UNIVERSITY OF CALIFORNIA

Santa Barbara

Developing a systems biology framework to engineer anaerobic gut fungi

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Chemical Engineering

by

St. Elmo Wilken

Committee in charge:

Professor Michelle A. O'Malley, Co-chair

Professor Linda R. Petzold, Co-chair

Professor M. Scott Shell

Professor Elizabeth G. Wilbanks

September 2020

The dissertation of St. Elmo Wilken is approved.

_____

Elizabeth G. Wilbanks

_____

M. Scott Shell

_____

Linda R. Petzold, Committee co-chair

_____

Michelle A. O'Malley, Committee co-chair

September 2020

Developing a systems biology framework to engineer anaerobic gut fungi

by

St. Emo Wilken

ACKNOWLEDGEMENTS

First, I would like to acknowledge and thank God for giving me the opportunity and ability to pursue my Ph. D. at UCSB.

Next, I would like to thank my advisors, Michelle O'Malley and Linda Petzold for their guidance, support and encouragement in completing the work that went into this thesis. They always supported me in my desire to explore new avenues of inquiry and were patient when things did not work out as expected. The Dow Discovery Fellowship was also instrumental in the research freedom I had, and I gratefully acknowledge its support. I would also like to acknowledge the entire O'Malley lab, but specifically Susanna Seppälä and Tom Lankiewicz for their many insightful discussions about biology and their patience in explaining it to an engineer. I would also like to thank Jon Monk for guiding me through the process of developing the first genome-scale model of a decidedly non-model anaerobic gut fungus.

I would not have been able to complete my Ph. D. without the support of my family, and specifically my parents and brother. They were always there for me, ready to listen and provided much needed emotional support as I completed my degree so far away from home. I would also like to acknowledge my friends, and specifically George Degen, for all the fun times we had in California. Most of all I want to thank my wife, Sina, for all her support and encouragement. She endured three years of a long-distance relationship, as well as all the highs and lows associated with completing a Ph. D., all the while being a constant source of strength. Thank you for everything you did for me!

VITA OF St. Elmo Wilken

September 2020

EDUCATION

2015 – 2020	Doctor of Philosophy in Chemical Engineering University of California, Santa Barbara
2017 – 2020	Graduate Program in Management Practice Certificate University of California, Santa Barbara
2014 – 2015	Master of Chemical Engineering, Control Engineering The University of Pretoria, Pretoria With distinction
2013 – 2014	Bachelor of Science Honors, Applied Mathematics The University of Pretoria, Pretoria With distinction
2012 – 2013	Bachelor of Engineering Honors, Control Engineering The University of Pretoria, Pretoria With distinction
2009 – 2012	Bachelor of Engineering, Chemical Engineering The University of Pretoria, Pretoria With distinction

PROFESSIONAL EMPLOYMENT

2015 – 2020	Graduate Research Fellow and Teaching Assistant Department of Chemical Engineering, UC Santa Barbara Research Advisors: Michelle A. O'Malley & Linda R. Petzold
2013 – 2015	Graduate Researcher Department of Chemical Engineering Research Advisors: Carl Sandrock & Pieter J. de Villiers
2011	Chemical Engineering Intern Sublime Technologies, South Africa
2010	Chemical Engineering Intern Nuclear Energy Company of South Africa, South Africa

PUBLICATIONS

1. **St. Elmo Wilken**, Patrick Leggieri, Corey Kerdman-Andrade, Matthew Reilly, Michael K. Theodorou, Michelle A. O'Malley, An Arduino based Automatic Pressure Evaluation System (A-APES) to quantify growth of non-model anaerobes in culture, *AIChe Journal* (2020)
2. **St. Elmo Wilken**, Susanna Seppälä, Thomas S. Lankiewicz, Mohan Saxena, John K. Henske, Asaf A. Salamov, Igor V. Grigoriev, Michelle A. O'Malley, Genomic and proteomic biases inform metabolic engineering strategies for anaerobic fungi, *Metabolic Engineering Communications* (2019)
3. Sean P. Gilmore, Thomas S. Lankiewicz, **St. Elmo Wilken**, Jennifer L. Brown, Jessica A. Sexton, John K. Henske, Michael K. Theodorou, David L. Valentine, Michelle A. O'Malley, Top-down enrichment guides in formation of synthetic microbial consortia for biomass degradation, *ACS Synthetic Biology* (2019)
4. **St. Elmo Wilken**, Candice L. Swift, Igor A. Podolsky, Susanna Seppälä, Michelle A. O'Malley, Linking "omics" to function unlocks the biotech potential of non-model fungi, *Current Opinion in Systems Biology* (2019)
5. **St. Elmo Wilken**, Mohan Saxena, Linda R. Petzold, Michelle A. O'Malley, *In Silico* Identification of Microbial Partners to Form Consortia with Anaerobic Fungi, *Processes* (2018)
6. John K. Henske, **St. Elmo Wilken**, Kevin V. Solomon, Chuck R. Smallwood, Vaithiyalingam Shutthanandan, James E. Evans, Michael K. Theodorou, Michelle A. O'Malley, Metabolic characterization of anaerobic fungi provides a path forward for two-stage bioprocessing of crude lignocellulose, *Biotechnology and Bioengineering* (2017)
7. Susanna Seppälä*, **St. Elmo Wilken***, Doriv Knop, Kevin V. Solomon, Michelle A. O'Malley, The importance of sourcing enzymes from non-conventional fungi for metabolic engineering & biomass breakdown, *Metabolic Engineering* (2017) ***equal author contributions**

PUBLICATIONS IN PREPARATION OR IN REVISION

1. Xuefeng Peng, **St. Elmo Wilken,** Thomas S. Lankiewicz, Sean P. Gilmore, Jennifer L. Brown, John K. Henske, Candice L. Swift, Asaf Salamov, Kerrie Barry, Igor V. Grigoriev, Michael K. Theodorou, David L. Valentine, Michelle A. O'Malley, Sculpting gut microbial communities alters fermentation products and methane release, *In revision at Nature Microbiology*
2. **St. Elmo Wilken**, Jonathan M. Monk, Patrick A. Leggieri, Christopher Lawson, Thomas S. Lankiewicz, Susanna Seppälä, Stephen J. Mondo, Kerrie W. Barry, Igor V. Grigoriev, John K. Henske, Michael K. Theodorou, Bernhard O. Palsson, Linda R. Petzold, Michelle A. O'Malley, Experimentally validated reconstruction and analysis of a genome-scale metabolic model of an anaerobic Neocallimastigomycota fungus, *In preparation for mSystems Journal*

