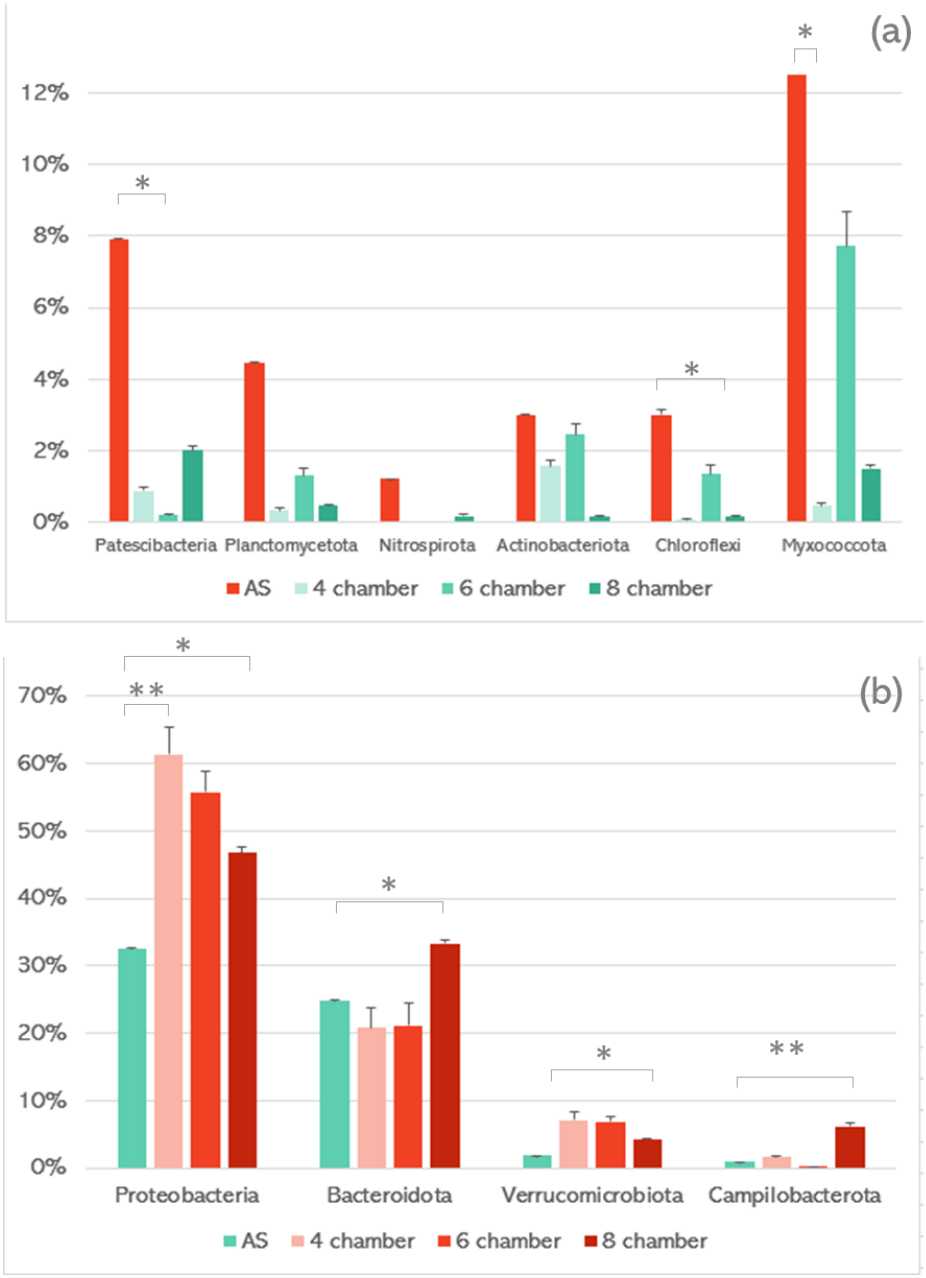


Figure 10. The microbial class assortment distribution of the 4-, 6-, and 8-chambered PFR systems and primary effluent from UOSA (AS). Phyla expected to be in higher abundance in AS (a) and phyla expected to be higher in AGS (b) are provided.

*P ≤ 0.05; **P ≤ 0.02

The higher abundance of Polyangia, Parcubacteria, Anaerolineae, and Actinobacteria classes in AS (Figure 11) support previously published findings on the relative abundance of these classes in AS compared to AGS[27,77]. These trends continue to support the distinction in communities between the PFR and UOSA facility microbes.

About 84% of Proteobacteria is composed of Gammaproteobacteria, so identical trends are observed in Figure 11. Bacteria belonging to Gammaproteobacteria in particular contain filamentous microbes that are some of the main contributors of sludge bulking issues in WWTPs[39]. These issues arise when an overgrowth of filamentous microbes disrupt the compact structure of granular aggregates, resulting in fluffy flocs with poor settling characteristics that negatively impact the efficiency of the WWTP. Through properly coordinated feast and famine periods, the proliferation of these filamentous groups is controlled[11] to ensure proper settleability of the biomass. If the feast period is too long relative to famine, then filamentous groups are not properly controlled, resulting in activated sludge with large flocs instead of dense granules. The trends between the PFRs (Figure 15) are well-aligned with these concepts. The highest percent abundance of Gammaproteobacteria is observed in the 4-chambered PFR, followed by the 6-chambered PFR. The poor morphology and settling results of the biomass in these PFRs correspond to failed or compromised granulation. A predominance of filamentous microbes in these PFRs could explain the large, fluffier flocs, low specific densities, and low settling volumes. The lower abundance of Gammaproteobacteria in the 6-chambered PFR and partial settling success can be explained by the intermediate feast/famine ratio employed

in this system, as the apparent threshold for successful granulation ≤ 0.5. The lowest abundance observed in the 8-chambered PFR further supports this idea, as the best morphology and settling parameters were observed for these aggregates. It can be concluded that the 0.33 feast/famine ratio is sufficiently low enough to control the abundance of certain filamentous groups like Gammaproteobacteria to promote successful granulation.

If filamentous bacteria were solely deleterious to aerobic granules, it would be expected that the abundance of Gammaproteobacteria be lower than 39% in AGS. On the contrary, filamentous groups are central to successful aerobic granulation in moderate abundances. Filaments are necessary at the start of granulation to extend into substrate and absorb nutrients during the feasting phase[24] (Figure 13). In fact, previous studies have found that Gammaproteobacteria may even increase slightly during transitions from AS to AGS, indicating that this class essential for proper operation of successful AGS[27]. This would explain why the lowest abundance observed is in the AS samples in this study. The overall trend observed across PFRs and AS suggests that while overrepresentation of these groups can be detrimental to the aerobic granulation process, the presence of filamentous groups such as Gammaproteobacteria are needed for successful AGS.

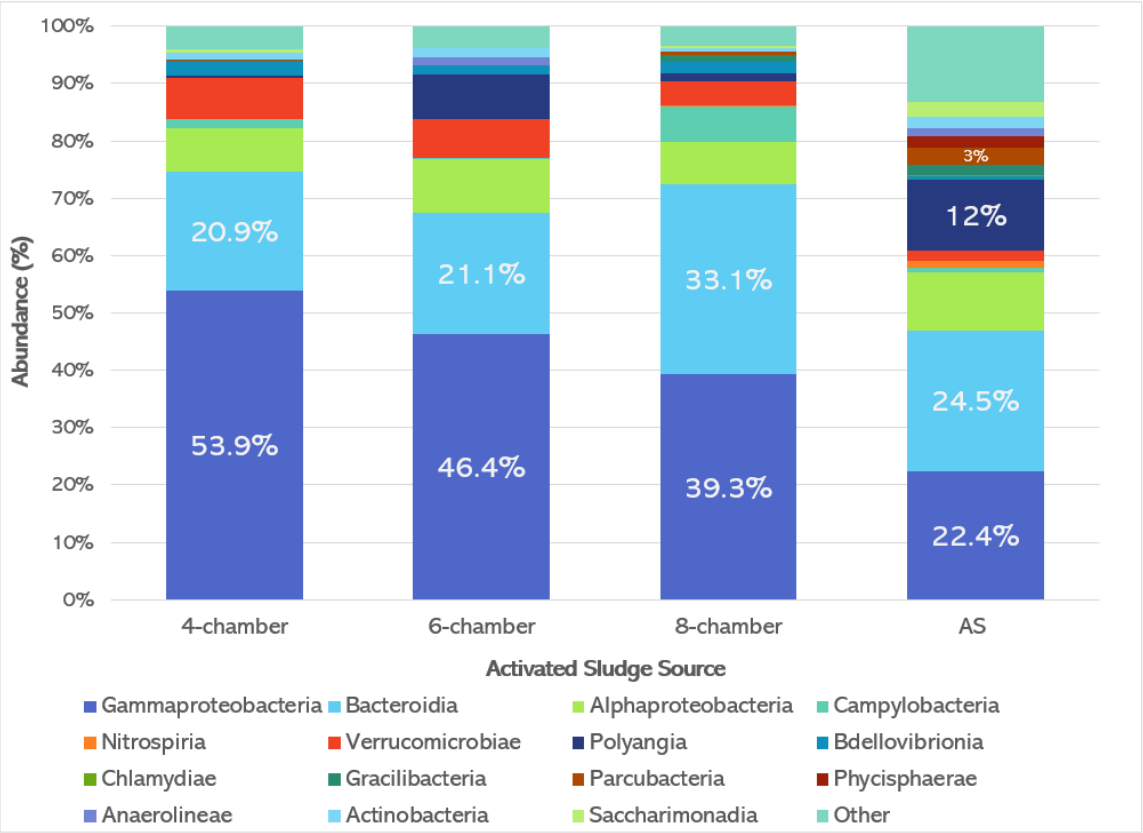


Figure 11. The microbial class assortment distribution of the 4-, 6-, and 8-chambered PFR systems (with feast/famine ratios of 1:1, 1:2, and 1:3, respectively), AS (AS) from UOSA, and partially nitrifying aerobic granules. Only the top 15 classes of the total bacterial abundance are presented.

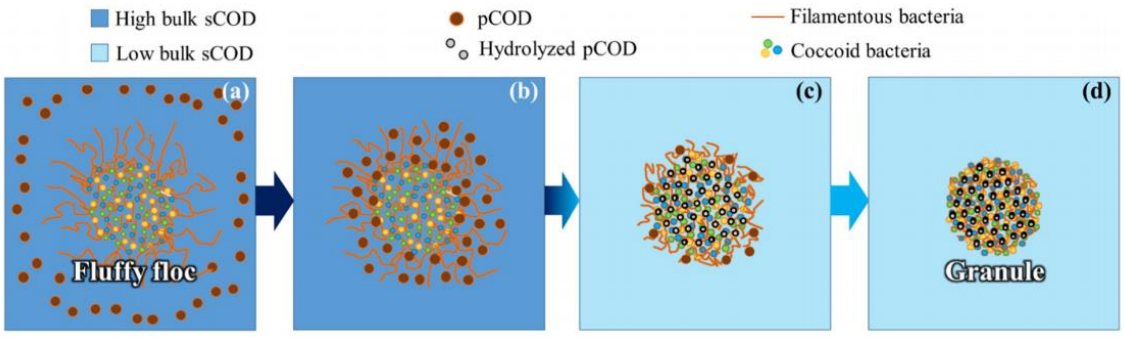


Figure 12. Chronological diagram of the structural and metabolic transformation of a microbial aggregate in a feast (a-b) and famine (c-d) period. In the feast phase when sCOD is highly concentrated, filamentous bacteria are fully extended from the floc (a) which allows for them to capture particulate COD (b). In the famine phase when sCOD is in low concentration, the filamentous groups congregate inwards in the aggregate for endogenous metabolism of the captured particulate COD (pCOD) (c). In the final maturation step, the filamentous bacteria are fully retratcted and are held together by a mature EPS matrix (d). Courtesy of Sun et al. (2021).

About 70% of Gammaproteobacteria is composed of Burkholderiales. Identical trends discussed for Proteobacteria and Gammaproteobacteria are observed in Burkholderiales (Figure 13). AGS communities predominantly composed of Burkholderiales correlated with slow-settling and fluffy granules[78]. Thus, the same relationship between the morphological results of the PFRs and abundance of Gammaproteobacteria applies to Burkholderiales.

Campylobacterales (from the aforementioned phylum Campilobacterota), Sphingomonadales, Pseudomonadales, and Flavobacteriales are found in much higher abundance in the 8-chambered PFR than AS and other PFR systems. The latter of these three orders are known to produce EPS[79], and microbes belonging to Campylobacterales (such as *Arcobacter*) demonstrate a high correlation to granulation and have been exclusively identified in AGS bioreactors[27,74,80]. Thus, at the order level, it becomes clearer that there are distinct functional differences between the 8-chambered PFR and the other PFR systems. Other orders related to EPS activity are indiscriminately represented in the three PFR systems, such as Thiotrichales and Verrucomicrobiales. 100% of Thiotrichales is composed of *Thiothrix*, a genus that has been linked EPS production[68] and successful granulation in systems where it accounted for over half of the total microbial community[69]. *Thiothrix* is found exclusively in the PFRs; its abundance was undetectable in AS.

An interesting shift between Sphingobacteriales and Flavobacteriales is observed between AS and the PFRs. A previous study evaluating the effect of cell hydrophobicity on community structure and EPS production in AGS found that when hydrophobicity was increased, a shift from Sphingobacteriales to Flavobacteriales was observed. Following the proliferation of Flavobacteriales was an increase in EPS production. When the cellular community was hydrophilic and dominated by Sphingobacteriales, there was minimal EPS production and failed granulation[66]. This finding it notable because cell surface hydrophobicity is important for biofilm attachment[83]. A similar trend is observed in this study, particularly in the 8-chambered PFR due to the high abundance of Flavobacteriales. The percent abundance of Flavobacteriales in the 4-chambered PFR is comparable to AS,

and abundance in the 6-chambered PFR is between the two PFRs. The relationship between the relative abundance of Sphingobacteriales in AS and Flavobacteriales in the PFRs suggests highest cell hydrophobicity in the 8-chambered PFR, which correlates to the most optimal environment for EPS production. This is supported by the high EPS content reported for the 8-chambered PFR (Table 1) and morphological results. The intermediate abundance of Flavobacteriales in the 6-chambered PFR is also supported by the intermediate success of morphological and settling results reported. The compositional profiles in AS and the 4-chambered PFR suggest an inadequate community profile for proper cell hydrophobicity, which would then lead to poor EPS production. This explains the poor EPS content reported in this study for the 4-chambered PFR.

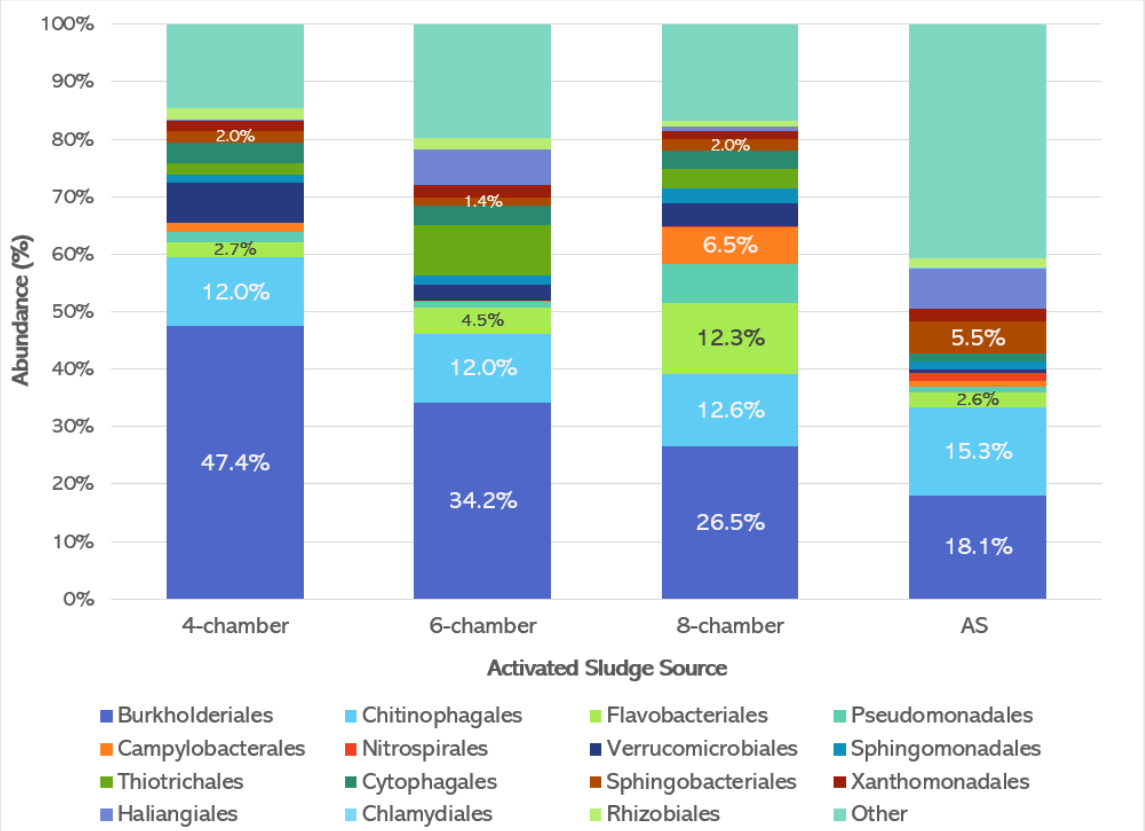


Figure 13. The microbial class assortment distribution of the 4-, 6-, and 8-chambered PFR systems (with feast/famine ratios of 1:1, 1:2, and 1:3, respectively), AS from UOSA (AS), and partially nitrifying aerobic granules. Only the top 15 orders of the total bacterial abundance are presented.

4.2.2 Family and genus distributions

A variety of low level taxa are reported to be in higher abundance in either AS or AGS in the literature. Important taxa found to be in higher abundance in metagenomic reports on continuous flow WWTPs are bolded in the AS column of Table 3, while taxa that were found in higher abundance from SBR WWTPs are bolded for the 8-chambered

PFR column, where successful granulation is observed. If the expected trend from the literature disagreed with the findings of this study, the taxa are highlighted in blue

Table 3. Percentage of the most important classes, orders, and genera between AS and the three different PFR benchtop reactors. Taxa that have been differentially identified in AS in continuous flow reactors are bolded in the AS column. Taxa that have been differentially identified AGS systems from SBRs are bolded in the 8-chambered PFR (AGS) column. *P ≤ 0.05; **P ≤ 0.01; ***P ≤ 0.001.

				Number of Chambers			
Class (phylum)	**Order (family)**	**Genus**	**AS**	**4**	**6**	**8 (AGS)**	
Actinobacteria (Actinobacteria)			**1.90 ± 0.03**	1.1 ± 0.2	1.7 ± 0.2	0.56 ± 0.03	*
	Micrococcales	*Tetrasphaera*	0.13 ± 0.05	0	0	**0.00 ± 0.00**	***
		Intrasporangiaceae	**0.236 ± 0.004**	0	0	0.005 ± 0.002	
Nitrospiria (Nitrospirota)			**1.22 ± 0.01**	0	0	0.18 ± 0.03	
	Nitrospirales	*Nitrospira*	**2.43 ± 0.02**	0	0	0.26 ± 0.04	***
Anaerolineae (Chloroflexi)			**1.53 ± 0.06**	0.09 ± 0.01	1.3 ± 0.2	0.08 ± 0.01	*
Acidimicrobiia (Actinobacteria)			**2.67 ± 0.01**	0.41 ± 0.04	0.8 ± 0.1	0.80 ± 0.05	
	Microtrichales	*Candidatus Microthrix*	**0.29 ± 0.03**	0	0	0.033 ± 0.007	
Gammaproteobacteria (Proteobacteria)	Pseudomonadales	*Acinetobacter*	0.34 ± 0.04	0.80 ± 0.05	0.8 ± 0.2	**2.3 ± 0.2**	**
	Burkholderiales	*Zoogloea*	**11.25 ± 0.09**	0.43 ± 0.05	0.35 ± 0.04	1.3 ± 0.1	***
		Dechloromonas	0.68 ± 0.10	0.9 ± 0.1	0.12 ± 0.02	**0.62 ± 0.03**	*
		Candidatus Accumulibacter	**1.61 ± 0.08**	0.020 ± 0.01	0	0.007 ± 0.004	***
	Burkholderiales (Comamonadaceae)		5.4 ± 0.2	46.6 ± 4.0	31.6 ± 4.5	**24.4 ± 1.6**	***
		Hydrogenophaga	1.07 ± 0.04	2.6 ± 0.5	22.4 ± 2.5	**8.2 ± 0.6**	
		Acidovorax	0.50 ± 0.09	3.3 ± 0.2	2.2 ± 0.3	**3.1 ± 0.4**	
	Xanthomonadales (Xanthomonadaceae)		0.08 ± 0.01	0.34 ± 0.04	1.8 ± 0.2	**0.90 ± 0.07**	*
	Competibacterales	*Candidatus Competibacter*	0	0	0	**0.010 ± 0.002**	**
	Thiotrichales	*Thiothrix*	0	2.7 ± 0.2	11 ± 1	**4.7 ± 0.2**	
Bacteroidia (Bacteroidota)			24.46 ± 0.03	20.9 ± 3.0	21.2 ± 3.3	**33.1 ± 0.4**	*
	Flavobacteriales		2.64 ± 0.05	2.7 ± 0.3	4.5 ± 0.8	**12.4 ± 0.3**	***
		Flavobacterium	1.771 ± 0.001	3.4 ± 0.4	5.1 ± 0.9	**11.2 ± 0.3**	***
	Cytophagales		1.41 ± 0.07	3.6 ± 1.0	3.3 ± 1.0	**3.1 ± 0.2**	***
	Sphingobacteriales		**5.6 ± 0.2**	2.0 ± 0.3	1.4 ± 0.2	2.0 ± 0.1	**
Alphaproteobacteria (Proteobacteria)			**10.16 ± 0.02**	7.6 ± 0.5	9.3 ± 0.5	7.4 ± 0.5	***
	Sphingomonadales		1.39 ± 0.09	1.4 ± 0.2	1.7 ± 0.3	**2.5 ± 0.2**	***
		Sphingopyxis	0.03 ± 0.03	0.01 ± 0.01	0	**0.06 ± 0.03**	
	Caulobacterales		0.41 ± 0.03	0.41 ± 0.3	1.0 ± 0.1	**1.01 ± 0.08**	***
	Rhodobacterales	*Rhodobacter*	**0.44 ± 0.06**	0.8 ± 0.1	0.5 ± 0.1	0.19 ± 0.03	***
	Rhizobiales (Hyphomicrobiaceae)		1.615 ± 0.005	2.03 ± 0.07	2.2 ± 0.2	1.10 ± 0.06	**

Differentially prevalent in AS

The taxa that are more abundant in the UOSA AS samples overwhelmingly agree with the expected trends based on the literature of continuous flow WWTPs (Table 3).

Many of these groups (Hyphomicrobiaceae, *Candidatus Microthrix*, and *Zoogloea*) are attributed to filament or adhesin production[84] known to cause bulking, foaming issues, poor settling times, and poor compaction ability in AS WWTPs[31,44,105], which aligns with the expected morphology of AS compared to AGS. Furthermore, the low abundance in the PFR samples suggests that feast/famine conditions play a role in controlling these filamentous groups that are detrimental to successful granulation. It is interesting to note that out of the three PFRs, these taxa are most abundant in the 8-chambered PFR (Table 3). This may be due to the role of substrate acquisition of filamentous groups in the early feasting phases of SBR systems. Filamentous bacteria are often encountered on the surface of granules during feasting phases in order to obtain nutrients needed for endogenous metabolism of the granule core during subsequent famine[11]. Furthermore, findings indicate that *Zoogloea* is found early in granule maturation, decreases as maturation progresses, and is capable of EPS production[46]. This suggests that some amount of *Zoogloea* may be integral for both capturing nutrients and granular structure, while excess abundance may be attributable to uncontrolled filamentous growth. Thus, the dual functionality of filament and EPS production may explain the moderate presence of Zoogloea in extended famine conditions.

Decreased *Nitrospira* abundance in the 8-chambered PFR is possibly due to the change in biomass structure from the AS to an AGS system; *Nitrospira sp*. are nitrite-oxidizing bacteria that utilize oxygen in their environment. In AS, free diffusion allows for easier access of oxygen. In aerobic granules, oxygen penetration is restricted, worsening conditions for growth for this genus[27,86].

The high relative abundance of PAOs in AS samples in this study (*Candidatus Accumulibacter*, Intrasporangiaceae, *Tetrasphaera, Dechloromonas*) suggests a negative selection pressure of these taxa in the PFRs. This aligns with literature that indicates aerobic granulation discourages the growth of most PAOs in SBR systems[27,74]. Despite being more abundant than the PFR systems, it is notable that the abundance of PAOs in AS is much lower than anticipated for continuous flow systems[27]. This may be due to the method of phosphorus removal used at the UOSA. As discussed in Section 2.3.2 Community structure of AGS WWTPs, some WWTPs employ microbes for phosphorus removal through an initial anaerobic cycle[45] with phosphorus enrichment[87] that selects for proliferation of anaerobic microbes such as the aforementioned PAOs. However, the UOSA system removes phosphorus by feeding primary effluent through a high-lime process[88]. The lack of an initial anaerobic cycle may explain the poor abundance of these taxa in AS. Furthermore, this same primary effluent is immediately subjected to aerobic conditions when fed to the PFRs, which explains the further drop in abundance observed in all PFR systems. *Flavobacterium, Thiothrix,* and *Sphingopyxis* are known to accumulate phosphorus as well[27,81], and the dominating abundance of these genera in the 8-chambered reactor (>15% total abundance) suggest competitive activity with other PAOs commonly observed in AGS and AS. Hence, other microbes outcompete the already poorly represented PAOs for substrate[89].

Differentially prevalent in AGS

Hydrogenophaga, *Acidovorax*, *Cloacibacterium*, Xanthomonadaceae, *Aeromonas*, *Acinetobacter*, *Flavobacterium*, and *Sphingopyxis* are all commonly identified in granular

sludge reactors[78,90–95] and are known to produce EPS[46,54,96]. Notably integral for bacterial aggregation is the Xanthomonadaceae family, which produces *N*-acyl-homoserine-lactone, a component of EPS and a QS molecule. The drastic increases in abundances of these groups from AS to the 8-chambered PFR suggest that the low feast/famine ratio strongly selects for these EPS-producing bacteria. Interestingly, the abundance of some groups such as *Hydrogenophaga* and Xanthomonadaceae are highest in the 6-chambered PFR. This suggests that the threshold feast/famine ratio of 0.5 may be low enough to positively select for some EPS-producing groups as well. Jointly, the trend of increased abundance of EPS-producing bacteria in the 6- and 8-chambered PFRs support the morphological findings of the aggregates in these systems, thus relating the abundance of EPS producers to successful granulation.

Many of these EPS-producing genera are present in the top 15 most abundant genera in the 8-chambered PFR (Figure 14). Other genera include *Bdellovibrio*, which are highly motile and flagellated[97], and *Novosphingobium*, which are non-filamentous, known to form microcolonies[98], and possess strong cell surface hydrophobicity[99]. The strong differences in abundance of these genera between the 8-chambered PFR and AS show the distinct selection for bacteria that are either capable of EPS production or facilitate the formation of biofilms in the PFR system.

Interestingly, *Sphaerotilus* is the largest genus in the 8-chambered PFR (Figure 14). Many microbes belonging to this genus are largely associated with the production of long filaments (up to 1000 µm) in wastewater that contribute substantially to sludge bulking and foaming issues in AS[100]. , although some reports have identified *Sphaerotilus* in dense

bioflocs[101]. Such a high predominance of this genus in the 8-chambered reactor reflects poorly on the structural stability of granule. More puzzling is the non-existent abundance of *Sphaerotilus* in AS, which indicates that a variable in the PFRs positively selected for the overgrowth of these microbes. Overgrowth of *Sphaerotilus* has been linked to high organic loading rates of simple biodegradable substances like acetate and glucose[102,103]. These microbes are known to store polysaccharides and poly-β-hydroxybutyrate in granules as reserve material[72], so it is possible that such conditions in the PFRs erroneously selected for this genus. It is unclear why the abundance of Sphaerotilus is so poorly represented in the 6-chambered PFR compared to the 4- and 8-chambered PFRs (Table 3).

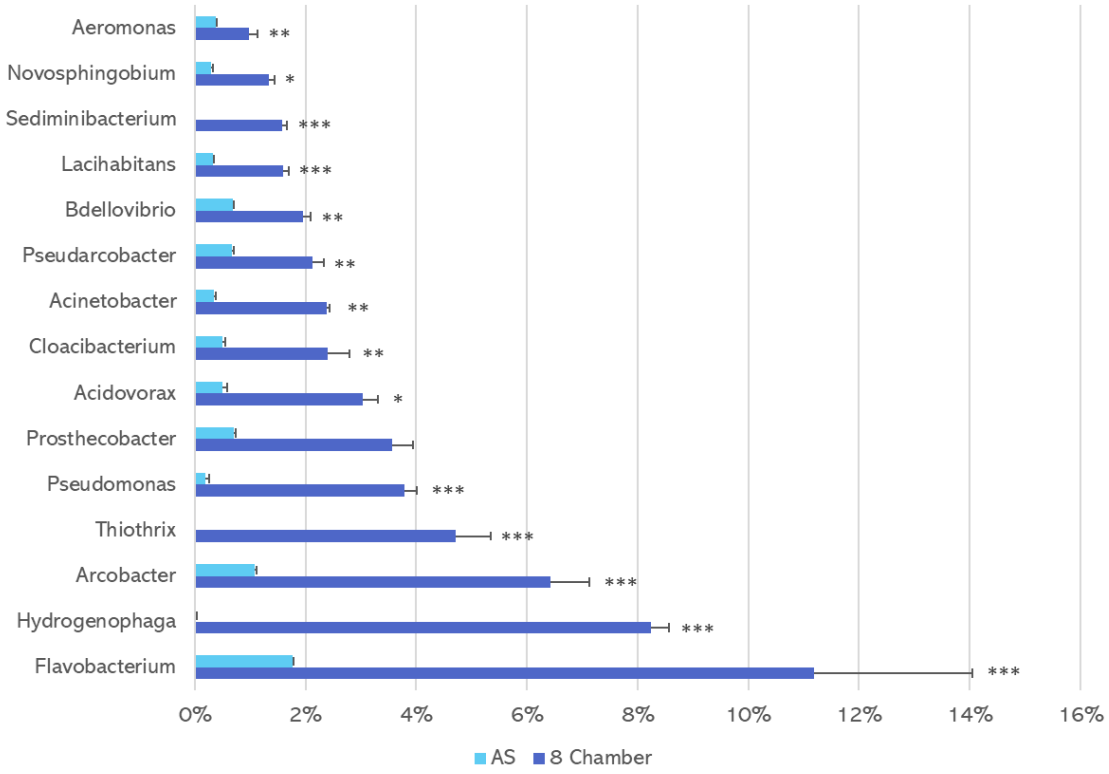


Figure 14. Percent abundance of the 15 most abundant genera in the 8-chambered PFR (excluding Sphaerotilus). The abundances in AS are also provided for comparison.

4.3 Community profile on the dynamic response of aerobic granulation

Figure 15 shows the community profile of Gammaproteobacteria in the individual chambers of all three PFR systems. It is notable that there is more dramatic variation of Gammaproteobacteria in the 4- and 6-chambered PFRs (σ = 11.2, 10.9) compared to the 8-chambered PFR (σ = 3.4). The decrease and rise in abundance observed in these PFRs does not appear in the 8-chambered PFR. These results show a compositional instability in the failed AGS reactors (4-6 chambers) and are accompanied by higher average abundances.

Conversely, greater stability in the 8-chambered PFR with AGS is accompanied by lower average abundance. This suggests a stronger compositional integrity in systems with successful AGS, whereas microbial flocs are more likely to shift dramatically with changes in environmental conditions. The variability observed in the 4- and 6-chambered PFRs "boom-and-bust" shift in community structure has been previously described in aggregated microbial communities during changes to nutrient availability[104]. The induction of feast and famine in the PFR may explain the diversity in abundance in the direction of plug flow, due to nutrient availability reduction. Conversely, it has been revealed that microbial community diversity in AGS decreases sharply in the early stages of maturation and stabilizes around key groups such as EPS producers[105]. This would explain the static microbial community composition observed in the 8-chambered PFR, since the aerobic granules were sampled at steady-state.

The apparent heterogeneity of the microbial community in the PFRs with failed granulation (4 and 6 chambers) indicate variable biological response during changes to nutritional availability between chambers. Implementation of appropriate feast/famine conditions in the 8-chambered PFR led to less pronounced community changes, likely due to the stabilizing characteristics identified in AGS. Metagenomic analysis is insufficient in explaining the factors at play in maintaining a consistent microbial community structure. These microbial mechanisms are elucidated in Section 4.4 Predictive functional analysis through predictive functional analysis.

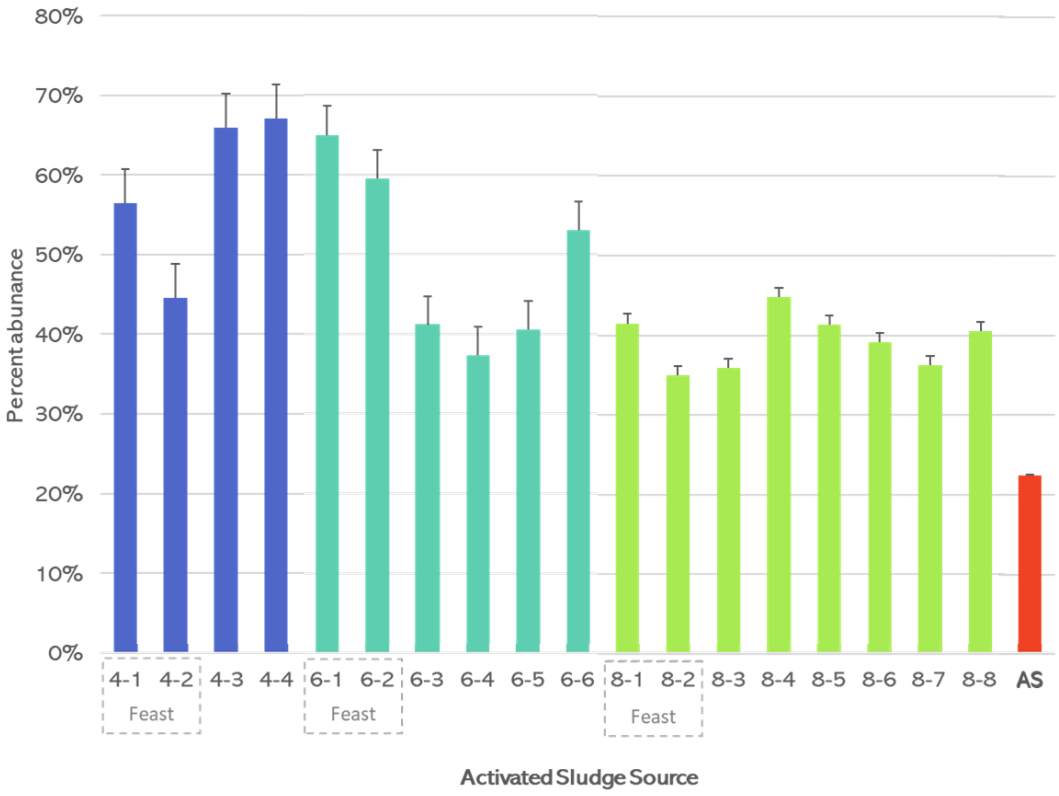


Figure 15. Percent abundance of Gammaproteobacteria across all chambers of the three PFR systems, accompanied by the AS samples (average is shown). Data labels are provided for percent abundances associated with transitions from feast to famine conditions. Triplicate samples (4-2 and 6-4) are presented as averages.

For the majority of other taxa, there is little compositional change between chambers. Plug flow models are not designed to change the composition of microbes during the hydraulic retention period, only the concentration of nutrients. As microbes flow through a decreasing gradient of nutrients, it is more likely that the epigenetic response of

the microbial community will change, rather than the actual community composition. It is notable that the standard error rates of percent abundance between chambers decreases as the number of chambers increases, further indicating that the highest compositional stability occurs in AGS.

4.4 Predictive functional analysis

Metagenomic analysis, particularly at family and genus ranks, provide associative relationships between classification and function. Further elucidation of function is typically determined through RNA sequencing and subsequent metatranscriptomic analysis. Unfortunately, RNA extractions yielded materials of insufficient quality for sampling, as determined by poor RIN values during Tapestation analysis. Despite three months of alterations to extraction protocols, evidence of high degradation persisted. It is hypothesized that the RNA degraded during the sampling process, prior to extraction. After collection from the PFRs, biomass was left for an indiscriminate amount of time to allow aggregates to settle prior to being placed in a -80°C freezer; this is problematic because RNA begins degrading within 30 minutes at room temperature[106]. It is also likely that ribonucleases (RNases) present in the biomass during sampling contributed to the degradation in this time. Resampling with protective measures such as use of an RNA stabilizing reagent and flash freezing was proposed. However, due to state-wide laboratory shutdowns in the spring of 2020 and subsequent shutdown of the PFR systems, resampling could not be conducted. As an alternative to RNA sequencing, microbial community function was predicted by testing ASVs against the KEGG functional database in

PICRUSt2. Functions determined to be significantly different ($P < 0.05$) between the AS and AGS (8-chambered PFR) samples are shown in Figure 16.

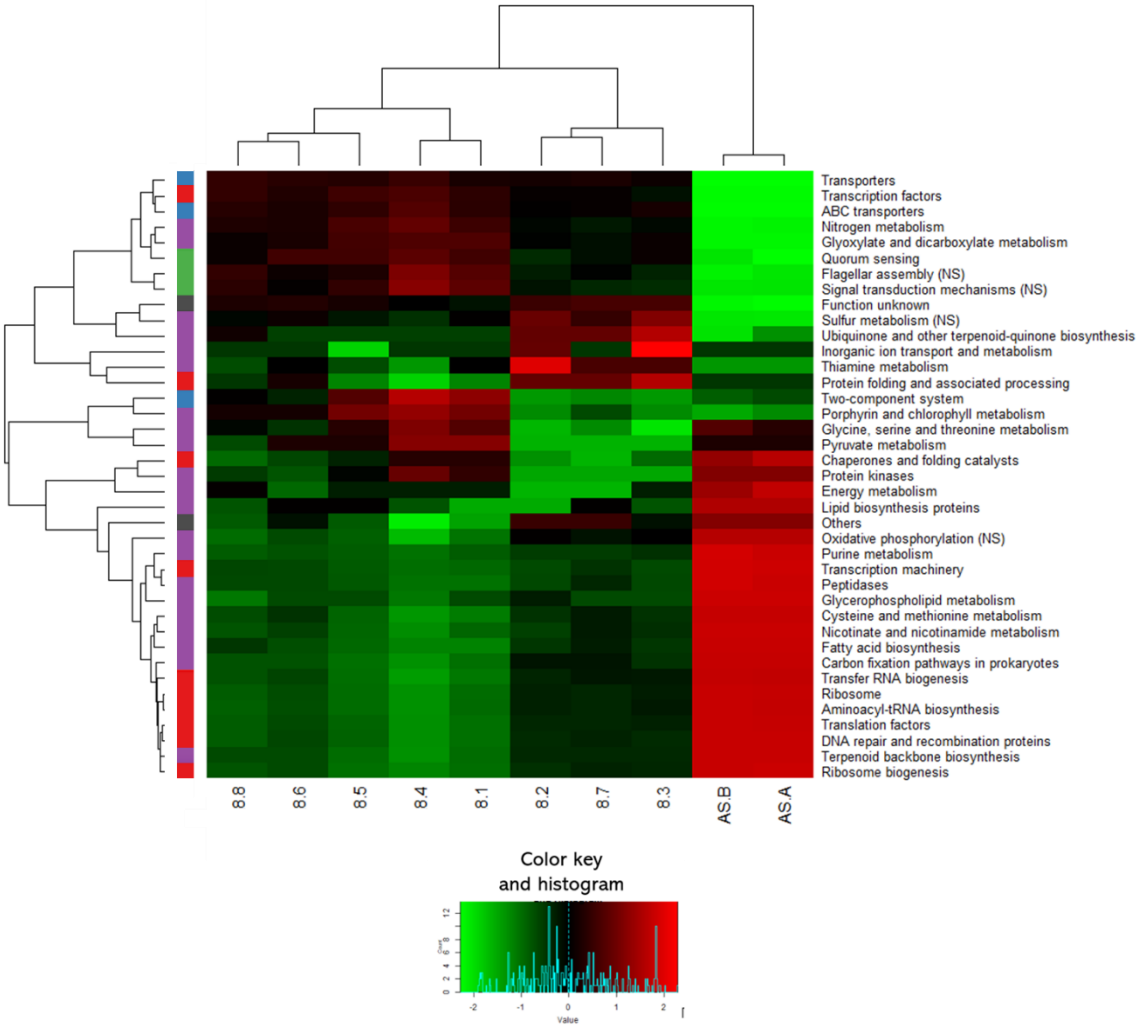


Figure 16. Heatmap of significantly different ($P < 0.05$) KEGG-predicted functions between AGS and AS samples. Functions and samples are clustered by similarity. KEGG SubPathways are represented by a colored legend: cellular processes (green),

environmental information processing (blue), genetic information processing (purple), and unclassified (grey). Notable functions with $P > 0.05$ are labeled NS (non-significant).

4.4.1 Biofilm formation

Secretion systems and transporters are integral for EPS production and dissemination; in conjunction with quorum sensing, upregulation of these functions in AGS indicates increased cell-to-cell communication, which aligns with the phenotype observed in mature biofilms[107] . Two-component systems (TCSs) are the predominant regulatory signal transduction systems that microbes use to detect and physiologically respond to environmental fluctuations. These cellular responses include biofilm formation, which is a common response to environmental stressors like nutrient deprivation. As such, TCS has been identified as the main pathway associated with the complex mechanism of biofilm formation in microbial communities[108]. The significant upregulation of these systems in AGS is then well-explained from both a biofilm formation perspective and as a response to famine periods in the PFR system. The taxa that contribute the most substantially to these functions include *Hydrogenophaga*, *Flavobacterium*, *Pseudomonas*, and *Acidovorax*.

The KEGG website identified a variety of biofilm formation pathways from the 8-chambered PFR results, such as the autoinducer-2 (AI-2) quorum sensing pathway that mediates interspecies communication for biofilm formation (Figure 17). AI-2 molecules entering the receiving cell (via transporters) are phosphorylated, allowing phospho-AI-2 to bind and inactivate LsrR. LsrR is a regulatory protein that represses transcription of genes

related to the biosynthesis of the EPS component, colanic acid. In addition to responding to incoming AI-2 molecules, it is clear that the microbial community in AGS are also producing AI-2 molecules through the activation of LuxS. This pathway was strongly identified in *Hydrogenophaga* and *Flavobacterium*. Other AI-2 quorum sensing pathways involving Lux proteins (such as the LuxO pathway) were identified in Pseudomonas and Aeromonas. Acyl-homoserine lactones (AHLs), a class of quorum sensing molecules found exclusively in gram-negative bacteria, were also identified through Las and Rhl systems in *Pseudomonas*. These systems are involved in the biosynthesis of Psl polysaccharide and rhamnolipids, which are both EPS components.

Cellulose and curli are components of EPS regulated by CsgD, the master biofilm transcriptional regulator from the LuxR protein family. CsgD can be activated by an array of environmental conditions, either directly or indirectly by Sigma-38, the main initiation factor involved in stationary phase microbial growth[109]. This pathway was identified in *Pseudomonas*, *Acinetobacter*, and *Aeromonas* in AGS.