SELECTED ORAL AND POSTER PRESENTATIONS

1. **St. Elmo Wilken**, Linda R. Petzold, Michelle A. O'Malley (2019). Developing a genome-scale model of Neocallimastigomycota fungi as a platform for metabolic engineering. Graduate Student Symposium at the University of California Santa Barbara, October 4th. Santa Barbara, California, USA. (**Best speaker award**)
2. **St. Elmo Wilken**, Linda R. Petzold, Michelle A. O'Malley (2018). Incorporating Flux Sampling into a Minimal Assumption Dynamic Flux Balance Analysis Algorithm. 5th Constraint-Based Reconstruction and Analysis Conference, October 14th – 16th. Seattle, Washington, USA.
3. **St. Elmo Wilken**, John K. Henske, Francis Cunningham, Michelle A. O'Malley (2017). Coarse Grained Model Development of Fungal Enzymatic Lignocellulosic Biomass Degradation. American Chemical Society Biotechnology Division, April 2nd – 6th. San Francisco, California, USA.

SELECTED HONORS AND AWARDS

2017-2020	Dow Discovery fellowship award
2019	Best speaker award at UCSB's chemical engineering graduate student symposium
2017	Distinguished service award from the University of California Santa Barbara's chemical engineering department
2014	South African Mathematical Society's award for the best honors student at the University of Pretoria
2014	Allan Gray Academic Achiever award
2012	Best undergraduate research project award from the University of Pretoria's chemical engineering department

ABSTRACT

Developing a systems biology framework to engineer anaerobic gut fungi

by

St. Elmo Wilken

Lignocellulose is a complex, energy-rich heterogenous polymer composed of cellulose (40-50%), hemicellulose (20-40%) and lignin (20-35%). Using the more than 1.6 billion tons of agricultural lignocellulosic waste generated worldwide each year is a promising avenue to explore for sustainable bioprocessing. While lignocellulose is abundant, lignin acts as a protective barrier that prevents its decomposition into fermentable sugars and poses significant challenges for the utilization of this energy-rich resource in biotechnological applications. On the other hand, anaerobic gut fungi specialize in bio-converting unpretreated lignocellulose into fermentable sugar monomers and represent a promising opportunity to exploit lignocellulosic plant biomass for bioprocesses.

Anaerobic gut fungi typically inhabit the digestive tracts of herbivores where they play an integral role in the decomposition of raw lignocellulose into its constitutive sugar monomers. The genomes of these fungi encode for the highest diversity, and largest number, of lignocellulolytic enzymes of any sequenced fungus to date. This, in combination with their filamentous morphology, causes them to excel at lignocellulose decomposition. Despite these advantages, anaerobic gut fungi are not utilized in bioprocesses due to challenges in

cultivating and engineering them. In this thesis a marriage between experimental and multi-omic datasets is used to develop techniques that can be used to better understand and engineer anaerobic gut fungi for lignocellulose decomposition in upstream bioprocesses.

A comprehensive metagenomic enrichment of goat fecal pellets was undertaken to better understand how rumen microbiome-based cultures respond to different environmental stressors. The outsize role anaerobic gut fungi play in these systems was highlighted by the markedly different way in which lignocellulosic carbon was metabolized to different fermentation products when anaerobic fungi were present. Overall, the analysis elucidated a natural compartmentalization that occurs between anaerobes during the degradation of lignocellulose, suggesting design rules that can be used to funnel carbon to different end-products based on the composition of the microbial consortia.

Despite the importance of anaerobic gut fungi in lignocellulolytic systems, no stable genetic engineering tools have been developed for this class of fungi to facilitate strain optimization. This is partially due to several unique genomic traits possessed by the gut fungi, namely an extreme bias towards AT bases in their genomes, as well as a disproportionate abundance of repetitive genomic regions. By making use of omics databases, the consequences of these features were investigated. It was found that the carbohydrate active enzymes encoded for by the gut fungi are likely heavily glycosylated, which has ramifications for heterologous expression strategies. A novel codon optimization table was also introduced to facilitate the quest to genetically engineer these fungi.

Genome-scale models form a cornerstone of modern metabolic engineering strategies, due to their ability to accurately model the metabolism of an organism from first principles. These models are most often made for well characterized organisms, but there is great benefit in

constructing such a model of an anaerobic gut fungus due to its ability to act as a scaffold to build understanding upon. To this end, the first genome-scale model of an anaerobic gut fungus was constructed using a combination of experimental and omics data. This model captures the primary metabolism of *Neocallimastix lanati*, a novel anaerobic gut fungus isolate, and sheds light on the inner workings of the carbon metabolism unique to the gut fungi.

Furthermore, genome-scale models can also be used to predict growth rate characteristics of organisms in *silico*. This aspect can be particularly useful when screening microbes for the development of stable consortia with the anaerobic gut fungi. A novel dynamic flux balance analysis algorithm, specifically geared towards the anaerobic gut fungi, was developed for this purpose. It was found that methanogens are likely the best partners due to their ability to metabolize by-products of the gut fungi and not compete with them for resources.

Finally, due to challenges associated with cultivating the anaerobic gut fungi, indirect measurements are typically used to infer their growth rate. These usually take the form of pressure measurements, performed with a digital handheld pressure transducer. While high resolution experiments afford the most insight into the impact of environmental perturbations on the growth rate of the fungi, these are very labor and time intensive. An automatic pressure measurement and venting device was designed and built to automate this process. Beyond the time savings afforded by the device, an extremely high-resolution growth curve can now be automatically constructed, shedding light on the growth dynamics of the anaerobic gut fungi.

In sum this thesis combines experimental and omics data to yield new insights in the behavior of anaerobic gut fungi and paves the way for their exploitation in biotechnology.