Glycogen biosynthesis and poly-*N*-acetyl-glucosamine biosynthesis, a component of EPS[110], are activated through the BarA/UvrY/CsrA pathway. BarA/UvrY is a TCS; BarA is a sensor transmembrane kinase that is activated by peroxide[111], formate, or acetate[112]. Autophosphorylated BarA then transphosphorylates UvrY[113], which then activates the transcription of csrB sRNA (small regulatory RNA). These sRNAs inactivate CsrA through sequestration. CsrA is a repressor RNA-binding protein that inhibits translation of mRNAs involved in glycogen and EPS biosynthesis and potentially quorum sensing[114]. When CsrA is inactive, translation of these various mRNAs can occur, resulting

in EPS production. The presence of all the necessary proteins for the inactivation of CsrA are identified in the KEGG database, indicating successful biofilm formation through this pathway. The taxa in this study linked to this pathway include *Pseudomonas*, *Acinetobacter*, *Arcobacter*, *Aeromonas*, and Campylobacterales.

Motility is downregulated during biofilm formation, which is primarily regulated by levels of cyclic-di-GMP (c-di-GMP). Many regulatory proteins in the c-di-GAMP pathway are not present in AGS, suggesting poor activation of flagellar assembly.

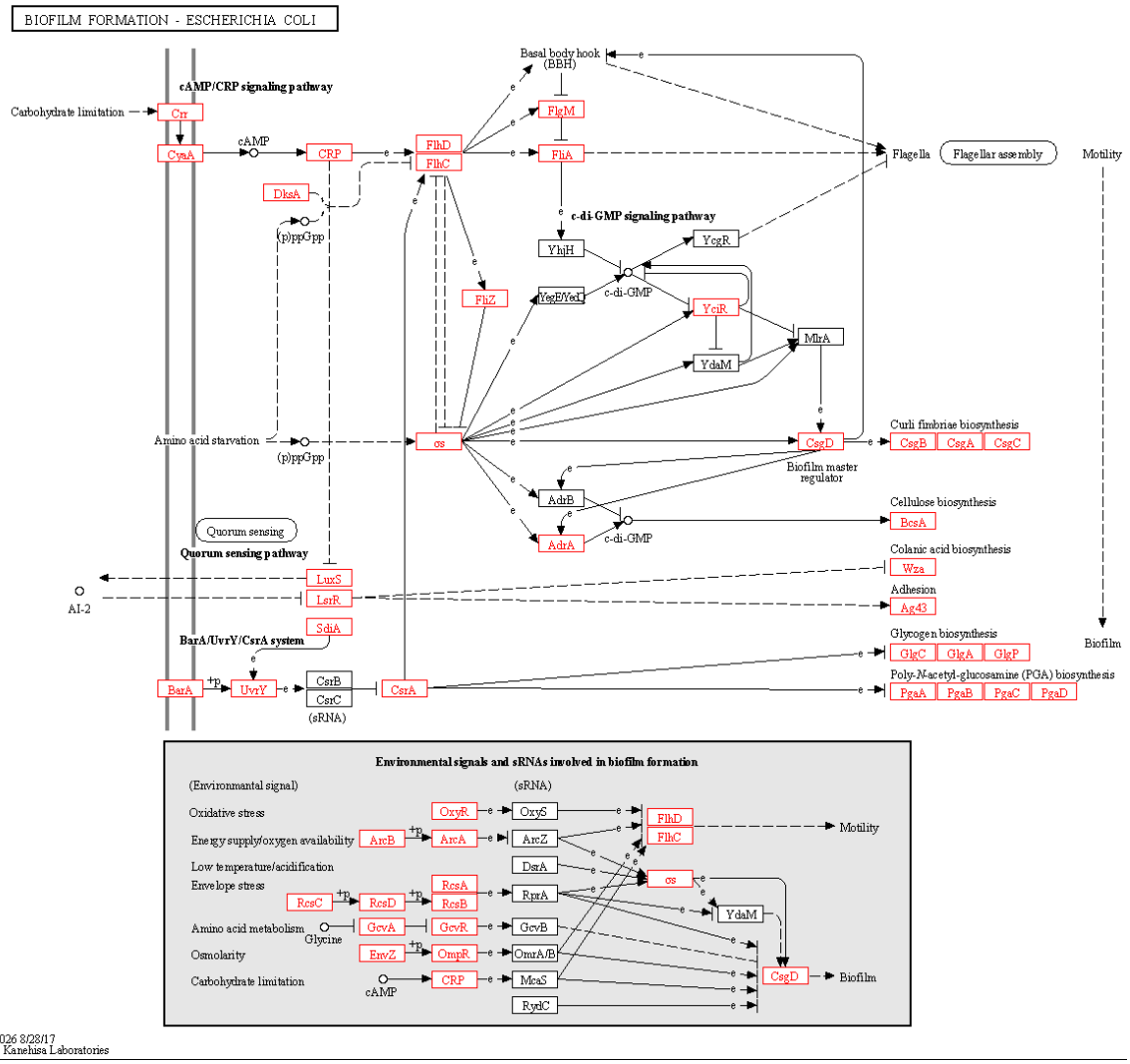


Figure 17. The KEGG reference pathway of biofilm formation (modeled after *E. coli)* in AGS. Activity at the top of the figure is associated with motility, and the bottom portion is associated with biofilm formation. Additional environmental signals that can activate CsgD are shown in the grey box. Proteins in the pathway are labeled in boxes unless otherwise indicated; non-proteins are illustrated with bullets. Proteins highlighted in red indicate a match in the KEGG database; proteins in black do not have a match.

4.4.2 Metabolic functions

Other functions upregulated in AGS include nitrogen and sulfur metabolism, particularly sulfate reduction. Denitrifiers and sulfate-reducing bacteria are typically found in the anoxic region and anaerobic core of aerobic granules (respectively) due to their anaerobic activity (Figure 20). Because of these anoxic/anaerobic regions, AGS provides an adequate environment for the proliferation of denitrifying and sulfate-reducing bacteria, resulting in higher denitrification and sulfate reduction compared to the flocs in AS[27,115]. The lack of compact structure in AS does not provide this environment in looser flocs. For nitrogen metabolism, this means only aerobic nitrifiers survive in AS conditions. One of the important advantages of AGS is simultaneous nitrification-denitrification, which is facilitated by the presence of the anoxic region. Through analysis of the predicted KEGG pathways, proteins for the full pathways of denitrification and nitrification were identified in the 8-chambered PFRs (Figure 18). The efficient removal of nitrates, ammonia, and sulfates is an important advantage of AGS over AS, so the upregulation of these pathways is a notable finding. The taxa that contribute the most substantially to these functions are *Hydrogenophaga* and *Sphaerotilus* (nitrogen metabolism) and *Pseudomonas* (sulfur metabolism).

Significant upregulation of oxidative phosphorylation, carbon fixation, and other energy metabolic functions was identified in the AS samples. These trends were somewhat unexpected given the superior COD removal rates observed in the AGS samples[18] (Figure 7d). This can be partially explained by the anaerobic metabolic ability of AGS, whereas

nearly all energy metabolism in flocs can be attributed to aerobic respiration. An additional explanation for the downregulation of metabolism in AGS is due to induction of famine periods, which reduce nutrient availability for metabolism. However, nutrients acquired during the feast phase are easily absorbed into the granule structure (Figure 12), providing nutrient sources through the famine period[27]. A more probable explanation is related to the growth rates between the bacteria in AGS and AS. Mature biofilms approach stationary phases of growth characterized by growth rates that are equivalent to decay rates (Figure 19). These growth restrictions do not apply to planktonic bacteria. This hypothesis is supported by the upregulation of replication proteins in AS, such as DNA repair/recombination proteins, translation factors, aminoacyl-tRNA synthetases, chaperones, and transcription machinery. Most importantly, ribosome expression and biosynthesis are upregulated in AS. Ribosomes are so strongly correlated to microbial growth that ribosome counts are commonly used in in situ studies to estimate growth rates of bacteria[116].

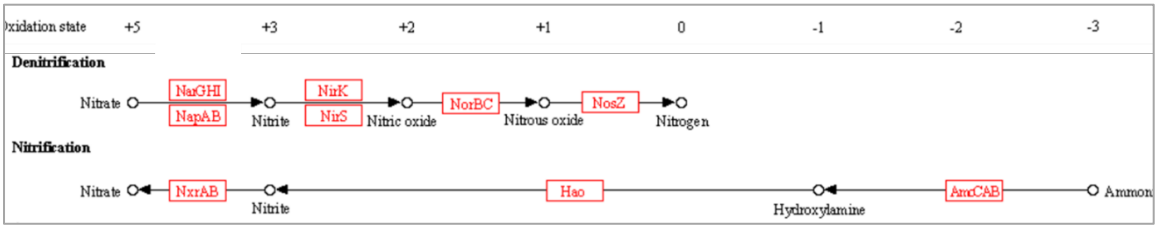


Figure 18. Schematic of the bacterial mechanism of nitrogen metabolism in AGS, as provided by KEGG[117]. Proteins in the pathway are labeled in boxes, and non-proteins are illustrated with bullets. Proteins highlighted in red indicate a match in the KEGG database.

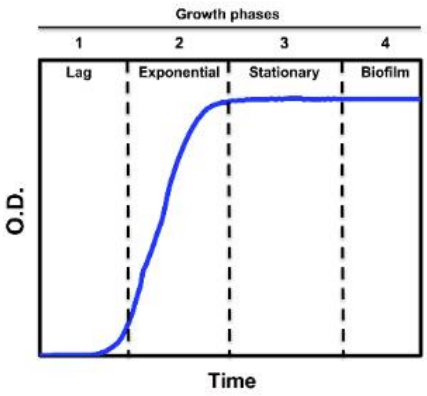


Figure 19. An example growth curve for bacteria over the course of biofilm establishment and maturation. Growth is measured by the ocular density (OD) at 600 nm. Absorbance at this wavelength captures light scattering associated with microbial growth in liquid media[118].

4.5 Future considerations

There is a strong relationship between EPS production and cell hydrophobicity. Because EPS production is crucial to granule formation, analyzing the cell hydrophobicity

across the different CSTR chambers of the PFR system would be an important validation technique to confirm EPS findings presented in this study.

There lacks a consensus in the literature about the validity of using the protein to polysaccharide (PN/PS) ratio in EPS as a way to predict sludge settleability. I propose that the controversy in the literature is the product of different compositions of protein present in EPS. Hydrophobic amino acid side chains can increase cell hydrophobicity, while hydrophilic amino acid side chains are found to be destabilizing to granular structures. The ratio of hydrophobic to hydrophilic amino acids can vary substantially with different environmental factors, and thus there is a vast assumption made in concluding that high PN/PS content is inherently destabilizing to aerobic granules. More accurate conclusions on the relative hydrophilicity can be achieved through the full amino acid profile in EPS protein. Hence, these results would be more informative than the total PN content.

Many AGS studies sample biomass throughout the aerobic granulation process. In this study, samples were taken only at steady state across the various CSTRs. While granule maturation occurs to an extent in this system, recycled granules in activated sludge at steady state are further along in the maturation process and represent a different community phenotype than maturing granules. To differentiate mature granules from maturing granules in AS which coexist in a steady state system, additional sampling throughout the maturation process may further elucidate community and motility functions needed to complete successful biofilm formation.

The composition of the microbial community is not homogenous within an aerobic granule. On the contrary, there is distinct stratification of microbial groups in a granule

based on function and metabolic ability. The arrangement of bacteria that make up the core and outer layers of the granule are largely determined by conditions such as shear force and carbon source (Figure 20). For example, the outer layer that has optimal accessibility to oxygen is dominated by aerobic bacteria responsible for nitrification and organic materials processing. The core is made up of anaerobic bacteria responsible for the majority of denitrification activity. In between these layers, hypoxic and anoxic groups associated with phosphorus removal are common. During sample preparation, the granules (or flocs) were thoroughly homogenized in order to ensure that the bacteria from all portions of the granular structure are equally represented during sequencing. However, sequencing different portions of the particulates separately would yield more insightful information on the dynamic function of the microbial community throughout the granulation process. Additionally, DNA from bacterial groups enveloped in EPS may not be adequately extracted during homogenization of the material. This could lead to underrepresentation of EPS-producing bacteria present in the granule structure. Substantial compositional variation amongst the triplicate samples that were sequenced (Figure 21) affirm that this attempt at homogenization was unsuccessful.

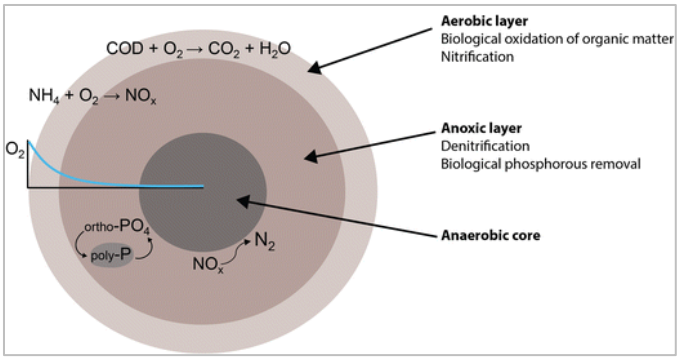


Figure 20. Schematic drawing of the stratifications in an aerobic granule. The aerobic, anoxic, and anaerobic layers are listed along with their associated functions in wastewater treatment.

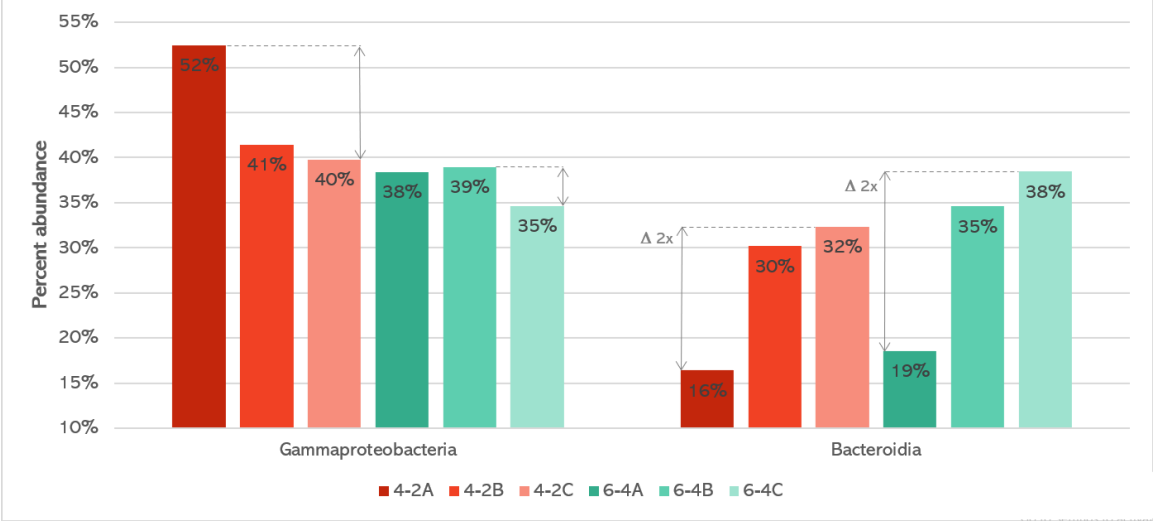


Figure 21. Percent abundance plots of triplicate samples from the 4-chambered PFR (4-2A to 4-2C) and the 6-chambered PFR (6-4A to 6-4C) for Gammaproteobacteria and Bacteroidia. The ranges for each triplicate dataset are provided.

The high abundance of *Sphaerotilus* in AGS may foreshadow sludge bulking issues under the current operational parameters of the PFR system. The starting AS abundance of *Sphaerotilus* was 0.10% but was found to be upwards of 20% in the feasting phase of the 8-chambered PFR. This indicates that the feasting conditions may apply a selective pressure encouraging the overgrowth of *Sphaerotilus*. High organic loading rate with readily biodegradable compounds like acetate and glucose have been linked to overgrowth of filamentous groups including *Sphaerotilus*. Alterations to these parameters could inhibit filamentous overgrowth and promote stable granule formation.

Quality assessment studies have reported that the largest influence on variation in microbiome profiling is the choice in primers[119]. The 341F and 785R primers used in this study were selected due to the high coverage reported in studies on the human microbiome[49,120]. However, it is possible that these findings are not accurate representations of the microbial community in wastewater. One study relating soil, plant, animal, and human microbiomes found that that the best primer pair was 515F/806R, yielding the largest diversity coverage, species richness, and number of ASVs[121]. Studies comparing the best primer pair for wastewater treatment plant samples are largely in disagreement; some suggest 27F/1492R, others suggest 341F/534R and 968F/1401R[120,122]. Most metagenomic studies on the microbial communities in AS use the hypervariable V3-

V4 region for amplification, although the specific primer pairs differ[74,77,80,123–128]. Others use different regions altogether[44,54,92]. The lack of a standard set or even region of primer pairs for amplification is potentially introducing biases across different studies that may render trends between communities unreliable.

In addition to selecting for 16S bacterial rRNA, it has been previously demonstrated that other non-prokaryotic organisms are integral to AGS. For example, differential abundances of archaeal groups are found in AS and AGS; Euryachaeota is strongly correlated to AGS, and Methanosaeta is predominant in AS[129]. Additionally, microorganisms such as tardigrades, ciliates, and rotifers are indicative of a healthy ecosystem in AGS. Tardigrades and rotifers in particular are known to feed on planktonic bacteria, while walking ciliate are known to feed on dead bacteria on the surface of aerobic granules[27]. Thus, these non-bacterial microorganisms also play an important role in the success of aerobic granulation, and analyzing their prevalence in novel wastewater treatment reactors may detail an additional layer of complexity of successful AGS.

5. CONCLUSION

The results of this study suggest that the complex community structure of the plug-flow reactor largely agrees with findings of other mature aerobic granular sludge systems under the correct combination of feast and famine conditions. Furthermore, the critical limit of a 0.5 feast/famine cycle has been reaffirmed in this study, accompanied by novel physical, metabolic, and metagenomic results that reflect an intermediate, transitional granular sludge phase in the 1:2 feast/famine PFR (6 chambers). As expected, aerobic granulation failed in the 4-chambered PFR with a 1.0 feast/famine ratio. Metagenomic results provide a microbiological explanation for the failure of this system with heavy similarities in the taxonomic groups of this PFR and AS samples.

By settling and standards, successful granulation was observed in the 8-chamberd PFR. This finding in itself is the first of its kind, as successful granulation is only consistently achieved in sequence batch reactors. The novel metagenomic analysis conducted on resultant AGS identify EPS-producing groups such as *Flavobacterium, Hydrogenophaga, Pseudomonas,* and others. Moreover, there was a substantial decrease in representation of filamentous groups such as Actinobacteria, Chloroflexi, *Microthrix,* and *Zoogloea*. Predictive functional analysis further reveals the upregulation of biofilm formation pathways and proteins related to intercellular communication in the 8-chambered PFR sample, suggesting that the 0.33 feast/famine ratio implemented in the PFR provided substantial nutritional deprivation to activate aerobic granulation.

APPENDIX A

There is a plethora of open-source bioinformatic tools available for the type of metagenomic analysis needed to complete this study. The purpose of this appendix is to provide the reader with the rationale for each tool and function used for the metagenomic analysis of data in this study. The tools needed for this microbial community study range from simple data pre-processing packages to more complex algorithms that infer amplicon sequence variants (ASVs), find chimeric sequences, and conduct taxonomic identification. These include software packages and programs such as Bioconductor, QIIME, UPARSE, MED, and more. Web servers such as MG-RAST and PATRIC have also emerged in recent years in order to provide a user interface that make metagenomic analysis more accessible for scientists with limited programming and/or cloud computing experience[130,131]. No single tool or web server is a superior option above the others, and the algorithms in these tools are consistently optimized to address biases or limitations. Furthermore, there exists a virtually unlimited combination of parameters within even a single tool. As a result, metagenomic analysis is a largely decentralized process that the researcher necessarily tailors to their type of data and expertise.

A.1 Data preprocessing

One of the few domains that remains relatively conserved among the different bioinformatics tools is data preprocessing. The goal of this step is to determine the quality of raw FASTQ reads and prime the reads for downstream analysis. Due to the large amount of data generated during sequencing, preprocessing is imperative to make information

more manageable. Quality tools include primer trimming, quality trimming, filtering, dereplication, and merging. These steps are well-defined and simple, and there are minimal differences between the tools belonging to different packages. However, the different tools do require the researcher to establish the desired parameters, and the different values and thresholds employed have the potential to influence results.

A.1.1 Primer trimming

Primer trimming is an important first step in quality management of FASTQ files, as primers account for 10% or more of base reads. Primer trimming tools can utilize a specific or non-specific methodology, depending on the researcher's preference. Specific trimming is the preferred method of primer trimming and appears in more published workflows than non-specific trimming. The advantage of specific trimming is that it simultaneously accounts for quality and discards reads that have a high likelihood of returning unreliable results. However, nucleotide insertions and deletions (collectively referred as indels) of primer sequences can occur during amplification and sequencing, and specific trimming techniques do not account for this. Reads that experience frameshifts due to insertions or deletions are automatically discarded in the specific trimming method on the assumption that the rest of the sequence is unreliable, which is not necessarily a sound assumption to make.

Specific trimming

Specific trimming protocols require the primer sequence used during amplification to be specified (both forward and reverse primers). The trimming tool will then search through the FASTQ sequences and excise the desired primer sequence from the forward or

reverse reads. During amplification, it is not uncommon for point mutations (also known as single nucleotide substitution errors[132]) to occur in the extension phase. As an example, *Taq* DNA polymerase I has a reported error rate between 8.0×10^{-5} to 7.2×10^{-5} errors per base per doubling[133–135]. Additionally, substitution errors are known to be the dominant error type observed during Illumina next-generation sequencing (NGS)[136]. The substitution error rate of Illumina NGS is 0.24% per base. To account for the propensity of substitution errors during amplification and sequencing, the primer trimming tool utilizes an error threshold. Error thresholds will automatically discard reads that have primer mismatches above a designated percentage. The standard threshold is 10%. This means that for a primer sequence that is 40 nucleotides (nts) in length, the trimming tool will tolerate up to four bases that do not match the specified primer sequence. Primer sequences with a mismatch of five or more bases will be discarded from the FASTQ file, and thus removed from further downstream analysis.