Table of contents

I. Introduction

1.1 Motivation and overview of the thesis

Lignocellulose is an abundant, renewable and largely underexploited natural resource (Sanderson, 2011). It is composed of the energy-rich polysaccharides cellulose and hemicellulose, which are protected from decomposition and utilization by microbes through the recalcitrant lignin that surrounds them. Bioconversion of the estimated 1.6 billion tons of agricultural lignocellulosic waste generated each year into value added chemicals is an important step towards a more sustainable chemical industry (Demirbas and Demirbas, 2010; Saini, Saini and Tewari, 2015). One of the primary obstacles in realizing this goal is the recalcitrance of the crude feedstock. Lignin, which acts as a barrier to enzymatic attack, has proven to be a significant obstacle in utilizing this resource to release fermentable sugars via enzymatic digestion. Moreover, the largest component of lignocellulose, by mass, is the energy rich glucose polymer cellulose, which by itself requires the simultaneous action of three distinct classes of enzymes to be converted into the metabolically accessible glucose monomer (Horn *et al.*, 2012). Expensive pretreatment is typically required to process crude lignocellulose into fermentable sugars for utilization by microorganisms.

On the other hand, nature already processes crude lignocellulose into biologically relevant compounds in the digestive tracts of ruminants (Weimer, Russell and Muck, 2009). In this thesis various omics[1] and experimental datasets, derived from microbes found in the rumen of large herbivores, are coupled and modeled to investigate and engineer systems that could

[1] Genomic, transcriptomic etc. datasets are collectively referred to as “omics” data.

be used to improve the current methods of utilizing lignocellulose as a feedstock for bioprocessing. Specific attention is paid to anaerobic gut fungi, in the phylum Neocallimastigomycota, which are of central importance in the gut microbiota of herbivores. The following subsections of Chapter I continue with more detailed background information regarding the subsequent chapters of this thesis. The rest of the thesis is outlined below.

The biotechnological promise of ruminant inspired systems for the bioconversion of lignocellulose is vast. Indeed, several recent studies have made use of modern sequencing techniques to elucidate this potential (Seshadri *et al.*, 2018; Stewart *et al.*, 2018). However, a major drawback of these studies is that they do not focus on the class of organisms primarily responsible for the lignocellulolytic action of ruminants: anaerobic gut fungi in the clade Neocallimastigomycota (Gruninger *et al.*, 2014). Chapter II addresses these shortcomings in literature by introducing a metagenome study that highlights the outsized impact eukaryotes have on the degradation of lignocellulose in ruminant herbivores. In sum the study reveals consortia membership design rules that can be used to channel lignocellulose into biotechnologically relevant end-products.

While a comprehensive database of cellulolytic sequences is a prerequisite first step in understanding ruminant based microorganisms, it is also necessary to be able to genetically engineer such systems. While robust genetic engineering techniques are under development, Chapter III introduces a systems level analysis aimed at using the currently available omics data to facilitate the development of genetic engineering techniques, both native and heterologous, for anaerobic gut fungi. It was found that their carbohydrate active enzymes possess a disproportionate abundance of repetitive elements and are likely heavily glycosylated, which has consequences for heterologous expression strategies. Additionally,

the first codon optimization table for an anaerobic gut fungus was also developed, guiding the way for future genetic engineering strategies.

Anaerobic gut fungi are central to unlocking the lignocellulolytic capabilities inherent to the ruminant microbiome. While troves of omics data have been collected, a systematic framework to synthesize this data to better understand the gut fungal metabolism is sorely lacking. Genome-scale models[2] can be constructed from such omics datasets and can be used to guide engineering efforts and elucidate metabolic features unique to an organism. Chapter IV introduces the first genome-scale metabolic model of an anaerobic gut fungus. This model captures the primary metabolism of *Neocallimastix lanati*, a novel anaerobic gut fungal isolate. The model highlights the metabolic degeneracy of *N. lanati*, a feature most likely used to optimally regulate its metabolism under different environmental conditions. Furthermore, the model is also well suited to focusing experimental effort on the areas that are most important to understand, and subsequently engineer, for bioprocessing.

Furthermore, the power of a mechanistically predictive model is that it can be used to accurately simulate cellular responses to perturbations, saving experimental time and effort. Chapter V introduces an algorithm that is specifically designed to interrogate genome-scale models of organisms in culture conditions typically used for rumen microbiome derived systems. This allows for the rapid screening of microbes that are more likely to result in stable pairings with the anaerobic gut fungi. These pairings can then be further investigated experimentally for bioprocessing applications.

[2] Genome-scale models are a mathematical representation of the metabolism of an organism.

Due to the unique culturing conditions required by rumen derived microorganisms, batch cultivation is a necessity. Moreover, only a few techniques exist to non-invasively measure growth, with intermittently measured gas production rate measurements being the standard. This type of manual measurement is typically time intensive and low resolution. Chapter VI describes the development and construction of a device that automatically measures the gas production rate as a proxy for fungal growth, requiring minimal manual oversight, and is capable of continuously monitoring the growth rate of a culture in high resolution.

Finally, Chapter VII concludes the thesis and points to future opportunities to exploit the capabilities of rumen microbiome inspired biotechnology. Subsequent portions of this motivating chapter are based on my review papers "The importance of sourcing enzymes from non-conventional fungi for metabolic engineering & biomass breakdown" (Metabolic Engineering, 2017) and "Linking 'omics' to function unlocks the biotech potential of non-model fungi" (Current Opinion in Systems Biology, 2019).

1.2 Fungi are a rich resource for sustainably utilizing lignocellulosic feedstocks for biotechnology

Modern biotechnology uses enzymes and engineered microbes to produce a wide variety of fuels, materials and chemicals from renewable feedstocks (Otero and Nielsen, 2010). In contrast, current commodity- and fine-chemical production relies on non-renewable petroleum feedstocks. Given the dwindling resources and the heavy carbon footprint of oil, the demand for environmentally friendly alternatives is urgent and ever increasing.

The so-called first generation biofuels were derived from crops that are rich in starch and sugars, such as corn and sugarcane (Saini, Saini and Tewari, 2015). As the world's population is predicted to increase to ~10 billion by the year 2050, this approach is not sustainable because it competes with food resources and for agricultural land (Bothast and Schlicher, 2005; Rogers *et al.*, 2017). Current efforts seek to convert lignocellulosic energy crops and residues from agriculture and forestry into hexose and pentose sugars (Sanderson, 2011; U.S. Department of Energy, 2016, 2017). Given the impetus of the European Union's goal to develop a bio-economy by 2050, as well as the estimated €2 trillion bio-market size in 2012, there are significant political and financial drivers to pursue these endeavors (Scarlat *et al.*, 2015). To fully realize the potential of sustainable bioproduction platforms, there is a great need to identify novel organisms, enzymes and molecules with activities that can be harnessed for a range of breakdown and conversion applications (Curran and Alper, 2012; Adrio and Demain, 2014; Monciardini *et al.*, 2014; Rocha-Martin *et al.*, 2014; Thies *et al.*, 2016). In particular, it is necessary to be able to cost-effectively convert diverse, underutilized plant biomass into tailor-made value-added compounds.

Lignocellulosic biomass, available worldwide in plant cell walls, is arguably the most promising feedstock for the sustainable production of bio-based chemicals and value-added products (Himmel *et al.*, 2007; Rogers *et al.*, 2017; U.S. Department of Energy, 2017). Underutilized lignocellulosic feedstock is abundant – it is estimated that 1.6 billion tons of agricultural waste is generated on an annual basis worldwide (Sarkar *et al.*, 2012; Saini, Saini and Tewari, 2015). It has been suggested that the US alone could sustainably produce equally as much biomass that could be funneled into bioprocessing applications on an annual basis (Himmel *et al.*, 2007; Rogers *et al.*, 2017). However, the inherent recalcitrance of plant cell walls presents a formidable challenge for biotechnological applications. Few organisms can fully degrade the highly heterogeneous and recalcitrant structures found in plant cell walls. Therefore, biomass-degrading organisms are highly sought after, as their enzymes can be directly harvested or used for consolidated bioprocessing (Hess *et al.*, 2011; Piao *et al.*, 2014; Zhang *et al.*, 2016).

Fungi play a major role in nutrient cycling and biogeochemical cycles in both aquatic and terrestrial environments (Dighton, 2007; Gadd, 2007; Gessner *et al.*, 2007). Already, most industrial enzymes for lignocellulosic bioprocessing are sourced from fungi but the fungal kingdom is vast, largely hidden, and exhibits a wide range of interesting bioactivities that remain underexploited (Banerjee, Scott-Craig and Walton, 2010; Payne *et al.*, 2015; Ramanjaneyulu and Rajasekhar Reddy, 2016; Falade *et al.*, 2017).

While biomass-degrading enzymes can be found in both bacteria and fungi, nearly all industrial enzymes are sourced from fungi. This is likely because the fungal enzymes are often stabilized by glycosylation and they have been shown to be active in the presence of proteases and surfactants, and at high temperature (Hong *et al.*, 2001; Beckham *et al.*, 2012; Ilmberger,

2013). Consequently, fungi excel at biomass degradation in nature and possess a wide variety of enzymes that depolymerize plant biomass with high efficiency (Dighton, 2007). As shown in Table 1.1, fungal biomass-degrading enzymes are heavily used for the processing of paper and pulp; for the production of food, feed, pharmaceuticals and cosmetics; as well as for bioremediation.

Table 1.1: Various industrial applications of fungal enzymes, see (Susanna Seppälä *et al.*, 2017) for a full set of references.

Enzyme	EC number	Reaction	Applications
		Hydrolases	
Cellulase	3.2.1.4	Hydrolysing the β-1,4-glycosidic bonds in cellulose	Food industry, textile manufacturing, detergent industry, paper and pulp industry, bioremediation, biofuel production
Xylanase (Hemicellulase)	3.2.1.8	Hydrolysing the β 1,4-glycosidic bonds in xylan	Food industry, biofuel production, paper and pulp industry, deinking, production of animal feed
Alpha-amylase	3.2.1.1	Hydrolysing the α-1,4-glycosidic bonds in starch	Food industry, starch conversion, biofuel production, detergent industry, paper and pulp industry
Invertase	3.2.1.26	Hydrolysing sucrose into glucose and fructose	Food industry, cosmetics, pharmaceutical industry, paper industry
Beta-galactosidase; Lactase	3.2.1.23	Hydrolysing lactose into glucose and galactose	Food industry
Lipases	3.1.1.3	Total or partial hydrolysis of fats and oils	Food industry, biofuel production, spills, detergent industry, paper and pulp industry,

			pharmaceutical industry
Phytase	3.1.3.8	Catalyzing phosphate monoester hydrolysis of phytic acid	Food industry, agriculture, production of animal feed
		Oxidases	
Laccase	1.10.3.2	Catalyzing the one-electron oxidation of four reducing-substrate molecules concomitant with the four-electron reduction of molecular O_2 to H_2O	Nanotechnology, synthetic chemistry, bioremediation, cosmetics
		Peroxidases	
Lignin peroxidase	1.11.1.14	Catalyzing the oxidation of various organic and inorganic substrates in the presence of H_2O_2 as electron acceptor via long-range electron transfer (LRET)	Paper and pulp industry, textile industry, pharmaceutical industry, bioremediation, biomass conversion, cosmetics
Manganese peroxidase	1.11.1.13	Catalyzing the oxidation of Mn (II) to Mn (III), as well as a variety of low redox potential organic substrates, in the presence of H_2O_2 as electron acceptor	Paper and pulp industry, textile industry, pharmaceutical industry, bioremediation, biomass conversion
Versatile peroxidase	1.11.1.16	Catalyzing the oxidation of various high and low redox potential organic substrates in the presence of H_2O_2 as electron acceptor in either a manganese-mediated reaction or a manganese-independent reaction via LRET	Paper and pulp industry, textile industry, pharmaceutical industry, bioremediation, biomass conversion

1.3 Fungi excel at decomposing recalcitrant lignocellulose

Plant cell walls are complex and dynamic structures made mainly of cellulose (40-50%), hemicellulose (20-40%) and lignin (20-35%) (Houston *et al.*, 2016; Liao *et al.*, 2016), which together form a formidable barrier against chemical and enzymatic degradation, see Figure 1.1. Cellulose is an unbranched polymer of D-glucose moieties that are linked by β(1→4) bonds; the cellulose chains may contain thousands of glucose units and aggregate into crystalline microfibrils. In contrast, hemicelluloses are a heterogeneous group of branched polysaccharides composed of various 5- and 6-carbon sugars *e.g.* xylose, mannose, arabinose and galactose (Rubin, 2008). In plant cell walls, cellulose microfibrils are surrounded by a network of hemicelluloses. Further, the energy-rich cellulose and hemicellulose are encapsulated by lignin, which is a complex aromatic polymer resulting from the oxidative combinatorial coupling of p-coumaryl-, coniferlyl-, and sinapyl alcohols (Haghighi Mood *et al.*, 2013). In addition to providing structural support to the plant, the chemically recalcitrant lignin protects the cellulose and hemicellulose polymers from enzymatic hydrolysis and most microbial invaders.