High-quality sequences might not need as low of a threshold as the 10% default. These data can afford to increase the threshold value without losing a substantial number of reads. As such, the threshold set can be altered to the researcher's discretion depending on the quality of the reads and the subsequent number of sequences that are eliminated from the FASTQ file.

Specific trimming is advantageous because primers are one of the few controls in the reads. If the primer is inaccurate, it may be indicative of a poor read. Thus, the quality of the primer can serve as a quality filter for reads that demonstrate a high probability of error.

Non-specific trimming

Non-specific trimming does not require the primer sequence as input. This method of trimming only accounts for the nucleotide length of the primer. The number of specified bases will then be indiscriminately excised from all reads. For example, if the primer length is specified as 40 bases, all reads contained in the FASTQ file will have the first 40 bases removed. Because there is no threshold requirement, no reads are discarded at this stage.

The advantage of this method is that it accounts for oligonucleotide insertions and deletions encountered during amplification and sequencing[135]. While it is significantly more common to encounter substitution errors than insertions and deletions, the specific trimming method does not account for the presence of insertions or deletions and tolerates these reads very poorly. The addition of one or more nucleotides in the primer sequence effectively leads to a frameshift mutation and subsequently shifts the primer sequence. With the specific trimming method, these reads will fail the tolerance filter, as all bases following the position of the frameshift will be treated as mismatches. Specific trimming tools assume that these primer errors reflect poor quality in the rest of the read and therefore are discarded. This assumption is not universally applicable, as the rest of the read still reports the quality scores for each nucleotide position, and these quality scores can still be high. In such a scenario, the single frameshift is not representative of an entirely poor-quality read and therefore should not be discarded from consideration.

Because there is no threshold used in non-specific trimming, all reads are retained when using this method, including low quality reads. Excision of these reads is saved for the quality evaluation and quality trimming steps. This method prevents the bias against

reads with insertions or deletions that would otherwise be discarded using specific trimming.

The extent to which primers are subject to oligonucleotide insertions or deletions is somewhat dependent on the type NGS technology used. 454 pyrosequencing technology can be incredibly prone to indels; indel errors can range from 0.02% to nearly 50% per base[132]. Other technologies like Illumina NGS experience indel errors ranging from 2.8 x 10^{-6} to 5.1 x 10^{-6}, which is several of orders of magnitude less common than the error rate observed by 454 sequencing technology[137].

Amplification does not contribute largely to insertion or deletion errors. In fact, the errors experienced when using Taq polymerase I are predominantly substitutions by frequency (98.8%), followed by minimal deletions (1.2%), and no reported insertions[135].

In summary, the choice to use specific or non-specific primer trimming should be considered on a case-by-case basis depending on the proclivity for indel errors compared to point substitution errors. If Illumina NGS is used, the overwhelming majority of errors in the reads are substitution errors to the point that indel errors are negligible. Discarding these sequences with indels when they fail the threshold is not likely to significantly alter the outcome of the metagenomic data, as it represents around 1% of all total errors. Reads that fail the threshold in this case largely agree with the assumption that rest of the sequence is likely unreliable and should be discarded. If other forms of sequencing that are more likely to experience indel errors (such as 4454, Ion Torrent, or Pacific BioSciences), non-specific trimming should be considered instead.

Specific trimming: PANDAseq vs. Cutadapt

Because Illumina NGS is used for metagenomic sequencing in this study, specific trimming was chosen as the method for primer trimming. There are two popular trimming tools, PANDAseq and Cutadapt, used to complete this data preprocessing step.

PANDAseq in particular has attracted significant attention from bioinformaticians, with over a thousand of citations since its release in 2012[138]. Despite this popularity, the outputs from this trimming tool were substandard. Some functions did not yield the expected results, and thus it was not considered a reliable tool for metagenomics processing.

Many modifications can be made to the trimming function in PANDAseq to alter the number of accepted reads. One option is to merge the reads (also referred to as "assembly" or "alignment" in PANDAseq workflows) in addition to primer trimming. The default for PANDAseq is to trim the primers and then assemble the forward and reverse reads. This default command was tested with a 0.6 accuracy threshold, meaning that base errors of up to 40% in the primer sequence would be tolerated. Only 200 reads were discarded when using this pipeline (less than 0.2% of all reads). To trim after assembly, the `-a` command can also be specified. For reasons unknown, using the `-a` command increases the selectivity of the reads that are accepted even when the same 0.6 threshold was applied. 7000 reads were discarded, accounting for 5% of all reads in the FASTQ file.

The tool selected for primer trimming was the Python library, Cutadapt[139]. The limitations of PANDAseq were not apparent when using Cutadapt. The default error rate employed by Cutadapt is 10%. For our 17 and 21 base pair primers (forward and reverse,

respectively), this comes out to a tolerance of 1-2 base errors per read. A very low number of reads were discarded with this default value (less than 0.1% of all reads), so the tolerance threshold was increased to only allow a single base error in the primer sequence. To achieve this, the threshold was set to 5% and 6% for the forward and reverse FASTQ files, respectively. Only 1000 reads were discarded due to high error rates (1.1% of total reads).

A.1.2 Quality evaluation

Q scores

One of the benefits of the newer FASTQ format over the previously used FASTA format is that quality information is automatically included for each base sequenced. Quality information is formatted as Phred quality scores, also known as Q scores. Q scores logarithmically represent the probability, P, that a base is incorrectly reported during sequencing[140,141].

$$P = 10^{-Q/10}$$

(Equation 1)

$$Q = -10\log_{10}(P)$$

(Equation 2)

For example, a Q score of 30 has a probabilistic error of 0.001 ($10^{-30/10}$), or a 99.9% probabilistic accuracy. Q scores are displayed as ASCII (base 33) characters in FASTQ files and can be readily converted to the numerical Q score (Figure 22). The string of ASCII characters as they would appear in a FASTQ file are shown in Figure 23[142].

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Figure 22. The relationship between Q scores, probability, and ASCII characters[143]. FASTQ files report quality scores via their ASCII equivalents.

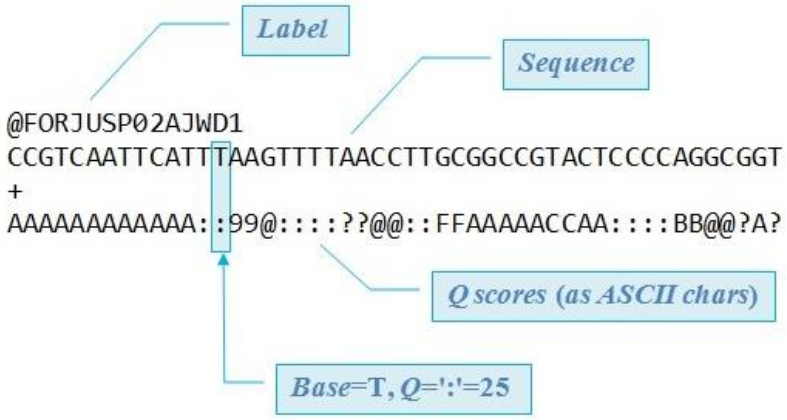


Figure 23. Example of a read in FASTQ format. ASCII-transcribed Q scores are found underneath the read, separated by a plus symbol. Q scores and bases are in line such that the nucleotide is directly above the ASCII character that reports its quality.

Phasing

Due to the limitations of Illumina NGS technology, there is typically a drop in quality towards the end of a read. This is due to a process called phasing[144]. During sequencing, DNA sequences (such as amplicons) are denatured to single-stranded molecules, exposing the base sequence of the strand. This exposed template strand is then washed with incoming nucleotides that are tagged with fluorescent signals and a terminator cap[145]. Fluorescent signals differentially bind to a specific nucleotide which allow for the specific detection of one of the four bases. The terminator cap prevents the addition of additional nucleotides, ensuring that only one fluorescent signal is measured per position (or cycle). In order to initiate the next cycle, this cap must be removed. Occasionally, the cap is not properly removed, causing the sequencing equipment to erroneously capture the previous fluorescent signal (n-1) a second time. Alternatively, an extra nucleotide can bind to the template strand (n+1) in the case that a cap is defective, causing one of the nucleotides to be skipped in the sequence. The former error is referred to as being out of phase, and the latter pre-phasing. Both out of phase and pre-phasing are phasing errors that can occur during sequencing. With a higher number of cycles (longer read length), phasing become more likely due to the cumulative nature of the errors[144]. Thus, the fluorescent signals received become more asynchronous, are less likely to be accurate readings, and generate lower quality scores towards the end of an amplicon read.

Tools for quality visualization

Q scores can be visualized with different tools such as DADA2 (from Bioconductor), FastQC, and VSEARCH (from QIIME1)[146–148]. These tools present the visuals in different ways, but the overall purpose is the same: to assess the quality of each base as a function of the read's length. Because of phasing, it is important to analyze the extent to which quality scores fall with respect to read length. A Q score above 20 is commonly recognized as "acceptable", and a Q score above 28-30 is commonly considered a high quality read[146,149]. In general, the drop is more drastic in reverse reads than in forward, as seen in the DADA2 visualizations in Figure 23. It is not uncommon to find quality scores below 20 at the end of reads, especially those with longer amplicons and in reverse reads. The presence of Q scores as low as 12 (error probability of 6.3%) can be observed in DADA2's heatmap (Figure 24b), further highlighting some of the shortcomings of NGS that must be accounted for during data preprocessing.

VSEARCH statistics

Programs such as VSEARCH analyze the Q score distributions numerically, which provide more insightful and quantitative results of the reads (*Tables 1-2*). There are a variety of simpler statistics, such as the number of bases associated with each Q score (N), the percent of bases with a given Q score (Pct), and the percent of bases that have a given Q score or higher (AccPct, for accumulated Q score) (Table 4). Through these statistics, we can infer that the 4-chambered reactor data being analyzed is very high quality. Over 96% of bases have a quality score of 32 or higher, which corresponds to a predicted error rate of 0.063%. Only 2.00% of all reads are below a quality score of 27 (AccPct), which is

considered the lower bound of a high quality read. This finding suggests that the overwhelming majority of bases sequenced are of high quality. However, these statistics do not provide insight for the quality of the reads with respect to the read length.

There is also an array of more complex statistical values: the percent of all reads with at least length "L" (PctRecs), average Q score in a given range of read lengths (AvgQ), probabilistic error rates (P(AvgQ)), average expected error (AvgEE), and the expected error growth rate (RatePct), as approximated from the AvgEE between position L and position L-1 (

Table 5). These statistics provide more extensive analysis on quality and sequence position. As the length of the read increases, the percent of reads that meet the length requirement falls. This is expected because not all reads will be a full 250 bases due to phasing or other NGS errors. However, 99.7% of bases are at least 248 bases in length. Of the reads that are at least 248 nts in length, the average quality score is a 37.2. Average quality scores (AvgQ) are not reliable to gauge the quality of a read because low quality reads are easily masked by an abundance of high-quality reads[150,151]:

Number of reads x Q score	AvgQ	AvgEE
(120 reads x Q37.2) + (30 reads x Q2)	30.2	19.0
150 reads x Q25	25	0.5

AvgEE is calculated as the sum of error probabilities, P[152].

As shown, a dataset may have a higher average quality score, but the true quality of the data is very poor: 19 expected errors per 150 bases. Because of this, it is essential to consider other statistics (like the expected error, AvgEE) that better represent the presence of error, particularly with respect to the read's length. A clear decline in quality can be observed in

Table 5 through the changes to probabilistic error rates, average expected error rates, and the expected error growth rate (P(AvgQ), AvgEE, and RatePct, respectively). Fortunately, these changes are insignificant. Even at length of 250 nts, the expected error (AvgEE) is 0.27. An AvgEE value of one means that one base is expected to be misreported in the entire read. Thus, a value below one means that the most probably number of errors is zero. It can be concluded from the dataset that the quality through the end of the read is very reliable and does not need quality trimming (Section A.1.3 Quality trimming).

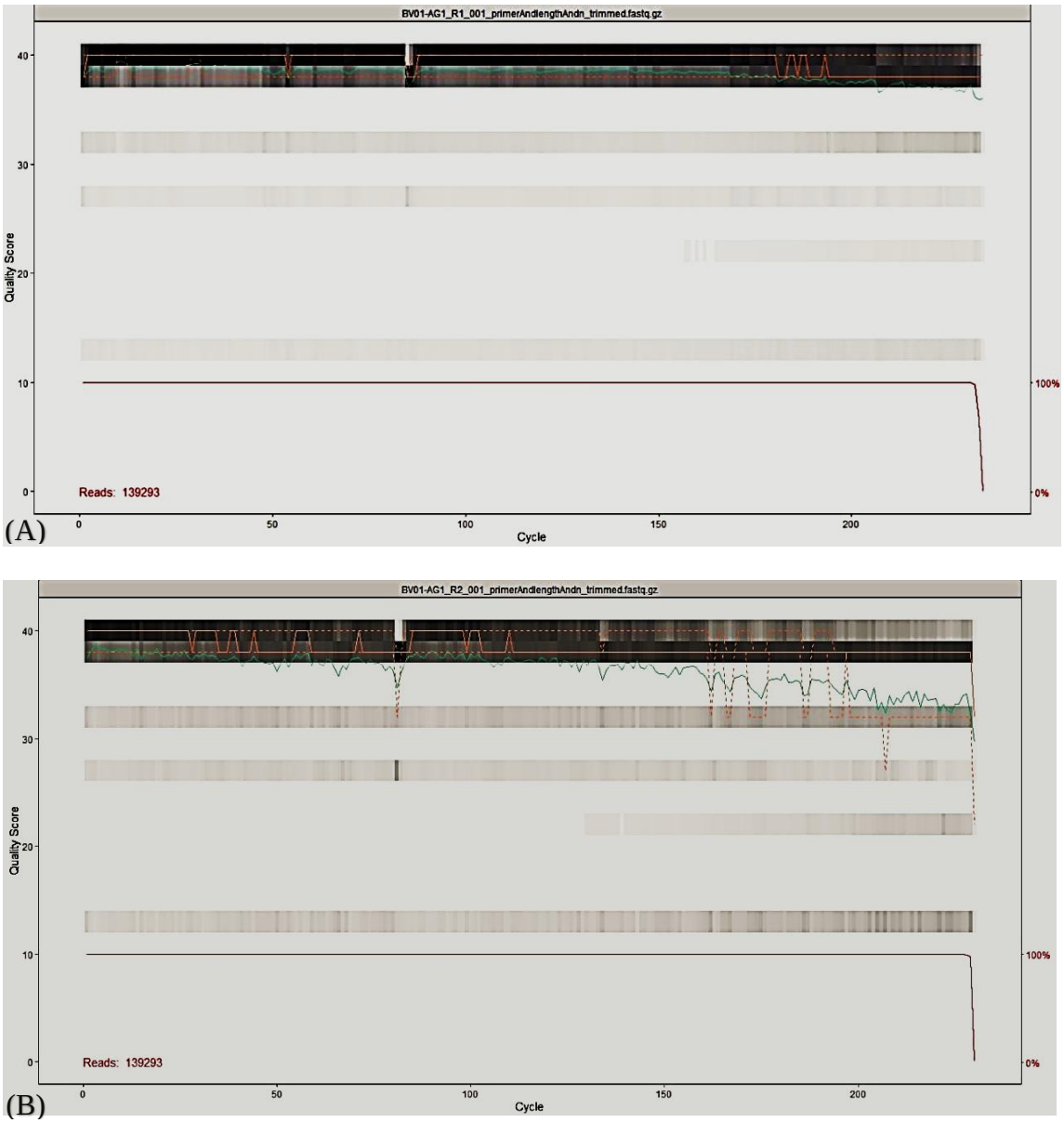


Figure 24. Quality score distribution plots generated by the DADA2 visualization tool, `plotQualityProfile`. Quality scores are plotted a function of the read length ("cycle"). Input data comes from forward (a) and reverse (b) FASTQ files from an 8-chambered reactor. Green lines represent the median quality scores for each base position; orange lines represent the quartiles. The black heatmaps represent the frequency of the

quality scores per base position. The forward reads are almost entirely contained in the 36-40 Q score range (a), whereas the reverse reads experience a steeper decline near cycle number 150, approaching median quality scores as low as 30.

Table 4. Simple quality distribution statistics provided by VSEARCH. Input data used for this example comes from a 4-chambered reactor system. The highest number of reads "N" are associated with the highest Q scores (32-40), making up 96.60% of all reads. "ASCII" corresponds to the base-33 conversion character for the Q score "Q"; "Pe" is the corresponding probability of error in the read for that given quality score.

ASCII	Q	Pe	N	Pct	AccPct
I	40	0.010%	20547955	58.50%	58.50%
G	38	0.016%	11938085	34.00%	92.50%
A	32	0.063%	1446491	4.10%	96.60%
<	27	0.200%	476386	1.40%	98.00%
7	22	0.631%	76573	0.20%	98.20%
.	13	5.012%	641178	1.80%	100.00%
#	2	63.096%	137	0.00%	100.00%

Table 5. Detailed VSEARCH statistical outputs for quality evaluation. Input data used for this example comes from a 4-chambered reactor system sequenced using 250 nt amplicon NGS. “L” represents the length of the read. As the value of L increases, the subsequent quality metrics show increased error probability: average quality score (AvgQ) drops with read length; probabilistic error rates, average expected error rates, and the expected error growth rate (P(AvgQ), AvgEE, and RatePct, respectively) all increase with read length.

L	PctRecs	AvgQ	P(AvgQ)	AvgEE	RatePct
2	100.0%	36.7	0.00021	0	0.093%
3	100.0%	36.9	0.00020	0	0.083%
4	100.0%	36.8	0.00021	0	0.082%
5	100.0%	37.1	0.00020	0	0.078%
6	100.0%	39.0	0.00013	0	0.074%
7	100.0%	38.6	0.00014	0.01	0.080%
8	100.0%	38.6	0.00014	0.01	0.080%
9	100.0%	38.6	0.00014	0.01	0.080%
10	100.0%	38.9	0.00013	0.01	0.077%
11	100.0%	39.0	0.00013	0.01	0.075%
12	100.0%	39.1	0.00012	0.01	0.073%
240	99.8%	36.7	0.00021	0.25	0.105%
241	99.8%	37.0	0.00020	0.25	0.105%
242	99.8%	36.8	0.00021	0.26	0.106%
243	99.8%	37.0	0.00020	0.26	0.106%
244	99.8%	36.7	0.00022	0.26	0.107%
245	99.8%	36.7	0.00021	0.26	0.107%
246	99.8%	36.6	0.00022	0.27	0.108%
247	99.8%	37.0	0.00020	0.27	0.108%
248	99.7%	37.2	0.00019	0.27	0.108%
249	97.5%	36.1	0.00024	0.27	0.109%
250	65.9%	35.8	0.00026	0.27	0.109%

A.1.3 Quality trimming

After reviewing the quality of the reads with the tools described in A.1.2 Quality evaluation, reads should be trimmed to remove the lowest quality bases at the end of a read. The amount trimmed will depend on the quality of the data, the size of the amplicons, and the length of the overlap between the forward and reverse reads. For example, an amplicon with 341F and 785R primers will have a sequence 444 nts in length in between the primers (shown in black in Figure 25). However, because the primers are also included during sequencing (17 and 21 bp in length for the forward and reverse primer, respectively), the true coverage is substantially larger: 482 bases. For amplicon sequencing of a coverage of 482 nts, 2 x 250 base pair (bp) amplicon would be selected for forward and reverse readings (as shown by the orange and green arrows, respectively). This yields a theoretical coverage of 500 nts, which is an overlap of 18 bases more than is required for full coverage of the sequence (shown in blue in Figure 25). Some overlap is needed in order to merge forward and reverse reads (discussed in Section A.1.6 Merging), but a large overlap allows for low-quality bases at the end of the read to be trimmed if needed without detriment to downstream steps such as merging. This example provides a moderate overlap where it would be acceptable to trim up to 6 bases total.

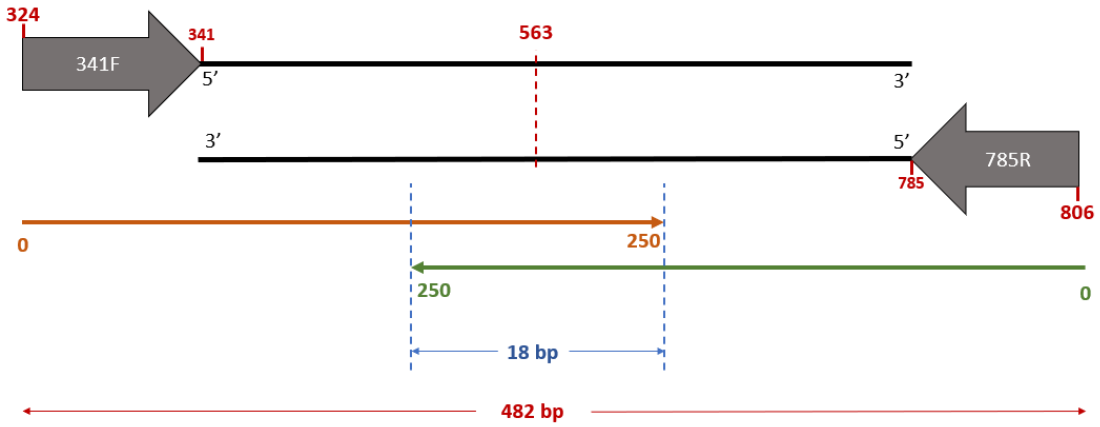


Figure 25. An example diagram of the overlap obtained during amplicon sequencing. When using 341F and 785R primers (shown in gray), the total length of the amplicon, including the primers, is 482 bps (shown in red). Thus, 250 bp sequencing in each direction (orange and green arrows) is utilized to achieve proper coverage of the amplicon. This yields an 18 bp overlap (shown in blue). Position 563 is the centermost base.

A.1.4 Filtering reads

After removing primers and quality trimming reads, short sequences can be removed through a processing called filtering. The purpose of this step is to prepare reads for merging. By excising reads that are far too short to be merged, the merging algorithm can be executed more quickly and with higher accuracy. This can be achieved using `Cutadapt`.