Despite its recalcitrance, fungi have a natural advantage against crude biomass – they break it down both physically and enzymatically. For example, fungi may burrow into the biomass, increasing its surface area and making it more accessible to biomass-degrading enzymes from fungi as well as from other neighboring microbes, as shown in Figure 1.2. As fungi cannot take up all polymeric compounds from their environment, they secrete extracellular enzymes that degrade the polymers to short oligomers and monomers that are imported through targeted transporters and metabolized in the cells (Seppälä *et al.*, 2016).

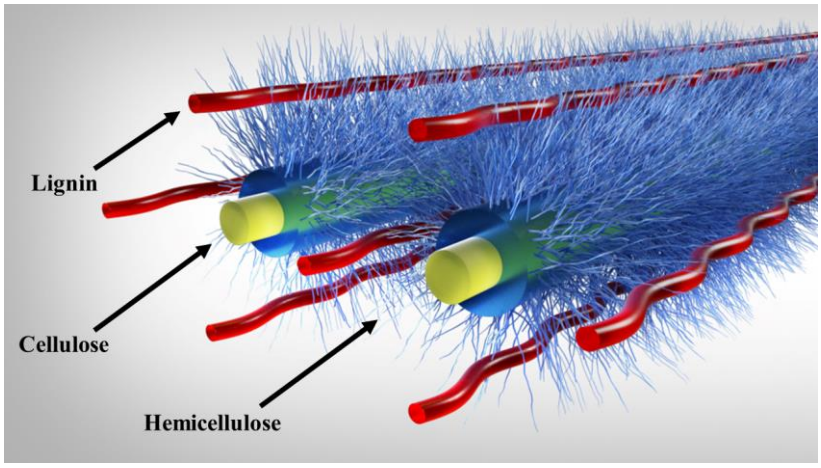

Figure 1.1: Qualitative diagram of lignocellulose composition. Cellulose is polymer of D-glucose linked by β-1,4-glucosidic bonds with a degree of polymerization of up to 10,000 or higher. Cellulose chains typically aggregate into crystalline fibrils of up to 36 chains. Hemicellulose is much more heterogenous and is composed of a variety of monomers (typically D-xylose or D-mannose). The degree of polymerization is usually significantly less than that of cellulose, around 200 or lower. Lignin is a complex polymer composed of phenyl propane units; it is the most abundant non-polysaccharide in lignocellulosic plant biomass (Jorgensen, Kristensen and Felby, 2007). Figure reproduced from (Susanna Seppälä *et al.*, 2017).

In order to accomplish this complex task, fungi harbor a variety of biomass degrading enzymes, where each enzyme excels in the biocatalysis of one (or more) specific compounds within the lignocellulosic structure (Dashtban, Schraft and Qin, 2009). Cellulose is degraded by glycoside hydrolases (GHs) that are either endocellulases or exocellulases: endocellulases bind anywhere along the length of the cellulose molecule and hydrolyze the β-1,4 glycosidic linkage, whereas exocellulases -cellodextrinases and cellobiohydrolases- bind at the ends of the cellulose polymer and release glucose or unit-length oligosaccharide products (Li, Chen and Ljungdahl, 1997). β-glucosidases break down cellooligosaccharides and cellobiose into glucose monomers (Sørensen *et al.*, 2013). Hemicellulose breakdown requires the added action of hemicellulases such as xylanases and mannanases (Bhattacharya, Bhattacharya and

Pletschke, 2015). Known mechanisms of fungal lignin degradation require laccases that couple the oxidation of substrates to the reduction of oxygen (Singh Arora and Kumar Sharma, 2010) and peroxidases that couple the oxidation of substrates to the reduction of hydrogen peroxide (Hofrichter *et al.*, 2010). The fungal biomass-degrading machinery typically consists of a cocktail of powerful cellulases, β-glucosidases, hemicellulases, and lignin-modifying enzymes that act synergistically.

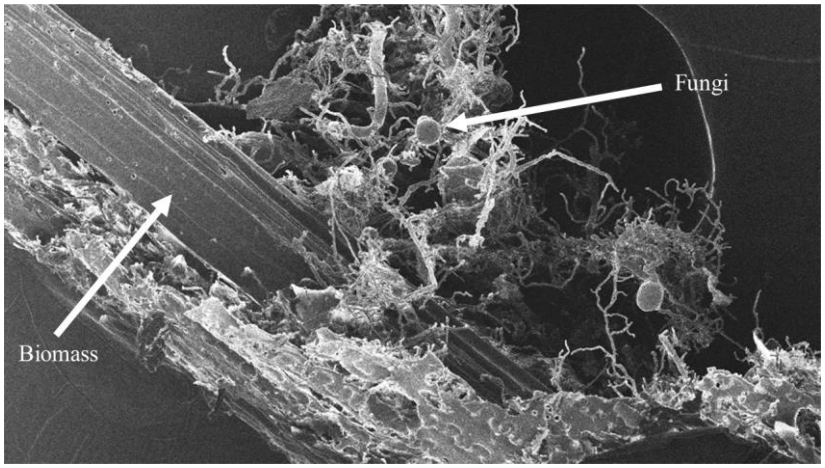

Figure 1.2: Micrograph showing invasive anaerobic fungal growth on lignocellulosic plant biomass. During the fungal growth cycle a root-like rhizoidal network is formed, which disrupts the biomass structure. The hyphae secrete the cellulolytic enzymes in close proximity to the substrate, greatly enhancing biomass degradation compared to the free diffusing enzyme system of aerobic fungi (Gilmore, Henske and O'Malley, 2015). Figure reproduced from (Susanna Seppälä *et al.*, 2017).

Given the ubiquitous interwoven enzymatic activities required for lignocellulose hydrolysis, the CAZy (carbohydrate-active enzymes) database (www.cazy.org) serves as an excellent resource that classifies (and updates) all known enzyme families involved in cellulolysis, hemicellulolysis, and, by a recent addition, the degradation of lignin (Levasseur et al., 2013). Apart from fungi, various bacterial genera have been observed to metabolize

lignin and shown to be competent to release ^{14}C-labeled CO_2 from labeled lignin (Kerr, Kerr and Benner, 1983; Kern and Kirk, 1987; Y. Chen et al., 2012; Brown and Chang, 2014). Although the bacterial catabolism of lignin is not as complete compared to fungal systems, it seems clear that bacteria can react with lignin and possibly produce smaller aromatics that can be imported into the cell for aromatic catabolism (Kanaly and Harayama, 2000; Chakraborty and Coates, 2004).

1.4 Current industrial fungal workhorses for biomass degradation have limitations

Industrial biomass conversion typically requires physical and chemical pretreatment to separate individual biopolymer constituents followed by enzymatic processing using, typically, fungal enzymes (Galbe and Zacchi, 2012). Physical pretreatment is resource intensive, and generally involves milling, or the utilization of hot steam to increase the surface area of the lignocellulosic biomass (Chandra *et al.*, 2007). Chemical pretreatment involves incubating the biomass with an acid or alkali (Sanderson, 2011; Davis *et al.*, 2013). After pretreatment, the biomass is hydrolyzed either by cellulolytic microbes or purified enzyme cocktails. As it has been estimated that the cost of enzymes make up a significant part of the cost of bioethanol production there is great interest to identify enzymes with increased degradation activity that can be produced at high titers (Dashtban, Schraft and Qin, 2009; Klein-Marcuschamer *et al.*, 2012; Liu, Zhang and Bao, 2016).

Currently, a handful of fungi are directly utilized on an industrial level for biomass hydrolysis. Foremost in this regard is the filamentous fungus *Trichoderma reesei*; other notable species include the filamentous *Aspergillus niger* and the thermophilic *Humicola insolens* (Sukumaran, Singhania and Pandey, 2005; Bischof, Ramoni and Seiboth, 2016;

Paloheimo *et al.*, 2016). Typically, these organisms are employed because they are natural hypersecreters of cellulolytic enzymes, and the secretion system of *T. reesei* has been subject to extensive investigation for decades (for recent reviews, see *e.g.* (Bischof, Ramoni and Seiboth, 2016; Paloheimo *et al.*, 2016)). A complication is that fungal enzyme production is typically subject to carbon catabolite repression and the molecular mechanisms behind these processes are complex and difficult to engineer (Amore, Giacobbe and Faraco, 2013).

Current enzyme yields stand at around 100 g/L, and it has been suggested that further improvements are likely to be modest (Banerjee, Scott-Craig and Walton, 2010). An alternative to improving enzyme yields is to improve the performance of the saccharolytic enzymes directly though protein engineering. For example, through directed evolution, the cellulases of *Hypocrea jecorina*, a teleomorph of *T. reesei*, were evolved to function optimally at 70°C instead of 60°C (Wu and Arnold, 2013). Also, recently the AA9 (formerly GH61) family was shown to greatly increase the cellulolytic capabilities of the cellulases secreted by *T. reesei* (Langston *et al.*, 2011).

Although the enzymes of *T. reesei* - and moreover its ability to secrete astonishing quantities of enzymes - remain useful for industrial applications, there are limits to what we can expect from this organism. Notably, it has become apparent that the genome of *T. reesei* encodes for the smallest diversity of cellulases and hemicellulases of any sequenced fungus capable of plant cell wall degradation (Martinez *et al.*, 2008). Further, these enzymes cannot act on lignin, which is generally separated from most industrial substrates and burned as an energy source (Kuhad, Gupta and Singh, 2011). Therefore, bioprospecting in other fungal clades for novel proteins that can be utilized for biomass breakdown is an appealing path forward.

1.5 Lesser known fungi could improve the efficiency of biomass deconstruction

Most fungi degrade cellulose and some, such as the white-rot fungus *Phanerochaete chrysosporium*, are able to efficiently depolymerize lignin (for a review, see (Chandel *et al.*, 2015). In recent years the advent of new biotechnological tools related to next-generation sequencing and 'omics' techniques enabled the discovery of new biomass degradation enzymes in the more basal clades (Floudas *et al.*, 2012; Youssef *et al.*, 2013; Riley *et al.*, 2014; Haitjema *et al.*, 2017a). It has become apparent that anaerobic fungi, in the early diverging phylum Neocallimastigomycota, possess a largely untapped source of biomass degrading enzymes (Solomon *et al.*, 2016; Haitjema *et al.*, 2017a). A recent comprehensive look at transcriptomic data collected from three gut fungal strains suggested that the anaerobic fungi are prodigious producers of glycoside hydrolases (Solomon *et al.*, 2016), and further may possess novel receptors and transporters for carbohydrates that hold biotechnological promise, as shown in Figure 1.3 (Seppälä *et al.*, 2016) .