Minimum and maximum read lengths are proportional to the average length of the read, which has likely decreased during primer trimming and quality trimming. The default in `Cutadapt` is 10%, signifying that the lower boundary of accepted reads is 10% below

the average length, and the upper bound of accepted reads is 10% above the average length. Because it is known that the overwhelming majority of samples are either 249 or 250 bases long (97.5% and 65.9% of reads, respectively, as seen in

Table 5), the average length is between 249 and 250. Below is an example for the boundary determination of a forward read with a length of 17 bps:

(Read length – primer length) * 10% = proportion value

$$(250\ bp - 17\ bp) * 10\% = 22\ bp$$

(Equation 3)

(Average length – primer length) – proportion value = minimum bound

$$(250\ bp - 17\ bp) - 22bp = 211\ bp$$

(Equation 4)

(Average length – primer length) + proportion value = maximum bound

$$(250\ bp - 17\ bp) + 22bp = 252\ bp$$

(Equation 5)

Thus, the maximum and minimum values to be passed through the `Cutadapt` filter tool are 252 and 211, respectively. This must be repeated for the reverse read, which likely has a smaller average length due to more extensive trimming and a different primer length.

A.1.5 Dereplication

Dereplication eliminates replicate sequences found in a FASTQ file and notates that there is a given abundance of repeats for every given sequence. The purpose of dereplication is to simplify data in a way that makes analysis less computationally intensive. Data storage is optimized if a single representative sequence is listed 100 copies rather than actually storing the same sequence 100 separate times. This step requires a 100% match between sequences and is relatively straightforward. This step can be completed through all of the major software packages such as Bioconductor and QIIME.

A.1.6 Merging

Merging marks the final step in quality processing. The tools described in *Sections 3.1* serve to improve the relative quality of the reads by discarding poor quality reads, discarding reads of insufficient length, trimming primers, and trimming read length to remove poor Q scores. Once these steps are completed with the raw data of the forward and reverse reads, they can be merged to form a single-stranded sequence. In fact, many pipelines wait to merge forward and reverse sequences until after sequence error analysis ("learning errors") and denoising functions have been called.

As mentioned in Section A.1.3 Quality trimming, there must be a proper overlap between the forward and reverse reads in order to ensure the pairs match accurately. Similar to primer trimming (Section A.1.1 Primer trimming), there is an error threshold that determines the number of bases that are permitted to mismatch during merging. The default error threshold in DADA2 is zero; this means that reads with any base pair mismatches are to be excised from the dataset for further analysis. This parameter is intentionally strict in

order to prevent mispairing between two unrelated reads, which can drastically alter downstream results such as taxonomic identifications. This default can be changed to allow for errors. If this option is selected, the merging tool will evaluate the mismatched bases and choose the correct base in the merged sequence based off of the quality score between the two bases.

Additionally, there is a length parameter that defines the minimum overlap required in order to match forward and reverse reads. The default value in DADA2 is a 12 bp overlap; the default value in the QIIME1 pipeline is 10 bp.

Quality evaluation after merging

Quality statistics can be generated following the merging step in order to determine average lengths and changes to quality. The outputs of these statistics are identical to those described in Section A.1.2 Quality evaluation.

A.2 Learning errors

Sequence error estimation is an algorithm apart of DADA2 used to predict errors introduced during PCR amplification and sequencing[147]. These errors vary drastically depending on the dataset. As a result, the developers of DADA2 designed a machine-learning algorithm that learns the errors from the data directly by alternating the estimation of error rates and inference of sample composition (discussed in Section A.3.1 Clustering). This learning process continues until a consistent solution is achieved[147,152]. This algorithm uses a parametric model to predict these errors, as shown in Figure 26. Estimating error parameters from an individual's set of data results in sample inference results that are more sensitive and specific than any other clustering algorithm[147].

Error rate estimation is executed with the `learnErrors` function in DADA2. The only inputs for this function are the output files from quality preprocessing discussed in Section A.1 Data preprocessing. There are no parameters that need to be specified with this function.

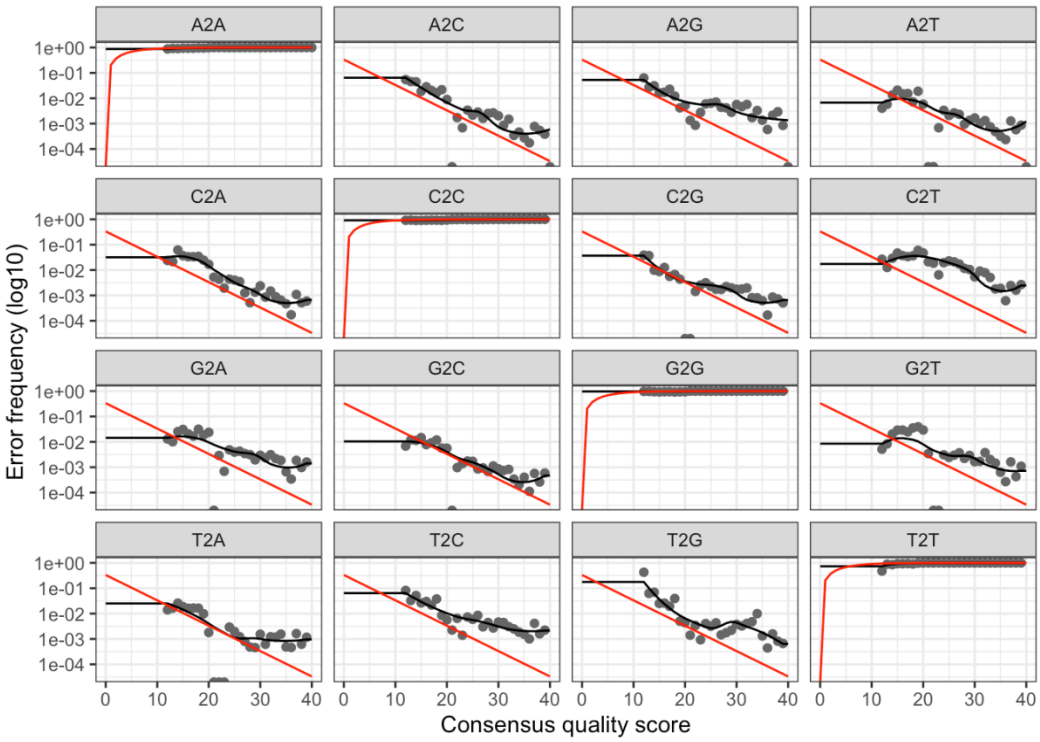


Figure 26. A visualization of estimated error rates as determined by the learnErrors function in DADA2. Points in black represent the observed error rates for each consensus quality score. The black line represents the estimated parametric model of the error rates as was determined by the machine-learning algorithm. The red line represents the expected

relationship between error rate and Q score for comparison between the expected error rate (red) and learned error rate (black)[147].

A.3: Sample inference: Denoising or clustering

A.3.1 Clustering

Leading up to sample inference, samples have been passed through many different tools and functions to reduce the probability of errors. However, Illumina-based NGS is not perfect, and assigning taxonomies to dereplicated sequences can cause sequencing errors to be erroneously interpreted as genomic variation[147]. In order to correct for this, clusters of similar sequences are grouped together in operational taxonomic units (OTUs). The threshold requirement of similarity is typically set to 97%. This grouping step is called clustering, and it is one way to execute sample inference.

The limitation of OTU clustering is its low resolution. In attempt to correct for sequence errors, clustering algorithms fail to take advantage of the fine-scale variations that can be detected by Illumina NGS[147]. These variations can provide insightful information on population structure. These higher resolution variations are compromised by conventional OTU clustering algorithms such as `uclust` (used in QIIME1), UPARSE, Mothur, and MED (minimum entropy decomposition).

A.3.2 DADA: A new tool for sample inference

DADA2 is an open-source R package that contains tools for filtering, trimming, dereplication, merging, assigning taxonomy, identifying chimeras, producing

visualizations, and more. Arguably, the most influential tool in the DADA2 package is the divisive amplicon denoising algorithm (DADA). The DADA algorithm is designed to infer composition differently than traditional clustering algorithms. Samples are analyzed by dividing amplicon reads into small segments that match the learned error rates obtained from the `learnErrors` function. Instead of a clustering method that generates OTUs, the DADA algorithm uses a denoising method which generates ASVs (actual sequence variants).

The division of amplicon reads into small segments is not a new method of sample inference. MED utilizes this method as well, and as a result, this algorithm is able to identify fine-scale variations[153]. MED and DADA2 were found to have comparable ability to identify fine-scale variation, but DADA2 generated a lower number of false positives[147]. This is significant because prior to the development of DADA2, MED was considered the algorithm with the lowest false positive rate of all published sample inference algorithms[153].

Additionally, when the DADA algorithm was tested on vaginal microbiome samples (n = 142 women), six separate and novel variants of *Lactobacillus crispatus* were identified. Vaginal microbiome samples selected for evaluation due to the relatively low level of microbial diversity compared to other body habitats[154]. Prior to this analysis, there had only been one predominant *Lactobacillus* OTU identified through metagenomic analysis[155]. This finding suggests that the DADA algorithm is more sensitive than other available clustering algorithms.

When compared to the UPARSE algorithm, DADA2 identified an equal or larger amount of reference strains in a mock dataset. DADA2 also outperformed UPARSE in specificity as observed by the number of false sequences identified by UPARSE compared to DADA2[147].

Perhaps the largest competitor of the DADA2 package is the software pipeline QIIME (quantitative insights into microbial ecology). Prior to the release of its second version, the QIIME1 platform used `uclust` as the default tool for OTU clustering[156,157]. When `uclust` and DADA2 were both tested on mock datasets, DADA2 was more specific and sensitive than QIIME, and QIIME was found to have a significantly higher propensity for identifying erroneous sequences[147]. Recent literature on QIIME2 detail the addition of new plugins that increase the options available for sample inference[158]. QIIME2 now includes improved clustering methods such as `q2-vsearch` and `q2-dbotu`, both of which are more accurate than their `uclust` predecessor[159]. Interestingly, plugins now exist for denoising methods such as DADA2 and Deblur.

A.3.3 Additional denoising packages

Comparative studies with DADA2 such as those described above typically compare the efficacy of OTU clustering methods with the one denoising algorithm, DADA. However, other denoising packages exist, such as Deblur and UNOISE3. The denoise method of sample inference is apparently superior to traditional clustering methods, but it is also important to compare the accuracy of different denoising methods against one another. A comparative study on DADA2, Deblur, and UNOISE3 concluded that DADA2 found the most ASVs of all three denoising methods when tested on mock datasets as well

as an array of real microbiome datasets[160]. It was suggested that this high resolution may lend DADA2 more vulnerable to false positive results. However, this was not reflected in the results obtained by the study and has yet to be demonstrated in the literature. Additionally, all three pipelines generated similar microbial community profiles.

At this time, DADA2 appears to be the best bioinformatics tool for sample inference due to its unmatched specificity and sensitivity in identifying sequence variants.

A.4 Removing chimeras

Chimeras are sequences that contain components from two different sequences. Incomplete amplification during PCR occurs occasionally and creates sequence fragments that act erroneously as primers for the next cycle of extension, leading to integration of two different templates during the annealing step.

The DADA algorithm is able to identify substitution, insertion, and deletion errors, but it is incapable of removing chimeras. Thus, the removal of chimeras is completed with a separate step. Algorithms such as `removeBimeraDenovo` identify chimeric sequences by attempting to reconstruct singular sequences from different segments of parent sequences. If they match, then they are excised from the dataset[147]. "De novo" methods signify that the algorithm does not reference an external database to find chimeras. "Closed" methods of chimera identification that use a database (such as ChimeraSlayer Gold[161]) also exist, but these are not consistently reliable since the process by which chimeras are generated is random.

An additional benefit of using denoising methods is that ASVs improve the accuracy by which algorithms can identify and remove chimeras from reads in comparison to OTUs generated from clustering methods.

A.5 rDNA gene databases

Reads have now been fully processed and are ready to be run against a database to be assigned taxonomies. Commonly used taxonomic classification 16S databases include SILVA, GreenGenes, and RDP (Ribosomal Database Project). Literature comparing the different databases is sparse and largely inconclusive[162].

The SILVA and RDP databases are both regularly updated to continually add additional taxonomies. The GreenGenes database, on the other hand, has not been updated since 2013[163]. The SILVA database is larger than RDP, which is advantageous from the perspective of increasing variety, but it does increase run time[162]. Additionally, SILVA sequences are quality checked, which ensures accuracy of the results obtained from taxonomic matches[164]. Recent publications tend to prefer SILVA over GreenGenes when choosing a 16S database.

A new paper published in 2020 details a reference database designed specifically for AS and influent wastewater microbiomes, called MiDAS (Microbial Database for AS)[72]. The taxonomy is based on results from 21 Danish WWTPs. It attempts to improve the taxonomic resolution for WWTP microbial communities which are lacking from 16S databases such as SILVA or GreenGenes, which only provide classifications to the genus level. It is argued that the diversity of wastewater microbial communities is integral to predicting physiological differences between microbes. Because of this, higher resolution

is needed to differentiate microbial members of the same genus. There is little comparative research conducted on this database due to the niche microbiome that it applies to. Additionally, the taxonomies have been trained exclusively on Danish municipal WWTPs. Hence, it is uncertain how applicable the results would be applied to pilot wastewater studies, or even municipal WWTPs in different countries that have different parameters for treatment.

APPENDIX B

The purpose of this appendix is to provide the reader with the code written in R employing the Bioconductor package, DADA2, as mentioned in *Section 4.3*. It also serves to explain the utilization of the AWS EC2 framework and RStudio AMI.

Launching the AWS EC2 Machine

The virtual machine created through the AWS EC2 server was instance type `c5.4xlarge`, which is a 16-core machine[165]

. When creating the virtual machine, the RStudio AMI was specified as provided by Louis Aslett[166]. The region was set to N. Virginia and then launched. The IPv4 public IP address generated then served as the web address used to access the remote RStudio. The requested username was provided by the AMI, and the password is the instance identification number as provided through AWS.

Next, FileZilla was used in order to connect local files and remote storage. Under `Site Manager`, the IPv4 public IP address was pasted to fulfill the host address. Then, any desired files stored locally can be transferred to the virtual machine storage for use in RStudio.

R Script

This script assumes previous installation of BiocManager and DADA2 packages.

```
library(BiocManager)
library(dada2)

# Organize files into R1 and R2 folders (Rstudio terminal)
mv /home/rstudio/primer_trimmed/*_R1_001_* /home/rstudio/primer_trimmed/R1
mv /home/rstudio/primer_trimmed/*_R2_001_* /home/rstudio/primer_trimmed/R2
```

```
# Tie all forward and reverse fastq files together.
pathF <- "/home/rstudio/primer_trimmed/R1"
pathR <- "/home/rstudio/primer_trimmed/R2"
filtpathF <- file.path(pathF, "filteredforward")
filtpathR <- file.path(pathR, "filteredreverse")
fastqFs <- sort(list.files(pathF, pattern="fastq.gz"))
fastqRs <- sort(list.files(pathR, pattern="fastq.gz"))

# Ensure that the number of R1 files equals R2 files.
if(length(fastqFs) != length(fastqRs)) stop("Forward and reverse files do
not match.")

# Filter reads for low quality and N nts
filterAndTrim(fwd=file.path(pathF, fastqFs), filt=file.path(filtpathF, fas
tqFs), rev=file.path(pathR, fastqRs), filt.rev=file.path(filtpathR, fastqR
s), compress=TRUE, verbose=TRUE, multithread=TRUE)

# Change file paths of filtpathF and filtpathR.
filtpathF <- "/home/rstudio/primer_trimmed/R1/filteredforward"
filtpathR <- "/home/rstudio/primer_trimmed/R2/filteredreverse"

# Ensure that filtered sample names in the filtered sequences are paired.
filtFs <- list.files(filtpathF, pattern="fastq.gz", full.names = TRUE)
filtRs <- list.files(filtpathR, pattern="fastq.gz", full.names = TRUE)
sample.names <- sapply(strsplit(basename(filtFs), "_R1"), `[`, 1)
sample.namesR <- sapply(strsplit(basename(filtRs), "_R2"), `[`, 1)
if(!identical(sample.names, sample.namesR)) stop("Forward and reverse file
s do not match.")

# Call "learnErrors" function.
names(filtFs) <- sample.names
names(filtRs) <- sample.names
set.seed(100)
errF <- learnErrors(filtFs, nbases = 1e8, multithread = TRUE)
errR <- learnErrors(filtRs, nbases = 1e8, multithread = TRUE)

# Perform dereplication and denoising on "err" objects. Use sam
to feed the output of the "derepFastq" function into the "dada" function.
mergers <- vector("list", length(sample.names))
names(mergers) <- sample.names
for(sam in sample.names) {
```

```
    cat("Processing:", sam, "\n")
    derepF <- derepFastq(filtFs[[sam]])
    ddF <- dada(derepF, err = errF, multithread = TRUE)
    derepR <- derepFastq(filtRs[[sam]])
    ddR <- dada(derepR, err = errR, multithread = TRUE)
}

# Merge "dd" and "derep" objects. Forward and reverse reads are separated.
merger <- mergePairs(ddF, derepF, ddR, derepR)
mergers[[sam]] <- merger

# Generate sequence table using “merger” object. Save R and locally.
seqtab_4and6chamber <- makeSequenceTable(mergers)
saveRDS(seqtab_4and6chamber, "home/rstudio/seqtab_4and6chamber.rds")

# Merge sequence tables from additional batches, if needed:
    # First upload any additional sequence tables to the console.
    seqtab_4and6chamber <- readRDS("home/rstudio/seqtab_4and6chamber.rds")
    seqtab_8chamber <- readRDS("home/rstudio/seqtab_8chamber.rds")

    # Merge the sequence tables from all batches. Save to R and locally.
    seqtab_468sequences <- mergeSequenceTables(seqtab_4and6chamber,
    seqtab_8chamber)
    saveRDS(seqtab_468sequences, "home/rstudio/seqtab_468sequences.rds")

# Once all sequence tables are loaded, remove chimeras.
seqtab_468sequences_nochim <- removeBimeraDenovo(seqtab_468sequences, meth
od="consensus", multithread=TRUE)

# The “_nochim” object is a sequence table used by downstream packages. Save
to R and locally.
saveRDS(seqtab_468sequences_nochim, "home/rstudio/seqtab_468sequences_noch
im_FINAL.rds")

# Assign taxonomy and save the output (to R and locally via FileZilla).
taxa_seqtab_468sequences <- assignTaxonomy(seqtab_468sequences_nochim, "ho
me/rstudio/silva_nr_v138_train_set.fa")
saveRDS(taxa_seqtab_468sequences, "home/rstudio/taxa_seqtab_468sequences.r
ds")
```

APPENDIX C

The purpose of this appendix is to provide the reader with the Python scripts used for the PICRUSt2 software as described in Section 3.4 Predictive functional analysis.

There are two input files needed to run the PICRUSt2 pipeline. The first is a FASTA file with ASV identification numbers and their corresponding sequences. The second is the tab-delimited ASV sequence table with sample names as column headers and the ASV identification number as row headers. A tab must be present as the first character in this text file. The ASV identification numbers must match exactly in both the FASTA file and the sequence table. The PICRUSt2 pipeline can be executed locally or using an AWS virtual machine. If it is used locally, it requires 16 gigabytes of RAM.

Python Script

This script assumes previous installation of PICRUSt2, which was completed using a `conda` environment in Ubuntu. The FASTA file of ASVs (`ASVs4and6.fa`) fulfills the `-s` flag, and the ASV sequence table (`ASVs_seqtable4and6.tsv`) fulfills the `-i` flag. The --stratified flag must be specified in order to generate stratified tables. The desired output file directory is established using the `-o` flag.

```
conda activate picrust2

# Import the FASTA file of ASVs.
cp /mnt/c/Users/Ali/Documents/research/ASVs4and6.fa .

# Import the ASV sequence table.
cp /mnt/c/Users/Ali/Documents/research/ASVs_seqtable4and6.tsv .

# Execute the PICRUSt2 pipeline (with stratified outputs.
picrust2_pipeline.py -s ASVs4and6.fa -i ASVs_seqtable4and6.tsv -
o picrust2_out_pipeline_with_stratified -p 1 --stratified --verbose

# Convert the stratified EC and KO files into "legacy" format that can be
read by programs like BURRITO.
convert_table.py picrust2_out_pipeline_with_stratified/KO_metagenome_out/p
red_metagenome_contrib.tsv.gz -c contrib_to_legacy -
o picrust2_out_pipeline_with_stratified/KO_metagenome_out/pred_metagenome_
contrib.legacy.tsv.gz
```

APPENDIX D

The purpose of this appendix is to provide the reader with the R script used to generate statistically significant groups with ALDEx2.

```
library(ALDEx2)

phyla8cham <- data.frame(fread("C:\\Users\\Ali\\Documents\\research\\phylu
m_percabund_8chamber.txt", sep = "\t"), row.names = 1)

# Differentiate the two sample sets to compare within the phyla8cham
object by assigning the number of columns to each.
conds <- c(rep("chamb8",8), rep("AS",2))

# Execute the ALDEx function.
x.all <- aldex(phyla8cham, conds, mc.samples = 32, test="t", effect=TRUE,
include.sample.summary = FALSE,denom = "all",verbose = TRUE)

# Plot the results of ALDEx (x.all object).
par(mfrow=c(1,2))
aldex.plot(x.all, type="MA", test="welch", xlab="Log-
ratio abundance", ylab = "Difference")
aldex.plot(x.all, type="MW", test="welch", xlab="Dispersion", ylab = "Diff
erence")

# Display the rows that are significant (P < 0.05) either by the Welchs or
 Wilcoxon method.
found.by.one <- which(x.all$we.eBH < 0.05 | x.all$wi.eBH < 0.05)

# Display the rows that are significant by both Welchs and Wilcoxon
methods.
found.by.one <- which(x.all$we.eBH < 0.05 & x.all$wi.eBH < 0.05)

# Save the statistical outputs of the ALDEx function.
write.table(x.all, "C:\\Users\\Ali\\Documents\\research\\phylum_8AS_aldex_
output.tsv", sep = '\t', quote = FALSE, row.names = FALSE)
```

APPENDIX E

The purpose of this appendix is to provide the reader with the scripts used for the various visuals provided in this document. These include scripts for generating matrices for ordinance plots, generating PCoA plots, and heatmaps.