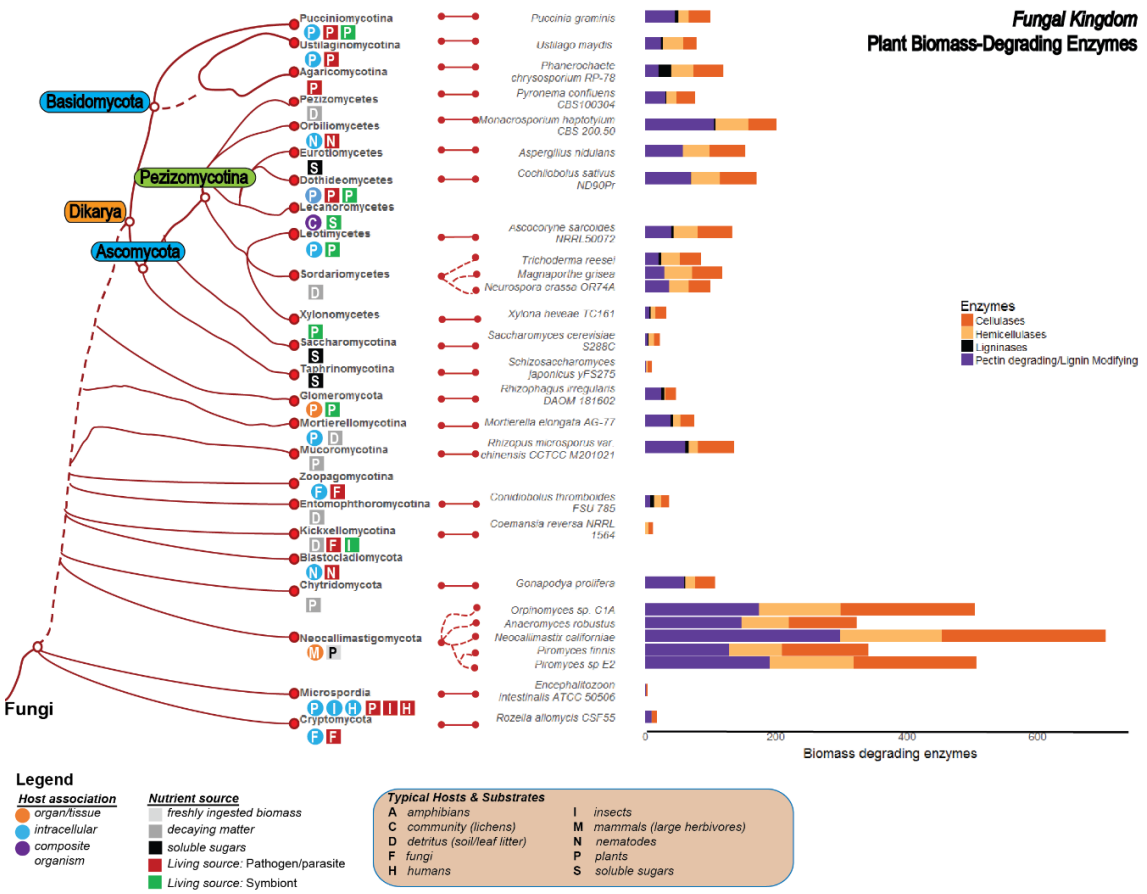

Figure 1.3: Plant biomass-degrading enzyme content & diversity in the fungal kingdom. Genomes of 27 fungi presenting different divisions of fungal tree of life revealed that the genomes of gut fungi contain the largest number of biomass-degrading enzymes, mainly cellulases (orange), hemicellulases (yellow) and pectin degrading/lignin modifying enzymes (purple) (Goffeau *et al.*, 1996; Galagan *et al.*, 2003; Dean *et al.*, 2005; Kämper *et al.*, 2006; Martinez *et al.*, 2008; Arnaud *et al.*, 2010; Corradi *et al.*, 2010; Rhind *et al.*, 2011; Duplessis *et al.*, 2011; Ohm *et al.*, 2012, 2014; Crous *et al.*, 2012; Gianoulis *et al.*, 2012; James *et al.*, 2013; Meerupati *et al.*, 2013; Tisserant *et al.*, 2013; Traeger *et al.*, 2013; Wang *et al.*, 2013; Youssef *et al.*, 2013; Grigoriev *et al.*, 2014; Chang *et al.*, 2015; Solomon *et al.*, 2016; Gazis *et al.*, 2016; Haitjema *et al.*, 2017a; Uehling *et al.*, 2017). The tree was modified from Mycocosm (Grigoriev *et al.*, 2014). Figure reproduced from (Susanna Seppälä *et al.*, 2017).

1.6 Anaerobic gut fungi are an untapped resource that excel at biomass deconstruction

An inviting alternative to current methods of biomass breakdown is to find superior biomass-degrading enzymes from natural sources where crude biomass degradation is

abundant. Anaerobic gut fungi inhabit the intestines of a range of large herbivores, from cows, horses and sheep to elephants, camels and even iguanas (Hibbett *et al.*, 2007; Liggenstoffer *et al.*, 2010). In spite of their key role in conversion and recycling of photosynthetically fixed plant biomass, research on anaerobic gut fungi did not gain steam until the late 1970s, when Colin Orpin established that the 'protozoan flagellates' that were found in rumen fluid are in fact fungi (Orpin, 1975, 1977). Even today, anaerobic gut fungi remain understudied, primarily owing to challenges in isolation and in maintaining the fungi in culture (Haitjema *et al.*, 2014). However, powerful integrated 'omics' approaches are rapidly filling the knowledge gaps (Mondo *et al.*, 2017; Youssef *et al.*, 2013; Solomon *et al.*, 2016; Haitjema *et al.*, 2017a).

Recent advances in next generation sequencing have revolutionized the study of non-conventional fungi, and anaerobic gut fungi in particular. In the past four years, researchers have increased the inventory of known anaerobic fungal enzymes from a mere handful to thousands (Youssef *et al.*, 2013; Solomon *et al.*, 2016). These enzymes have been validated with proteomics, and biochemical characterization has revealed that anaerobic fungal enzymes exhibit little substrate preference, in contrast to their later-diverging fungal cousins that strongly favor cellulose-rich substrates (Solomon *et al.*, 2016). This ability is a consequence of extensive horizontal gene transfer where the anaerobic fungi have complemented their cellulolytic capabilities with hemicellulases derived from bacteria (Haitjema *et al.*, 2017a). Thus, anaerobic fungi possess the largest and most diverse inventory of biomass-degrading enzymes among fungi sequenced to date (see Figure 1.3), and their enzyme setup is seemingly regulated by growth substrate so that the optimal enzyme cocktail is produced at all times (Solomon *et al.*, 2016). Similarly, sequencing-enabled discovery

recently revealed the elusive fungal cellulosomal complex that synergistically combines diverse biomass-degrading enzymes for efficient biomass degradation (Haitjema *et al.*, 2017a). With these rich and growing databases of characterized anaerobic fungal enzymes, we are now poised to enter a new phase of innovation where anaerobic fungal enzymes may be harnessed for bioengineering applications.

1.7 Fungal cellulosomes greatly enhance the lignocellulolytic capabilities of anaerobic gut fungi

Cellulosomes are multi-enzyme complexes, first described in the anaerobic bacterium *Clostridium thermocellum* (Lamed *et al.*, 1985), that tether plant biomass-degrading enzymes together for improved hydrolysis. While aerobic fungi release their biomass degrading enzymes to the external *milieu*, it was recently discovered that anaerobic gut fungi secrete fungal cellulosomes that bear some architectural resemblance to the bacterial cellulosomes (Guerriero *et al.*, 2015; Haitjema *et al.*, 2017a). The power of the cellulosomal system lies in its modular ability to host a variety of enzymes that can break down unpretreated lignocellulosic plant biomass synergistically (Resch *et al.*, 2013).