Ordinance Plots

Below is the Python script used to generate merged count tables at each taxonomic level from ASV count tables and corresponding taxonomic files. The purpose of this code is to find the sum of sums of all taxonomic classifiers across all levels per samples. The summed counts (`output_df`) are then imported into Excel to generate ordinance plots.

```
import argparse
import pandas

TAXONOMIC_LEVELS = ['Kingdom', 'Phylum', 'Class', 'Order', 'Family', 'Genu
s']

def main():
    parser = argparse.ArgumentParser()

    # Parse some arguments.
    parser.add_argument("-i", "--
input", type = str, help = "Input file containing taxonomies and columns
to be summed.", dest = "input_file", default = "HC_abundance_v2.csv")
    parser.add_argument("-t", "--taxonomic-
level", type = str, help = "The taxonomic level of interest. For all,
specify 'ALL'.", dest = "taxonomic_level", default = "ALL")
    parser.add_argument("-d", "--
delimiter", type = str, help = "The delimiter of the input file. Defaults
to comma. For tab, write 'tab'.", dest = "delimiter", default = ',')
    parser.add_argument("-o", "--
output", type = str, help = "The base filename of the output files.
Defaults to '_counts'.", dest = "output_file", default = "_counts")
```

```
    args = parser.parse_args()

    sep = args.delimiter

    if sep == 'tab':
        sep = '\t'

    print("Input File: %s\nTaxonomic Level: %s\nDelimiter: \"%s\"" % (args
.input_file, args.taxonomic_level, args.delimiter))

    df = pandas.read_csv(args.input_file, delimiter = sep)

    if args.taxonomic_level == 'ALL' or args.taxonomic_level == 'all':
        for level in TAXONOMIC_LEVELS:
            print("Computing sums for taxonomic level: \"%s\"" % level)
            compute_sum(df, level, args.output_file, args.delimiter)
            print("... Done!\n")
    else:
        print("Computing sums for taxonomic level: \"%s\"" % args.taxonomi
c_level)
        compute_sum(df, args.taxonomic_level, args.output_file, args.delim
iter)
        print("... Done!\n")

def compute_sum(df, taxonomic_level, output_base, delimiter):
    """
    Compute the sum for the given taxonomic level.

    Arguments:
        df: The dataframe.

        taxonomic_level: The taxonomic level for which the sum(s) will be
computed.
    """
    # Get the unique values for the kingdom/phylum/class/order/family/
      genus column.
    classifiers = df[taxonomic_level].unique()

    # Create a copy of the master list of taxonomic levels.
    drop_labels = TAXONOMIC_LEVELS.copy()

    # Remove the level we're interested in from this list, as we're going
```

```
    to drop all of the columns in this list.
drop_labels.remove(taxonomic_level)
print("Drop labels: " + str(drop_labels))

 # Drop columns for taxonomic levels that we are not interested in.
   Also drop the ASV ID column, which is the first column.
df2 = df.drop(df.columns[0], axis = 'columns')
df2 = df2.drop(drop_labels, axis = 'columns')

# The columns are what we'll be computing the sums for.
columns = None

# Collect data that we'll use to create a dataframe.
data = []

# For each of the unique values in the phylum/class/etc. column...
for classifier in classifiers:
    sums = []

    # Select only the rows from the dataframe corresponding to the
      current classifier.
    df3 = df2.loc[df2[taxonomic_level] == classifier]

    # We only need to populate the columns list once.
    if columns is None:
        columns = df3.columns[1:] # Ignore the taxonomic level column.

    # Compute the sum for each column.
    for column in columns:
        sums.append(df3[column].sum())

    # Compute the sum of sums (i.e., the total for this classifier).
      We have a 'TOTAL' column at the very end.
    sums.append(sum(sums))

    # Store the data.
    data.append([classifier] + sums)

output_file = taxonomic_level + output_base + ".csv"
print("Writing results to file \"%s\" now..." % output_file)

# Create dataframe from the data, then call to_csv to save it.
```

```
    output_df = pandas.DataFrame(data, columns = [taxonomic_level] +
    columns.values.tolist() + ["TOTAL"])
    output_df.to_csv(path_or_buf = output_file, sep = delimiter)

if __name__ == "__main__":
    main()
```

Principal coordinate analysis

Below is the R script used to generate the PCoA plot from sequence variants in a sequence table.

```
library(MicrobiotaProcess)

# Import count table.
seqtab_468sequences_nochim <- data.frame(fread("C:\\Users\\Ali\\Documents\
\research\\seqtab_468sequences_nochim_FINAL.rds", sep = "\t"), row.names =
 1)

# Import sample data table (simple table of 2*n dimensions where
n=samples, as row headers; second column represents feast/famine ratios).
sampleda <- data.frame(fread("C:\\Users\\Ali\\Documents\\research\\sampled
afinal.txt", sep = "\t"), row.names = 1)

# Create ps_dada2 object.
ps_dada2 <- import_dada2(seqtab=seqtab, sampleda=sampleda)

# Create PCoA plot.
pcoares <- get_pcoa(obj=ps_dada2, distmethod="euclidean", method="hellinge
r")
pcoaplot <- ggordpoint(obj=pcoares, biplot=TRUE, speciesannot=TRUE,
                       factorNames=c("Feast/famine"), ellipse=TRUE)
```

Heatmaps

Below is the R script used to generate heatmaps, both for taxonomic visualization as well as predictive functional visualizations.

```
library(gplots)
require(data.table)

# Import desired file (i.e., taxonomy count table at the phylum level).
phyla8cham <- data.frame(fread("C:\\Users\\Ali\\Documents\\research\\phylu
m8AS_significant_taxa_aldex.txt", sep = "\t"), row.names = 1)

# Scale the samples so that the dendrogram is not squished; requires
transposition function because cols are transformed by default.
z <- t(scale(t(phyla8cham)))

# Set custom distance and clustering functions for the dendrogram.
hclustfunc <- function(x) hclust(x, method="complete")
distfunc <- function(x) dist(x,method="maximum")
fit <- hclustfunc(distfunc(z))
clusters <- cutree(fit, 5)

# Plot the heatmap: without a column dendrogram, using the green/red palette
from gplots, and with smaller row text size.
heatmap(z, Colv = NA, col=greenred(256),cexRow = 0.8)

# Make the color key and histogram.
heatmap.2(z, col=greenred(256))
```

REFERENCES

(1) Burian, S. J.; Nix, S. J.; Pitt, R. E.; Durrans, S. R. Urban Wastewater Management in the United States: Past, Present, and Future. *J. Urban Technol.* **2000**, *7* (3), 33–62. https://doi.org/10.1080/713684134.

(2) *How Wastewater Treatment Works... The Basics*; United States Environmental Protection Agency: Office of Water, 1998.

(3) National Pollutant Discharge Elimination System (NPDES) Permit Writers' Manual. *US Environ. Prot. Agency Off. Wastewater Manag. Water Permits Div.* **2010**, 269.

(4) Kent, T. R.; Bott, C. B.; Wang, Z.-W. State of the Art of Aerobic Granulation in Continuous Flow Bioreactors. *Biotechnol. Adv.* **2018**, *36* (4), 1139–1166. https://doi.org/10.1016/j.biotechadv.2018.03.015.

(5) Liu, Y.; Tay, J.-H. The Essential Role of Hydrodynamic Shear Force in the Formation of Biofilm and Granular Sludge. *Water Res.* **2002**, *36* (7), 1653–1665. https://doi.org/10.1016/s0043-1354(01)00379-7.

(6) Beun, J. J.; Hendriks, A.; van Loosdrecht, M. C. M.; Morgenroth, E.; Wilderer, P. A.; Heijnen, J. J. Aerobic Granulation in a Sequencing Batch Reactor. *Water Res.* **1999**, *33* (10), 2283–2290. https://doi.org/10.1016/S0043-1354(98)00463-1.

(7) Beun, J. J.; Heijnen, J. J.; van Loosdrecht, M. C. N-Removal in a Granular Sludge Sequencing Batch Airlift Reactor. *Biotechnol. Bioeng.* **2001**, *75* (1), 82–92. https://doi.org/10.1002/bit.1167.

(8) de Kreuk, M. K.; Heijnen, J. J.; van Loosdrecht, M. C. M. Simultaneous COD, Nitrogen, and Phosphate Removal by Aerobic Granular Sludge. *Biotechnol. Bioeng.* **2005**, *90* (6), 761–769. https://doi.org/10.1002/bit.20470.

(9) Hasebe, Y.; Meguro, H.; Kanai, Y.; Eguchi, M.; Osaka, T.; Tsuneda, S. High-Rate Nitrification of Electronic Industry Wastewater by Using Nitrifying Granules. *Water Sci. Technol.* **2017**, *76* (11), 3171–3180. https://doi.org/10.2166/wst.2017.431.

(10) Antunes, L. P.; Martins, L. F.; Pereira, R. V.; Thomas, A. M.; Barbosa, D.; Lemos, L. N.; Silva, G. M. M.; Moura, L. M. S.; Epamino, G. W. C.; Digiampietri, L. A.; Lombardi, K. C.; Ramos, P. L.; Quaggio, R. B.; de Oliveira, J. C. F.; Pascon, R. C.; Cruz, J. B. da; da Silva, A. M.; Setubal, J. C. Microbial Community Structure and Dynamics in Thermophilic Composting Viewed through Metagenomics and Metatranscriptomics. *Sci. Rep.* **2016**, *6* (1), 38915. https://doi.org/10.1038/srep38915.

(11) Chiesa, S.; Irvine, R.; Manning, J. Feast/Famine Growth Environments and Activated Sludge Population Selection. *Biotechnol. Bioeng.* **1985**, *27*, 562–568. https://doi.org/10.1002/bit.260270503.

(12) Chen, C.; Bin, L.; Tang, B.; Huang, S.; Fu, F.; Chen, Q.; Wu, L.; Wu, C. Cultivating Granular Sludge Directly in a Continuous-Flow Membrane Bioreactor with Internal Circulation. *Chem. Eng. J.* **2016**, *309*. https://doi.org/10.1016/j.cej.2016.10.034.

(13) Juang, Y.-C.; Adav, S. S.; Lee, D.-J.; Tay, J.-H. Stable Aerobic Granules for Continuous-Flow Reactors: Precipitating Calcium and Iron Salts in Granular Interiors. *Bioresour. Technol.* **2010**, *101* (21), 8051–8057. https://doi.org/10.1016/j.biortech.2010.05.078.

(14) Li, D.; Lv, Y.; Zeng, H.; Zhang, J. Startup and Long Term Operation of Enhanced Biological Phosphorus Removal in Continuous-Flow Reactor with Granules. *Bioresour. Technol.* **2016**, *212*, 92–99. https://doi.org/10.1016/j.biortech.2016.04.008.

(15) Hunt, S. M.; Werner, E. M.; Huang, B.; Hamilton, M. A.; Stewart, P. S. Hypothesis for the Role of Nutrient Starvation in Biofilm Detachment. *Appl. Environ. Microbiol.* **2004**, *70* (12), 7418–7425. https://doi.org/10.1128/AEM.70.12.7418-7425.2004.

(16) Pronk, M.; de Kreuk, M. K.; de Bruin, B.; Kamminga, P.; Kleerebezem, R.; van Loosdrecht, M. C. M. Full Scale Performance of the Aerobic Granular Sludge Process for Sewage Treatment. *Water Res.* **2015**, *84*, 207–217. https://doi.org/10.1016/j.watres.2015.07.011.

(17) Lecture 25: Plug flow reactors and comparison to continuously stirred tank reactors https://chem.libretexts.org/Courses/New_York_University/CHEM-UA_652%3A_Thermodynamics_and_Kinetics/Lecture_25%3A_Plug_flow_reactors_and_comparison_to_continuously_stirred_tank_reactors (accessed Apr 4, 2021).

(18) Sun, Y.; Angelotti, B.; Brooks, M.; Wang, Z.-W. Feast/Famine Ratio Determined Continuous Flow Aerobic Granulation. *Sci. Total Environ.* **2021**, *750*, 141467. https://doi.org/10.1016/j.scitotenv.2020.141467.

(19) Sun, Y.; Angelotti, B.; Wang, Z.-W. Continuous-Flow Aerobic Granulation in Plug-Flow Bioreactors Fed with Real Domestic Wastewater. *Sci. Total Environ.* **2019**, *688*, 762–770. https://doi.org/10.1016/j.scitotenv.2019.06.291.

(20) Himeoka, Y.; Mitarai, N. Dynamics of Bacterial Populations under the Feast-Famine Cycles. *Phys. Rev. Res.* **2020**, *2* (1), 013372. https://doi.org/10.1103/PhysRevResearch.2.013372.

(21) Gao, D.; Liu, L.; Liang, H.; Wu, W.-M. Aerobic Granular Sludge: Characterization, Mechanism of Granulation and Application to Wastewater Treatment. *Crit. Rev. Biotechnol.* **2011**, *31* (2), 137–152. https://doi.org/10.3109/07388551.2010.497961.

(22) Sun, Y.; Gomeiz, A. T.; Van Aken, B.; Angelotti, B.; Brooks, M.; Wang, Z.-W. Dynamic Response of Aerobic Granular Sludge to Feast and Famine Conditions in Plug Flow Reactors Fed with Real Domestic Wastewater. *Sci. Total Environ.* **2021**, *758*, 144155. https://doi.org/10.1016/j.scitotenv.2020.144155.

(23) Lee, C. C.; Lin, S. D. *Handbook of Environmental Engineering Calculations 2nd Ed.*, 2nd edition.; McGraw-Hill Education, 2007.

(24) Liu, Y.; Liu, Q.-S. Causes and Control of Filamentous Growth in Aerobic Granular Sludge Sequencing Batch Reactors. *Biotechnol. Adv.* **2006**, *24* (1), 115–127. https://doi.org/10.1016/j.biotechadv.2005.08.001.

(25) López–Palau, S.; Dosta, J.; Mata-Álvarez, J. Start-up of an Aerobic Granular Sequencing Batch Reactor for the Treatment of Winery Wastewater. *Water Sci. Technol.* **2009**, *60* (4), 1049–1054. https://doi.org/10.2166/wst.2009.554.

(26) López-Palau, S.; Pinto, A.; Basset, N.; Dosta, J.; Mata-Álvarez, J. ORP Slope and Feast–Famine Strategy as the Basis of the Control of a Granular Sequencing Batch Reactor Treating Winery Wastewater. *Biochem. Eng. J.* **2012**, *68*, 190–198. https://doi.org/10.1016/j.bej.2012.08.002.

(27) Świątczak, P.; Cydzik-Kwiatkowska, A. Performance and Microbial Characteristics of Biomass in a Full-Scale Aerobic Granular Sludge Wastewater Treatment Plant. *Environ. Sci. Pollut. Res.* **2018**, *25* (2), 1655–1669. https://doi.org/10.1007/s11356-017-0615-9.

(28) Liu, Y.; Tay, J.-H. State of the Art of Biogranulation Technology for Wastewater Treatment. *Biotechnol. Adv.* **2004**, *22* (7), 533–563. https://doi.org/10.1016/j.biotechadv.2004.05.001.

(29) dos Santos, A. L. S.; Galdino, A. C. M.; de Mello, T. P.; Ramos, L. de S.; Branquinha, M. H.; Bolognese, A. M.; Columbano, J.; Roudbary, M. What Are the Advantages of Living in a Community? A Microbial Biofilm Perspective! *Mem. Inst. Oswaldo Cruz* **2018**, *113* (9). https://doi.org/10.1590/0074-02760180212.

(30) Liu, X.; Dong, C. Simultaneous COD and Nitrogen Removal in a Micro-Aerobic Granular Sludge Reactor for Domestic Wastewater Treatment. *Syst. Eng. Procedia* **2011**, *1*, 99–105. https://doi.org/10.1016/j.sepro.2011.08.017.

(31) Wei, D.; Qiao, Z.; Zhang, Y.; Hao, L.; Si, W.; Du, B.; Wei, Q. Effect of COD/N Ratio on Cultivation of Aerobic Granular Sludge in a Pilot-Scale Sequencing Batch Reactor. *Appl. Microbiol. Biotechnol.* **2013**, *97* (4), 1745–1753. https://doi.org/10.1007/s00253-012-3991-6.

(32) Tay, J.-H.; Liu, Q.-S.; Liu, Y. The Effects of Shear Force on the Formation, Structure and Metabolism of Aerobic Granules. *Appl. Microbiol. Biotechnol.* **2001**, *57* (1), 227–233. https://doi.org/10.1007/s002530100766.

(33) Liu, Y.-Q.; Tay, J.-H. Fast Formation of Aerobic Granules by Combining Strong Hydraulic Selection Pressure with Overstressed Organic Loading Rate. *Water Res.* **2015**, *80*, 256–266. https://doi.org/10.1016/j.watres.2015.05.015.

(34) Wingender, J.; Neu, T. R.; Flemming, H.-C. What Are Bacterial Extracellular Polymeric Substances? In *Microbial Extracellular Polymeric Substances: Characterization, Structure and Function*; Wingender, J., Neu, T. R., Flemming, H.-C., Eds.; Springer: Berlin, Heidelberg, 1999; pp 1–19. https://doi.org/10.1007/978-3-642-60147-7_1.

(35) Liu, Y.-Q.; Lan, G.-H.; Zeng, P. Resistance and Resilience of Nitrifying Bacteria in Aerobic Granules to PH Shock. *Lett. Appl. Microbiol.* **2015**, *61* (1), 91–97. https://doi.org/10.1111/lam.12433.

(36) Wu, X.; Li, H.; Lei, L.; Ren, J.; Li, W.; Liu, Y. Tolerance to Short-Term Saline Shocks by Aerobic Granular Sludge. *Chemosphere* **2020**, *243*, 125370. https://doi.org/10.1016/j.chemosphere.2019.125370.

(37) Jiang, H.-L.; Tay, J.-H.; Tay, S. T.-L. Changes in Structure, Activity and Metabolism of Aerobic Granules as a Microbial Response to High Phenol Loading. *Appl. Microbiol. Biotechnol.* **2004**, *63* (5), 602–608. https://doi.org/10.1007/s00253-003-1358-8.

(38) Pentachlorophenol. *Environ. Prot. Agency* **2016**.

(39) Conco, T.; Kumari, S.; Stenström, T.; Bux, F. Epibiont Growth on Filamentous Bacteria Found in Activated Sludge: A Morphological Approach. *Arch. Microbiol.* **2018**, *200* (3), 493–503. https://doi.org/10.1007/s00203-017-1461-3.
(40) Glymph, T. Identification and Control of Filamentous Bacteria, 2013.
(41) Nielsen, S. *Food Analysis Laboratory Manual*, 2nd ed.; Food Science Text Series; Springer US, 2010. https://doi.org/10.1007/978-1-4419-1463-7.
(42) Jenkins, D.; Richard, M. G.; Daigger, G. T. *Manual on the Causes and Control of Activated Sludge Bulking, Foaming, and Other Solids Separation Problems*, 3rd edition.; CRC Press: Boca Raton, Fla, 2003.
(43) Tandoi, V.; Jenkins, D.; Wanner, J. Activated Sludge Separation Problems. *Act. Sludge Sep. Probl. Theory Control Meas. Pract. Exp.* **2006**, 35–46.
(44) Xu, S.; Yao, J.; Ainiwaer, M.; Hong, Y.; Zhang, Y. Analysis of Bacterial Community Structure of Activated Sludge from Wastewater Treatment Plants in Winter. *BioMed Res. Int.* **2018**, *2018*, e8278970. https://doi.org/10.1155/2018/8278970.
(45) Seviour, R. J.; Mino, T.; Onuki, M. The Microbiology of Biological Phosphorus Removal in Activated Sludge Systems. *FEMS Microbiol. Rev.* **2003**, *27* (1), 99–127. https://doi.org/10.1016/S0168-6445(03)00021-4.
(46) Xia, J.; Ye, L.; Ren, H.; Zhang, X.-X. Microbial Community Structure and Function in Aerobic Granular Sludge. *Appl. Microbiol. Biotechnol.* **2018**, *102* (9), 3967–3979. https://doi.org/10.1007/s00253-018-8905-9.
(47) DNeasy PowerSoil Pro Kit Handbook - QIAGEN https://www.qiagen.com/us/resources/resourcedetail?id=9bb59b74-e493-4aeb-b6c1-f660852e8d97&lang=en (accessed Nov 29, 2020).
(48) BeadBug™ 3 Position Bead Homogenizer https://www.thomassci.com/Equipment/Homogenizers/_/BeadBug-Microtube-Homogenizer (accessed Feb 23, 2021).
(49) Thijs, S.; Op De Beeck, M.; Beckers, B.; Truyens, S.; Stevens, V.; Van Hamme, J. D.; Weyens, N.; Vangronsveld, J. Comparative Evaluation of Four Bacteria-Specific Primer Pairs for 16S RRNA Gene Surveys. *Front. Microbiol.* **2017**, *8*. https://doi.org/10.3389/fmicb.2017.00494.
(50) EC2 On-Demand Instance Pricing – Amazon Web Services https://aws.amazon.com/ec2/pricing/on-demand/ (accessed Feb 25, 2021).
(51) McNally, C. P.; Eng, A.; Noecker, C.; Gagne-Maynard, W. C.; Borenstein, E. BURRITO: An Interactive Multi-Omic Tool for Visualizing Taxa–Function Relationships in Microbiome Data. *Front. Microbiol.* **2018**, *9*. https://doi.org/10.3389/fmicb.2018.00365.
(52) Fernandes, A. D.; Macklaim, J. M.; Linn, T. G.; Reid, G.; Gloor, G. B. ANOVA-like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PloS One* **2013**, *8* (7), e67019. https://doi.org/10.1371/journal.pone.0067019.
(53) Xu, S.; Yu, G. *MicrobiotaProcess: An R Package for Analysis, Visualization and Biomarker Discovery of Microbiome*; Bioconductor version: Release (3.12), 2021. https://doi.org/10.18129/B9.bioc.MicrobiotaProcess.

(54) Li, J.; Ding, L.-B.; Cai, A.; Huang, G.-X.; Horn, H. Aerobic Sludge Granulation in a Full-Scale Sequencing Batch Reactor. *BioMed Res. Int.* **2014**, *2014*, 268789. https://doi.org/10.1155/2014/268789.
(55) Cassidy, D. P.; Belia, E. Nitrogen and Phosphorus Removal from an Abattoir Wastewater in a SBR with Aerobic Granular Sludge. *Water Res.* **2005**, *39* (19), 4817–4823. https://doi.org/10.1016/j.watres.2005.09.025.
(56) Tao, J.; Qin, L.; Liu, X.; Li, B.; Chen, J.; You, J.; Shen, Y.; Chen, X. Effect of Granular Activated Carbon on the Aerobic Granulation of Sludge and Its Mechanism. *Bioresour. Technol.* **2017**, *236*, 60–67. https://doi.org/10.1016/j.biortech.2017.03.106.
(57) Zheng, Y.-M.; Yu, H.-Q.; Sheng, G.-P. Physical and Chemical Characteristics of Granular Activated Sludge from a Sequencing Batch Airlift Reactor. *Process Biochem.* **2005**, *40* (2), 645–650. https://doi.org/10.1016/j.procbio.2004.01.056.
(58) Su, K.-Z.; Yu, H.-Q. Formation and Characterization of Aerobic Granules in a Sequencing Batch Reactor Treating Soybean-Processing Wastewater. *Environ. Sci. Technol.* **2005**, *39* (8), 2818–2827. https://doi.org/10.1021/es048950y.
(59) Schuler, A. J.; Jang, H. Density Effects on Activated Sludge Zone Settling Velocities. *Water Res.* **2007**, *41* (8), 1814–1822. https://doi.org/10.1016/j.watres.2007.01.011.
(60) Appendix I: F/M, HRT, MCRT, MLVSS, Sludge Age, SVI. In *Settleability Problems and Loss of Solids in the Activated Sludge Process*; John Wiley & Sons, Ltd, 2002; pp 153–156. https://doi.org/10.1002/047147164X.app1.
(61) McSwain, B. S.; Irvine, R. L.; Hausner, M.; Wilderer, P. A. Composition and Distribution of Extracellular Polymeric Substances in Aerobic Flocs and Granular Sludge. *Appl. Environ. Microbiol.* **2005**, *71* (2), 1051–1057. https://doi.org/10.1128/AEM.71.2.1051-1057.2005.
(62) Wang, Z.-W.; Liu, Y.; Tay, J.-H. Biodegradability of Extracellular Polymeric Substances Produced by Aerobic Granules. *Appl. Microbiol. Biotechnol.* **2007**, *74* (2), 462–466. https://doi.org/10.1007/s00253-006-0686-x.
(63) Sponza, D. T. Investigation of Extracellular Polymer Substances (EPS) and Physicochemical Properties of Different Activated Sludge Flocs under Steady-State Conditions. *Enzyme Microb. Technol.* **2003**, *32* (3), 375–385. https://doi.org/10.1016/S0141-0229(02)00309-5.
(64) Dignac, M.-F.; Urbain, V.; Rybacki, D.; Bruchet, A.; Snidaro, D.; Scribe, P. Chemical Description of Extracellular Polymers: Implication on Activated Sludge Floc Structure. *Water Sci. Technol.* **1998**, *38* (8), 45–53. https://doi.org/10.1016/S0273-1223(98)00676-3.
(65) Chen, Y.-C.; Lin, C.-J.; Chen, H.-L.; Fu, S.-Y.; Zhan, H.-Y. Cultivation of Biogranules in a Continuous Flow Reactor at Low Dissolved Oxygen. *Water Air Soil Pollut. Focus* **2009**, *9* (3), 213–221. https://doi.org/10.1007/s11267-009-9216-z.
(66) Guo, F.; Zhang, S.-H.; Yu, X.; Wei, B. Variations of Both Bacterial Community and Extracellular Polymers: The Inducements of Increase of Cell Hydrophobicity from

Biofloc to Aerobic Granule Sludge. *Bioresour. Technol.* **2011**, *102* (11), 6421–6428. https://doi.org/10.1016/j.biortech.2011.03.046.
(67) Jia, X. S.; Furumai, H.; Fang, H. H. P. Extracellular Polymers of Hydrogen-Utilizing Methanogenic and Sulfate-Reducing Sludges. *Water Res.* **1996**, *30* (6), 1439–1444. https://doi.org/10.1016/0043-1354(96)00028-0.
(68) Tang, C.-J.; Zheng, P.; Wang, C.-H.; Mahmood, Q.; Zhang, J.-Q.; Chen, X.-G.; Zhang, L.; Chen, J.-W. Performance of High-Loaded ANAMMOX UASB Reactors Containing Granular Sludge. *Water Res.* **2011**, *45* (1), 135–144. https://doi.org/10.1016/j.watres.2010.08.018.
(69) Mikkelsen, L. H.; Keiding, K. Physico-Chemical Characteristics of Full Scale Sewage Sludges with Implications to Dewatering. *Water Res.* **2002**, *36* (10), 2451–2462. https://doi.org/10.1016/s0043-1354(01)00477-8.
(70) Deng, S.; Wang, L.; Su, H. Role and Influence of Extracellular Polymeric Substances on the Preparation of Aerobic Granular Sludge. *J. Environ. Manage.* **2016**, *173*, 49–54. https://doi.org/10.1016/j.jenvman.2016.03.008.
(71) Ye, L.; Mei, R.; Liu, W.-T.; Ren, H.; Zhang, X.-X. Machine Learning-Aided Analyses of Thousands of Draft Genomes Reveal Specific Features of Activated Sludge Processes. *Microbiome* **2020**, *8* (1), 16. https://doi.org/10.1186/s40168-020-0794-3.
(72) Nierychlo, M.; Andersen, K. S.; Xu, Y.; Green, N.; Jiang, C.; Albertsen, M.; Dueholm, M. S.; Nielsen, P. H. MiDAS 3: An Ecosystem-Specific Reference Database, Taxonomy and Knowledge Platform for Activated Sludge and Anaerobic Digesters Reveals Species-Level Microbiome Composition of Activated Sludge. *Water Res.* **2020**, *182*, 115955. https://doi.org/10.1016/j.watres.2020.115955.
(73) Waite, D. W.; Chuvochina, M.; Pelikan, C.; Parks, D. H.; Yilmaz, P.; Wagner, M.; Loy, A.; Naganuma, T.; Nakai, R.; Whitman, W. B.; Hahn, M. W.; Kuever, J.; Hugenholtz, P. Proposal to Reclassify the Proteobacterial Classes Deltaproteobacteria and Oligoflexia, and the Phylum Thermodesulfobacteria into Four Phyla Reflecting Major Functional Capabilities. *Int. J. Syst. Evol. Microbiol.* **2020**, *70* (11), 5972–6016. https://doi.org/10.1099/ijsem.0.004213.
(74) Hamza, R. A. S. E. Development of Upflow Aerobic Granular Sludge Bioreactor (UAGSBR) for Treatment of High-Strength Organic Wastewater. *Univ. Calgargy* **2019**, 242.
(75) Fudou, R.; Jojima, Y.; Iizuka, T.; Yamanaka, S. Haliangium Ochraceum Gen. Nov., Sp. Nov. and Haliangium Tepidum Sp. Nov.: Novel Moderately Halophilic Myxobacteria Isolated from Coastal Saline Environments. *J. Gen. Appl. Microbiol.* **2002**, *48* (2), 109–116. https://doi.org/10.2323/jgam.48.109.
(76) Chen, H.; Wang, M.; Chang, S. Disentangling Community Structure of Ecological System in Activated Sludge: Core Communities, Functionality, and Functional Redundancy. *Microb. Ecol.* **2020**, *80* (2), 296–308. https://doi.org/10.1007/s00248-020-01492-y.
(77) Yang, Y.; Wang, L.; Xiang, F.; Zhao, L.; Qiao, Z. Activated Sludge Microbial Community and Treatment Performance of Wastewater Treatment Plants in

Industrial and Municipal Zones. *Int. J. Environ. Res. Public. Health* **2020**, *17* (2). https://doi.org/10.3390/ijerph17020436.

(78) Wilén, B.-M.; Liébana, R.; Persson, F.; Modin, O.; Hermansson, M. The Mechanisms of Granulation of Activated Sludge in Wastewater Treatment, Its Optimization, and Impact on Effluent Quality. *Appl. Microbiol. Biotechnol.* **2018**, *102* (12), 5005–5020. https://doi.org/10.1007/s00253-018-8990-9.

(79) Vu, B.; Chen, M.; Crawford, R. J.; Ivanova, E. P. Bacterial Extracellular Polysaccharides Involved in Biofilm Formation. *Molecules* **2009**, *14* (7), 2535–2554. https://doi.org/10.3390/molecules14072535.

(80) Chinh, T. T.; Hieu, P. D.; Cuong, B. V.; Linh, N. N.; Lan, N. N.; Nguyen, N. S.; Hung, N. Q.; Hien, L. T. T. Sequencing Batch Reactor and Bacterial Community in Aerobic Granular Sludge for Wastewater Treatment of Noodle-Manufacturing Sector. *Appl. Sci.* **2018**, *8* (4), 509. https://doi.org/10.3390/app8040509.

(81) de Graaff, D. R.; van Loosdrecht, M. C. M.; Pronk, M. Stable Granulation of Seawater-Adapted Aerobic Granular Sludge with Filamentous Thiothrix Bacteria. *Water Res.* **2020**, *175*, 115683. https://doi.org/10.1016/j.watres.2020.115683.

(82) Lee, H.; Choi, E.; Yun, Z.; Park, Y. K. Microbial Structure and Community of RBC Biofilm Removing Nitrate and Phosphorus from Domestic Wastewater. *J. Microbiol. Biotechnol.* **2008**, *18* (8), 1459–1469.

(83) Krasowska, A.; Sigler, K. How Microorganisms Use Hydrophobicity and What Does This Mean for Human Needs? *Front. Cell. Infect. Microbiol.* **2014**, *4*. https://doi.org/10.3389/fcimb.2014.00112.

(84) Rainey, F. A.; Ward-Rainey, N.; Gliesche, C. G.; Stackebrandt, E. Phylogenetic Analysis and Intrageneric Structure of the Genus Hyphomicrobium and the Related Genus Filomicrobium. *Int. J. Syst. Bacteriol.* **1998**, *48 Pt 3*, 635–639. https://doi.org/10.1099/00207713-48-3-635.

(85) Rossetti, S.; Tomei, M. C.; Nielsen, P. H.; Tandoi, V. "Microthrix Parvicella", a Filamentous Bacterium Causing Bulking and Foaming in Activated Sludge Systems: A Review of Current Knowledge. *FEMS Microbiol. Rev.* **2005**, *29* (1), 49–64. https://doi.org/10.1016/j.femsre.2004.09.005.

(86) Bin, Z.; Zhe, C.; Zhigang, Q.; Min, J.; Zhiqiang, C.; Zhaoli, C.; Junwen, L.; Xuan, W.; Jingfeng, W. Dynamic and Distribution of Ammonia-Oxidizing Bacteria Communities during Sludge Granulation in an Anaerobic-Aerobic Sequencing Batch Reactor. *Water Res.* **2011**, *45* (18), 6207–6216. https://doi.org/10.1016/j.watres.2011.09.026.

(87) Crocetti, G. R.; Hugenholtz, P.; Bond, P. L.; Schuler, A.; Keller, J.; Jenkins, D.; Blackall, L. L. Identification of Polyphosphate-Accumulating Organisms and Design of 16S RRNA-Directed Probes for Their Detection and Quantitation. *Appl. Environ. Microbiol.* **2000**, *66* (3), 1175–1182. https://doi.org/10.1128/AEM.66.3.1175-1182.2000.

(88) Treatment Process https://www.uosa.org/DisplayContentUOSA.asp?ID=490 (accessed Mar 19, 2021).

(89) Lopez-Vazquez, C. M.; Hooijmans, C. M.; Brdjanovic, D.; Gijzen, H. J.; van Loosdrecht, M. C. M. Temperature Effects on Glycogen Accumulating Organisms. *Water Res.* **2009**, *43* (11), 2852–2864. https://doi.org/10.1016/j.watres.2009.03.038.

(90) Aslam, Z.; Im, W.-T.; Kim, M. K.; Lee, S.-T. Flavobacterium Granuli Sp. Nov., Isolated from Granules Used in a Wastewater Treatment Plant. *Int. J. Syst. Evol. Microbiol.* **2005**, *55* (Pt 2), 747–751. https://doi.org/10.1099/ijs.0.63459-0.

(91) Kim, M. K.; Im, W.-T.; Ohta, H.; Lee, M.; Lee, S.-T. Sphingopyxis Granuli Sp. Nov., a Beta-Glucosidase-Producing Bacterium in the Family Sphingomonadaceae in Alpha-4 Subclass of the Proteobacteria. *J. Microbiol. Seoul Korea* **2005**, *43* (2), 152–157.

(92) Gómez-Acata, S.; Vital-Jácome, M.; Pérez-Sandoval, M. V.; Navarro-Noya, Y. E.; Thalasso, F.; Luna-Guido, M.; Conde-Barajas, E.; Dendooven, L. Microbial Community Structure in Aerobic and Fluffy Granules Formed in a Sequencing Batch Reactor Supplied with 4-Chlorophenol at Different Settling Times. *J. Hazard. Mater.* **2018**, *342*, 606–616. https://doi.org/10.1016/j.jhazmat.2017.08.073.

(93) Chung, B. S.; Ryu, S. H.; Park, M.; Jeon, Y.; Chung, Y. R.; Jeon, C. O. Hydrogenophaga Caeni Sp. Nov., Isolated from Activated Sludge. *Int. J. Syst. Evol. Microbiol.* **2007**, *57* (Pt 5), 1126–1130. https://doi.org/10.1099/ijs.0.64629-0.

(94) Faizan Khan, M.; Yu, L.; Hollman, J.; Hwa Tay, J.; Achari, G. Integration of Aerobic Granulation and UV/H 2 O 2 Processes in a Continuous Flow System for the Degradation of Sulfolane in Contaminated Water. *Environ. Sci. Water Res. Technol.* **2020**, *6* (6), 1711–1722. https://doi.org/10.1039/C9EW01048C.

(95) Müller, E.; Schade, M.; Lemmer, H. Filamentous Scum Bacteria in Activated Sludge Plants: Detection and Identification Quality by Conventional Activated Sludge Microscopy versus Fluorescence in Situ Hybridization. *Water Environ. Res. Res. Publ. Water Environ. Fed.* **2007**, *79* (11), 2274–2286. https://doi.org/10.2175/106143007x183943.

(96) Nouha, K.; Kumar, R. S.; Tyagi, R. D. Heavy Metals Removal from Wastewater Using Extracellular Polymeric Substances Produced by Cloacibacterium Normanense in Wastewater Sludge Supplemented with Crude Glycerol and Study of Extracellular Polymeric Substances Extraction by Different Methods. *Bioresour. Technol.* **2016**, *212*, 120–129. https://doi.org/10.1016/j.biortech.2016.04.021.

(97) Williams, H. N.; Baer, M. L.; Tudor, J. J. Bdellovibrio. In *Bergey's Manual of Systematics of Archaea and Bacteria*; American Cancer Society, 2015; pp 1–22. https://doi.org/10.1002/9781118960608.gbm01007.

(98) Kämpfer, P.; Witzenberger, R.; Denner, E. B. M.; Busse, H.-J.; Neef, A. Novosphingobium Hassiacum Sp. Nov., a New Species Isolated from an Aerated Sewage Pond. *Syst. Appl. Microbiol.* **2002**, *25* (1), 37–45. https://doi.org/10.1078/0723-2020-00083.

(99) Wongwongsee, W.; Chareanpat, P.; Pinyakong, O. Abilities and Genes for PAH Biodegradation of Bacteria Isolated from Mangrove Sediments from the Central of Thailand. *Mar. Pollut. Bull.* **2013**, *74* (1), 95–104. https://doi.org/10.1016/j.marpolbul.2013.07.025.

(100) Richard, M.; Hao, O.; Jenkins, D. Growth Kinetics of Sphaerotilus Species and Their Significance in Activated Sludge Bulking. *J. Water Pollut. Control Fed.* **1985**, *57* (1), 68–81.
(101) Weissbrodt, D. G.; Lochmatter, S.; Ebrahimi, S.; Rossi, P.; Maillard, J.; Holliger, C. Bacterial Selection during the Formation of Early-Stage Aerobic Granules in Wastewater Treatment Systems Operated Under Wash-Out Dynamics. *Front. Microbiol.* **2012**, *3*. https://doi.org/10.3389/fmicb.2012.00332.
(102) Guo, J.; Peng, Y.; Yang, X.; Wang, Z.; Zhu, A. Changes in the Microbial Community Structure of Filaments and Floc Formers in Response to Various Carbon Sources and Feeding Patterns. *Appl. Microbiol. Biotechnol.* **2014**, *98* (17), 7633–7644. https://doi.org/10.1007/s00253-014-5805-5.
(103) Gulez, G.; de Los Reyes, F. L. Multiple Approaches to Assess Filamentous Bacterial Growth in Activated Sludge under Different Carbon Source Conditions. *J. Appl. Microbiol.* **2009**, *106* (2), 682–691. https://doi.org/10.1111/j.1365-2672.2008.04049.x.
(104) Bousses, K. Molecular Analyses of Microbial Mats. Thesis, University of Delaware, 2018.
(105) Chen, H.; Li, A.; Cui, D.; Cui, C.; Ma, F. Evolution of Microbial Community and Key Genera in the Formation and Stability of Aerobic Granular Sludge under a High Organic Loading Rate. *Bioresour. Technol. Rep.* **2019**, *7*, 100280. https://doi.org/10.1016/j.biteb.2019.100280.
(106) How long is RNA stable at room temperature? | AAT Bioquest https://www.aatbio.com/resources/faq-frequently-asked-questions/How-long-is-RNA-stable-at-room-temperature (accessed Apr 11, 2021).
(107) Shrout, J. D.; Tolker-Nielsen, T.; Givskov, M.; Parsek, M. R. The Contribution of Cell-Cell Signaling and Motility to Bacterial Biofilm Formation. *MRS Bull. Mater. Res. Soc.* **2011**, *36* (5), 367–373. https://doi.org/10.1557/mrs.2011.67.
(108) Liu, C.; Sun, D.; Zhu, J.; Liu, W. Two-Component Signal Transduction Systems: A Major Strategy for Connecting Input Stimuli to Biofilm Formation. *Front. Microbiol.* **2019**, *9*. https://doi.org/10.3389/fmicb.2018.03279.
(109) Tanaka, K.; Takayanagi, Y.; Fujita, N.; Ishihama, A.; Takahashi, H. Heterogeneity of the Principal Sigma Factor in Escherichia Coli: The RpoS Gene Product, Sigma 38, Is a Second Principal Sigma Factor of RNA Polymerase in Stationary-Phase Escherichia Coli. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90* (8), 3511–3515. https://doi.org/10.1073/pnas.90.8.3511.
(110) Cerca, N.; Jefferson, K. K. Effect of Growth Conditions on Poly-N-Acetylglucosamine Expression and Biofilm Formation in Escherichia Coli. *FEMS Microbiol. Lett.* **2008**, *283* (1), 36–41. https://doi.org/10.1111/j.1574-6968.2008.01142.x.
(111) Mukhopadhyay, S.; Audia, J. P.; Roy, R. N.; Schellhorn, H. E. Transcriptional Induction of the Conserved Alternative Sigma Factor RpoS in Escherichia Coli Is Dependent on BarA, a Probable Two-Component Regulator. *Mol. Microbiol.* **2000**, *37* (2), 371–381. https://doi.org/10.1046/j.1365-2958.2000.01999.x.

(112) Chavez, R. G.; Alvarez, A. F.; Romeo, T.; Georgellis, D. The Physiological Stimulus for the BarA Sensor Kinase. *J. Bacteriol.* **2010**, *192* (7), 2009–2012. https://doi.org/10.1128/JB.01685-09.
(113) Pernestig, A. K.; Melefors, O.; Georgellis, D. Identification of UvrY as the Cognate Response Regulator for the BarA Sensor Kinase in Escherichia Coli. *J. Biol. Chem.* **2001**, *276* (1), 225–231. https://doi.org/10.1074/jbc.M001550200.
(114) Sonnleitner, E.; Romeo, A.; Bläsi, U. Small Regulatory RNAs in Pseudomonas Aeruginosa. *RNA Biol.* **2012**, *9* (4), 364–371. https://doi.org/10.4161/rna.19231.
(115) Hao, T.; Mackey, H. R.; Guo, G.; Liu, R.; Chen, G. Resilience of Sulfate-Reducing Granular Sludge against Temperature, PH, Oxygen, Nitrite, and Free Nitrous Acid. *Appl. Microbiol. Biotechnol.* **2016**, *100* (19), 8563–8572. https://doi.org/10.1007/s00253-016-7652-z.
(116) Yang, L.; Haagensen, J. A. J.; Jelsbak, L.; Johansen, H. K.; Sternberg, C.; Høiby, N.; Molin, S. In Situ Growth Rates and Biofilm Development of Pseudomonas Aeruginosa Populations in Chronic Lung Infections. *J. Bacteriol.* **2008**, *190* (8), 2767–2776. https://doi.org/10.1128/JB.01581-07.
(117) Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28* (1), 27–30. https://doi.org/10.1093/nar/28.1.27.
(118) El Khatib, M.; Tran, Q.-T.; Nasrallah, C.; Lopes, J.; Bolla, J.-M.; Vivaudou, M.; Pagès, J.-M.; Colletier, J.-P. Providencia Stuartii Form Biofilms and Floating Communities of Cells That Display High Resistance to Environmental Insults. *PLOS ONE* **2017**, *12*, e0174213. https://doi.org/10.1371/journal.pone.0174213.
(119) Hiergeist, A.; Reischl, U.; Gessner, A. Multicenter Quality Assessment of 16S Ribosomal DNA-Sequencing for Microbiome Analyses Reveals High Inter-Center Variability. *Int. J. Med. Microbiol.* **2016**, *306* (5), 334–342. https://doi.org/10.1016/j.ijmm.2016.03.005.
(120) Mancabelli, L.; Milani, C.; Lugli, G. A.; Fontana, F.; Turroni, F.; van Sinderen, D.; Ventura, M. The Impact of Primer Design on Amplicon-Based Metagenomic Profiling Accuracy: Detailed Insights into Bifidobacterial Community Structure. *Microorganisms* **2020**, *8* (1). https://doi.org/10.3390/microorganisms8010131.
(121) Wasimuddin; Schlaeppi, K.; Ronchi, F.; Leib, S. L.; Erb, M.; Ramette, A. Evaluation of Primer Pairs for Microbiome Profiling across a Food Chain from Soils to Humans within the One Health Framework. *bioRxiv* **2019**, 843144. https://doi.org/10.1101/843144.
(122) Fredriksson, N. J.; Hermansson, M.; Wilén, B.-M. The Choice of PCR Primers Has Great Impact on Assessments of Bacterial Community Diversity and Dynamics in a Wastewater Treatment Plant. *PLOS ONE* **2013**, *8* (10), e76431. https://doi.org/10.1371/journal.pone.0076431.
(123) Zhang, B.; Xu, X.; Zhu, L. Activated Sludge Bacterial Communities of Typical Wastewater Treatment Plants: Distinct Genera Identification and Metabolic Potential Differential Analysis. *AMB Express* **2018**, *8*. https://doi.org/10.1186/s13568-018-0714-0.
(124) Szabó, E.; Liébana, R.; Hermansson, M.; Modin, O.; Persson, F.; Wilén, B.-M. Microbial Population Dynamics and Ecosystem Functions of Anoxic/Aerobic

Granular Sludge in Sequencing Batch Reactors Operated at Different Organic Loading Rates. *Front. Microbiol.* **2017**, *8*. https://doi.org/10.3389/fmicb.2017.00770.

(125) Gómez-Basurto, F.; Vital-Jácome, M.; Gómez-Acata, E. S.; Thalasso, F.; Luna-Guido, M.; Dendooven, L. Microbial Community Dynamics during Aerobic Granulation in a Sequencing Batch Reactor (SBR). *PeerJ* **2019**, *7*, e7152. https://doi.org/10.7717/peerj.7152.

(126) Kowalska-Duplaga, K.; Kapusta, P.; Gosiewski, T.; Sroka-Oleksiak, A.; Ludwig-Słomczyńska, A. H.; Wołkow, P. P.; Fyderek, K. Changes in the Intestinal Microbiota Are Seen Following Treatment with Infliximab in Children with Crohn's Disease. *J. Clin. Med.* **2020**, *9* (3). https://doi.org/10.3390/jcm9030687.

(127) Shchegolkova, N. M.; Krasnov, G. S.; Belova, A. A.; Dmitriev, A. A.; Kharitonov, S. L.; Klimina, K. M.; Melnikova, N. V.; Kudryavtseva, A. V. Microbial Community Structure of Activated Sludge in Treatment Plants with Different Wastewater Compositions. *Front. Microbiol.* **2016**, *7*. https://doi.org/10.3389/fmicb.2016.00090.

(128) Speirs, L. B. M.; Rice, D. T. F.; Petrovski, S.; Seviour, R. J. The Phylogeny, Biodiversity, and Ecology of the Chloroflexi in Activated Sludge. *Front. Microbiol.* **2019**, *10*. https://doi.org/10.3389/fmicb.2019.02015.

(129) Liu, J.; Li, J.; Tao, Y.; Sellamuthu, B.; Walsh, R. Analysis of Bacterial, Fungal and Archaeal Populations from a Municipal Wastewater Treatment Plant Developing an Innovative Aerobic Granular Sludge Process. *World J. Microbiol. Biotechnol.* **2017**, *33* (1), 14. https://doi.org/10.1007/s11274-016-2179-0.

(130) Davis, J. J.; Wattam, A. R.; Aziz, R. K.; Brettin, T.; Butler, R.; Butler, R. M.; Chlenski, P.; Conrad, N.; Dickerman, A.; Dietrich, E. M.; Gabbard, J. L.; Gerdes, S.; Guard, A.; Kenyon, R. W.; Machi, D.; Mao, C.; Murphy-Olson, D.; Nguyen, M.; Nordberg, E. K.; Olsen, G. J.; Olson, R. D.; Overbeek, J. C.; Overbeek, R.; Parrello, B.; Pusch, G. D.; Shukla, M.; Thomas, C.; VanOeffelen, M.; Vonstein, V.; Warren, A. S.; Xia, F.; Xie, D.; Yoo, H.; Stevens, R. The PATRIC Bioinformatics Resource Center: Expanding Data and Analysis Capabilities. *Nucleic Acids Res.* **2020**, *48* (D1), D606–D612. https://doi.org/10.1093/nar/gkz943.

(131) Meyer, F.; Paarmann, D.; D'Souza, M.; Olson, R.; Glass, E.; Kubal, M.; Paczian, T.; Rodriguez, A.; Stevens, R.; Wilke, A.; Wilkening, J.; Edwards, R. The Metagenomics RAST Server – a Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes. *BMC Bioinformatics* **2008**, *9* (1), 386. https://doi.org/10.1186/1471-2105-9-386.

(132) Shao, W.; Boltz, V. F.; Spindler, J. E.; Kearney, M. F.; Maldarelli, F.; Mellors, J. W.; Stewart, C.; Volfovsky, N.; Levitsky, A.; Stephens, R. M.; Coffin, J. M. Analysis of 454 Sequencing Error Rate, Error Sources, and Artifact Recombination for Detection of Low-Frequency Drug Resistance Mutations in HIV-1 DNA. *Retrovirology* **2013**, *10*, 18. https://doi.org/10.1186/1742-4690-10-18.

(133) Cline, J.; Braman, J. C.; Hogrefe, H. H. PCR Fidelity of Pfu DNA Polymerase and Other Thermostable DNA Polymerases. *Nucleic Acids Res.* **1996**, *24* (18), 3546–3551. https://doi.org/10.1093/nar/24.18.3546.

(134) Ling, L. L.; Keohavong, P.; Dias, C.; Thilly, W. G. Optimization of the Polymerase Chain Reaction with Regard to Fidelity: Modified T7, Taq, and Vent DNA Polymerases. *PCR Methods Appl.* **1991**, *1* (1), 63–69. https://doi.org/10.1101/gr.1.1.63.
(135) Potapov, V.; Ong, J. L. Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLOS ONE* **2017**, *12* (1), e0169774. https://doi.org/10.1371/journal.pone.0169774.
(136) Allam, A.; Kalnis, P.; Solovyev, V. Karect: Accurate Correction of Substitution, Insertion and Deletion Errors for next-Generation Sequencing Data. *Bioinformatics* **2015**, *31* (21), 3421–3428. https://doi.org/10.1093/bioinformatics/btv415.
(137) Schirmer, M.; D'Amore, R.; Ijaz, U. Z.; Hall, N.; Quince, C. Illumina Error Profiles: Resolving Fine-Scale Variation in Metagenomic Sequencing Data. *BMC Bioinformatics* **2016**, *17* (1), 125. https://doi.org/10.1186/s12859-016-0976-y.
(138) Masella, A. P.; Bartram, A. K.; Truszkowski, J. M.; Brown, D. G.; Neufeld, J. D. PANDAseq: Paired-End Assembler for Illumina Sequences. *BMC Bioinformatics* **2012**, *13* (1), 31. https://doi.org/10.1186/1471-2105-13-31.
(139) Martin, M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet.journal* **2011**, *17* (1), 10–12. https://doi.org/10.14806/ej.17.1.200.
(140) Ewing, B.; Hillier, L.; Wendl, M. C.; Green, P. Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome Res.* **1998**, *8* (3), 175–185. https://doi.org/10.1101/gr.8.3.175.
(141) Ewing, B.; Green, P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Res.* **1998**, *8* (3), 186–194. https://doi.org/10.1101/gr.8.3.186.
(142) Edgar, R. FASTQ files https://www.drive5.com/usearch/manual/fastq_files.html.
(143) Edgar, R. Quality (Phred) scores https://www.drive5.com/usearch/manual/quality_score.html.
(144) Why does the per base sequence quality decrease over the read in Illumina? https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina (accessed Mar 9, 2021).
(145) Illumina. *Illumina Sequencing by Synthesis*; 2016.
(146) Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*; 2019.
(147) Callahan, B. J.; McMurdie, P. J.; Rosen, M. J.; Han, A. W.; Johnson, A. J. A.; Holmes, S. P. DADA2: High Resolution Sample Inference from Illumina Amplicon Data. *Nat. Methods* **2016**, *13* (7), 581–583. https://doi.org/10.1038/nmeth.3869.
(148) Rognes, T.; Flouri, T.; Nichols, B.; Quince, C.; Mahé, F. VSEARCH: A Versatile Open Source Tool for Metagenomics. *PeerJ* **2016**, *4*, e2584. https://doi.org/10.7717/peerj.2584.
(149) Phred-scaled quality scores https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores (accessed Mar 5, 2021).
(150) Edgar, R. Calculating average Q (Phred) scores is a bad idea https://www.drive5.com/usearch/manual/avgq.html.

(151) Edgar, R. Expected errors predicted by Phred (Q) scores https://www.drive5.com/usearch/manual/exp_errs.html.
(152) Edgar, R. C.; Flyvbjerg, H. Error Filtering, Pair Assembly and Error Correction for next-Generation Sequencing Reads. *Bioinformatics* **2015**, *31* (21), 3476–3482. https://doi.org/10.1093/bioinformatics/btv401.
(153) Eren, A. M.; Morrison, H. G.; Lescault, P. J.; Reveillaud, J.; Vineis, J. H.; Sogin, M. L. Minimum Entropy Decomposition: Unsupervised Oligotyping for Sensitive Partitioning of High-Throughput Marker Gene Sequences. *ISME J.* **2015**, *9* (4), 968–979. https://doi.org/10.1038/ismej.2014.195.
(154) Huttenhower, C.; Gevers, D.; Knight, R.; Abubucker, S.; Badger, J. H.; Chinwalla, A. T.; Creasy, H. H.; Earl, A. M.; FitzGerald, M. G.; Fulton, R. S.; Giglio, M. G.; Hallsworth-Pepin, K.; Lobos, E. A.; Madupu, R.; Magrini, V.; Martin, J. C.; Mitreva, M.; Muzny, D. M.; Sodergren, E. J.; Versalovic, J.; Wollam, A. M.; Worley, K. C.; Wortman, J. R.; Young, S. K.; Zeng, Q.; Aagaard, K. M.; Abolude, O. O.; Allen-Vercoe, E.; Alm, E. J.; Alvarado, L.; Andersen, G. L.; Anderson, S.; Appelbaum, E.; Arachchi, H. M.; Armitage, G.; Arze, C. A.; Ayyaz, T.; Baker, C. C.; Begg, L.; Belachew, T.; Bhonagiri, V.; Bihan, M.; Blaser, M. J.; Bloom, T.; Bonazzi, V.; Paul Brooks, J.; Buck, G. A.; Buhay, C. J.; Busam, D. A.; Campbell, J. L.; Canon, S. R.; Cantarel, B. L.; Chain, P. S. G.; Chen, I.-M. A.; Chen, L.; Chhibba, S.; Chu, K.; Ciulla, D. M.; Clemente, J. C.; Clifton, S. W.; Conlan, S.; Crabtree, J.; Cutting, M. A.; Davidovics, N. J.; Davis, C. C.; DeSantis, T. Z.; Deal, C.; Delehaunty, K. D.; Dewhirst, F. E.; Deych, E.; Ding, Y.; Dooling, D. J.; Dugan, S. P.; Michael Dunne, W.; Scott Durkin, A.; Edgar, R. C.; Erlich, R. L.; Farmer, C. N.; Farrell, R. M.; Faust, K.; Feldgarden, M.; Felix, V. M.; Fisher, S.; Fodor, A. A.; Forney, L. J.; Foster, L.; Di Francesco, V.; Friedman, J.; Friedrich, D. C.; Fronick, C. C.; Fulton, L. L.; Gao, H.; Garcia, N.; Giannoukos, G.; Giblin, C.; Giovanni, M. Y.; Goldberg, J. M.; Goll, J.; Gonzalez, A.; Griggs, A.; Gujja, S.; Kinder Haake, S.; Haas, B. J.; Hamilton, H. A.; Harris, E. L.; Hepburn, T. A.; Herter, B.; Hoffmann, D. E.; Holder, M. E.; Howarth, C.; Huang, K. H.; Huse, S. M.; Izard, J.; Jansson, J. K.; Jiang, H.; Jordan, C.; Joshi, V.; Katancik, J. A.; Keitel, W. A.; Kelley, S. T.; Kells, C.; King, N. B.; Knights, D.; Kong, H. H.; Koren, O.; Koren, S.; Kota, K. C.; Kovar, C. L.; Kyrpides, N. C.; La Rosa, P. S.; Lee, S. L.; Lemon, K. P.; Lennon, N.; Lewis, C. M.; Lewis, L.; Ley, R. E.; Li, K.; Liolios, K.; Liu, B.; Liu, Y.; Lo, C.-C.; Lozupone, C. A.; Dwayne Lunsford, R.; Madden, T.; Mahurkar, A. A.; Mannon, P. J.; Mardis, E. R.; Markowitz, V. M.; Mavromatis, K.; McCorrison, J. M.; McDonald, D.; McEwen, J.; McGuire, A. L.; McInnes, P.; Mehta, T.; Mihindukulasuriya, K. A.; Miller, J. R.; Minx, P. J.; Newsham, I.; Nusbaum, C.; O'Laughlin, M.; Orvis, J.; Pagani, I.; Palaniappan, K.; Patel, S. M.; Pearson, M.; Peterson, J.; Podar, M.; Pohl, C.; Pollard, K. S.; Pop, M.; Priest, M. E.; Proctor, L. M.; Qin, X.; Raes, J.; Ravel, J.; Reid, J. G.; Rho, M.; Rhodes, R.; Riehle, K. P.; Rivera, M. C.; Rodriguez-Mueller, B.; Rogers, Y.-H.; Ross, M. C.; Russ, C.; Sanka, R. K.; Sankar, P.; Fah Sathirapongsasuti, J.; Schloss, J. A.; Schloss, P. D.; Schmidt, T. M.; Scholz, M.; Schriml, L.; Schubert, A. M.; Segata, N.; Segre, J. A.; Shannon, W. D.; Sharp, R. R.; Sharpton, T. J.; Shenoy, N.; Sheth, N. U.; Simone, G. A.; Singh, I.; Smillie, C. S.; Sobel, J. D.; Sommer, D.

D.; Spicer, P.; Sutton, G. G.; Sykes, S. M.; Tabbaa, D. G.; Thiagarajan, M.; Tomlinson, C. M.; Torralba, M.; Treangen, T. J.; Truty, R. M.; Vishnivetskaya, T. A.; Walker, J.; Wang, L.; Wang, Z.; Ward, D. V.; Warren, W.; Watson, M. A.; Wellington, C.; Wetterstrand, K. A.; White, J. R.; Wilczek-Boney, K.; Wu, Y.; Wylie, K. M.; Wylie, T.; Yandava, C.; Ye, L.; Ye, Y.; Yooseph, S.; Youmans, B. P.; Zhang, L.; Zhou, Y.; Zhu, Y.; Zoloth, L.; Zucker, J. D.; Birren, B. W.; Gibbs, R. A.; Highlander, S. K.; Methé, B. A.; Nelson, K. E.; Petrosino, J. F.; Weinstock, G. M.; Wilson, R. K.; White, O.; The Human Microbiome Project Consortium. Structure, Function and Diversity of the Healthy Human Microbiome. *Nature* **2012**, *486* (7402), 207–214. https://doi.org/10.1038/nature11234.

(155) Ravel, J.; Gajer, P.; Abdo, Z.; Schneider, G. M.; Koenig, S. S. K.; McCulle, S. L.; Karlebach, S.; Gorle, R.; Russell, J.; Tacket, C. O.; Brotman, R. M.; Davis, C. C.; Ault, K.; Peralta, L.; Forney, L. J. Vaginal Microbiome of Reproductive-Age Women. *Proc. Natl. Acad. Sci.* **2011**, *108* (Supplement 1), 4680–4687. https://doi.org/10.1073/pnas.1002611107.

(156) Caporaso, J. G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F. D.; Costello, E. K.; Fierer, N.; Peña, A. G.; Goodrich, J. K.; Gordon, J. I.; Huttley, G. A.; Kelley, S. T.; Knights, D.; Koenig, J. E.; Ley, R. E.; Lozupone, C. A.; McDonald, D.; Muegge, B. D.; Pirrung, M.; Reeder, J.; Sevinsky, J. R.; Turnbaugh, P. J.; Walters, W. A.; Widmann, J.; Yatsunenko, T.; Zaneveld, J.; Knight, R. QIIME Allows Analysis of High-Throughput Community Sequencing Data. *Nat. Methods* **2010**, *7* (5), 335–336. https://doi.org/10.1038/nmeth.f.303.

(157) OTU picking strategies in QIIME http://qiime.org/tutorials/otu_picking.html (accessed Mar 10, 2021).

(158) Bolyen, E.; Rideout, J. R.; Dillon, M. R.; Bokulich, N. A.; Abnet, C. C.; Al-Ghalith, G. A.; Alexander, H.; Alm, E. J.; Arumugam, M.; Asnicar, F.; Bai, Y.; Bisanz, J. E.; Bittinger, K.; Brejnrod, A.; Brislawn, C. J.; Brown, C. T.; Callahan, B. J.; Caraballo-Rodríguez, A. M.; Chase, J.; Cope, E. K.; Da Silva, R.; Diener, C.; Dorrestein, P. C.; Douglas, G. M.; Durall, D. M.; Duvallet, C.; Edwardson, C. F.; Ernst, M.; Estaki, M.; Fouquier, J.; Gauglitz, J. M.; Gibbons, S. M.; Gibson, D. L.; Gonzalez, A.; Gorlick, K.; Guo, J.; Hillmann, B.; Holmes, S.; Holste, H.; Huttenhower, C.; Huttley, G. A.; Janssen, S.; Jarmusch, A. K.; Jiang, L.; Kaehler, B. D.; Kang, K. B.; Keefe, C. R.; Keim, P.; Kelley, S. T.; Knights, D.; Koester, I.; Kosciolek, T.; Kreps, J.; Langille, M. G. I.; Lee, J.; Ley, R.; Liu, Y.-X.; Loftfield, E.; Lozupone, C.; Maher, M.; Marotz, C.; Martin, B. D.; McDonald, D.; McIver, L. J.; Melnik, A. V.; Metcalf, J. L.; Morgan, S. C.; Morton, J. T.; Naimey, A. T.; Navas-Molina, J. A.; Nothias, L. F.; Orchanian, S. B.; Pearson, T.; Peoples, S. L.; Petras, D.; Preuss, M. L.; Pruesse, E.; Rasmussen, L. B.; Rivers, A.; Robeson, M. S.; Rosenthal, P.; Segata, N.; Shaffer, M.; Shiffer, A.; Sinha, R.; Song, S. J.; Spear, J. R.; Swafford, A. D.; Thompson, L. R.; Torres, P. J.; Trinh, P.; Tripathi, A.; Turnbaugh, P. J.; Ul-Hasan, S.; van der Hooft, J. J. J.; Vargas, F.; Vázquez-Baeza, Y.; Vogtmann, E.; von Hippel, M.; Walters, W.; Wan, Y.; Wang, M.; Warren, J.; Weber, K. C.; Williamson, C. H. D.; Willis, A. D.; Xu, Z. Z.; Zaneveld, J. R.; Zhang, Y.; Zhu, Q.; Knight, R.; Caporaso, J. G. Reproducible, Interactive, Scalable and Extensible Microbiome Data Science

Using QIIME 2. *Nat. Biotechnol.* **2019**, *37* (8), 852–857. https://doi.org/10.1038/s41587-019-0209-9.
(159) Bokulich, N. A.; Kaehler, B. D.; Rideout, J. R.; Dillon, M.; Bolyen, E.; Knight, R.; Huttley, G. A.; Gregory Caporaso, J. Optimizing Taxonomic Classification of Marker-Gene Amplicon Sequences with QIIME 2's Q2-Feature-Classifier Plugin. *Microbiome* **2018**, *6* (1), 90. https://doi.org/10.1186/s40168-018-0470-z.
(160) Nearing, J. T.; Douglas, G. M.; Comeau, A. M.; Langille, M. G. I. Denoising the Denoisers: An Independent Evaluation of Microbiome Sequence Error-Correction Approaches. *PeerJ* **2018**, *6*. https://doi.org/10.7717/peerj.5364.
(161) Edgar, R. C. *UCHIME2: Improved Chimera Prediction for Amplicon Sequencing*; preprint; Bioinformatics, 2016. https://doi.org/10.1101/074252.
(162) Balvočiūtė, M.; Huson, D. H. SILVA, RDP, Greengenes, NCBI and OTT — How Do These Taxonomies Compare? *BMC Genomics* **2017**, *18* (2), 114. https://doi.org/10.1186/s12864-017-3501-4.
(163) The Greengenes Database http://greengenes.secondgenome.com/?prefix=downloads/greengenes_database/gg_13_5/ (accessed Mar 11, 2021).
(164) Pruesse, E.; Quast, C.; Knittel, K.; Fuchs, B. M.; Ludwig, W.; Peplies, J.; Glöckner, F. O. SILVA: A Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB. *Nucleic Acids Res.* **2007**, *35* (21), 7188–7196. https://doi.org/10.1093/nar/gkm864.
(165) Amazon EC2 C5 Instances — Amazon Web Services (AWS) https://aws.amazon.com/ec2/instance-types/c5/ (accessed Feb 24, 2021).
(166) Aslett, L. RStudio Server Amazon Machine Image (AMI) - Louis Aslett https://www.louisaslett.com/RStudio_AMI/ (accessed Feb 24, 2021).

BIOGRAPHY

Alison Gomeiz received her Bachelor of Science from George Mason University in 2019. She was employed as a research assistant at George Mason University as an undergraduate under Dr. Gregory Foster in environmental chemistry. She later worked for Dr. Benoit Van Aken in plant toxicology and bacterial ecology as an undergraduate. As a graduate student, she continued her research under Dr. Benoit Van Aken in bioinformatics, metagenomics, and microbial genetics studies.