Although both fungal and bacterial cellulosomes digest plant cell wall material, there are significant differences between the kingdom-specific complexes. In addition to both the structural differences between the so-called dockerin and cohesin units, as well as their relative position on the cellulosome (Haitjema *et al.*, 2017a), the two cellulosome types differ in the diversity of glycoside hydrolase (GH) families present as catalytic components (Artzi, Bayer and Moraïs, 2017; Haitjema *et al.*, 2017a). In contrast to the bacterial cellulosomes, the fungal scaffoldin system is broadly conserved across the anaerobic fungal clade, allowing

interspecies cross-linking binding activity. The most profound biochemical difference between the bacterial and fungal cellulosomes is the end products produced during the hydrolysis of crystalline cellulose. Degradation of cellulose by Clostridial cellulosomes results in the production of cellobiose, which is taken up by the cell, hydrolyzed to glucose and metabolized (Schwarz, 2001). However, fungal cellulosomes produce glucose during cellulose degradation (Gilmore, Henske and O'Malley, 2015). Taken together, these differences may make fungal cellulosomes a superior candidate for metabolic engineering.

Recently, a comprehensive set of proteins that is critical to the assembly of fungal cellulosomes was described for the first time (Haitjema *et al.*, 2017a). This advancement is bound to lead to improvements in the "designer cellulosome" field (Gilmore, Henske and O'Malley, 2015), especially since the "parts" of the fungal cellulosome (i.e. cohesin, dockerin, scaffoldin) are markedly different from their bacterial counterparts. For example, fungal cohesion/dockerin units, which tend to be promiscuous by nature, could be used to engineer an extracellular production platform. By making use of a consortium of engineered organisms, each producing different enzymes required for a specific function, it is possible to assemble a complex *bona fide* cell factory line for almost limitless applications. This is important for developing modular non-interacting metabolon systems within the same organism that could be directly exploited in metabolic engineering applications.

1.8 Anaerobic gut fungi are challenging to work with, hampering the rate at which we can exploit them for biotechnology

Two of the greatest challenges facing the further development of anaerobic, and other non-conventional, fungi for metabolic engineering applications are their unique growth

requirements and their genetic intractability. Furthermore, owing to the difficulty in maintaining cultures of isolates (Haitjema *et al.*, 2014) it is not surprising that so few basal clades (like the anaerobic fungi) have been characterized. Similar to "unculturable" prokaryotes (Cowan *et al.*, 2005), the construction of viable culture collections is limited by the lack of information on the nutritional requirements of non-conventional fungi, *e.g.* dependence on another species to provide a key resource. Furthermore, the slow pace of discovery and biotech translation from these fungal clades is likely the result of the relatively small number of researchers dedicated to studying non-model fungi.

Apart from difficulties in isolating and maintaining cultures of slow-growing non-conventional fungi in a laboratory setting, the fungal isolates may be remarkably resistant to further genetic manipulations. Currently, only one report describes the transformation of anaerobic fungi with a plasmid that encodes a heterologous gene (Durand *et al.*, 1997). In the 1997 study, *N. frontalis* was bioholistically transformed with a plasmid containing the bacterial β-glucuronidase gene, downstream of a promoter sequence that was previously derived from the fungus (Fischer, Durand and Fèvre, 1995). While activity of the reporter protein was detected, gene expression was transient and lost after only a few generations, possibly due to the lack of a selection marker on the plasmid. The authors suggest that for improved stability, the reporter gene should be coupled to a selection marker such as a gene encoding resistance to an antibiotic, which would minimize the risk of plasmid loss over time (Durand *et al.*, 1997).

However, there are additional considerations, such as the mechanisms by which a plasmid is maintained in the host organism. A prerequisite for stability of the transformed strain is that the host is able to replicate and distribute the plasmid to progeny, and currently the gut fungal

mechanisms for plasmid replication and maintenance remain unknown. Further, given the complicated life cycle of gut fungi that involves both a single-celled motile zoospore-stage and a stage where a maturing sessile sporangium encapsulates hundreds of novel zoospores (Orpin, 1975), it is difficult to predict when, and how, the fungi should be transformed for optimal stability. These challenges may be overcome by more permanent genome engineering, which necessitates high-resolution genomic information and suggested regions of integration. However, the lagging development of a well-defined genetic toolkit for non-conventional fungi, such as promoters, transcriptional terminators and ribosomal binding sites, limits direct manipulation efforts.

In order to edit the gut fungal genome, it is necessary to get DNA across the gut fungal cell wall without compromising viability, and again the intricate and largely anaerobic life cycle of gut fungi serve as an example of the complications that should be taken into consideration. Moreover, compared to most model organisms, the gut fungal genomes are remarkably AT-rich (Nicholson, Theodorou and Brookman, 2005; Youssef *et al.*, 2013; Haitjema *et al.*, 2017a), and that may affect how the organism handles DNA that does not adhere to the apparent codon preference (Komar, 2016). Last, virtually nothing is known about the post-translational modifications or folding behavior of gut fungal proteins, and that should be taken into account when one wants to use a heterologous selection/reporter protein.

1.9 Heterologous production of gut fungal genes in model systems can be used to exploit their lignocellulolytic enzymes

While large-scale cultivation of anaerobic organisms remains challenging, and while most of them are still genetically intractable, it has been shown that gut fungal enzymes can be

successfully produced in engineered microbes. Early studies on the enzymatic activities of gut fungal culture fractions revealed that these organisms greatly contribute to biomass degradation in the animal host, and possess powerful enzymes that efficiently degrade α- and β-glucans, β-galatans, galactomannans, and arabinoxylans (Mountfort and Asher, 1985; Pearce and Bauchop, 1985; Wood *et al.*, 1986; Lowe, Theodorou and Trinci, 1987a; Williams and Orpin, 1987). It was soon discovered that the majority of the gut fungal enzymatic activities are extracellular or confined to the membrane fraction (Pearce and Bauchop, 1985), and a number of enzymes have been purified and characterized from gut fungal cultures since (see *e.g.* (Hebraud and Fevre, 1990a, 1990b; Borneman *et al.*, 1991; Li and Calza, 1991; Teunissen *et al.*, 1992; Garcia-Campayo and Wood, 1993; Vardakou *et al.*, 2008; Wang, Chen and Hseu, 2014)).

The advent of molecular tools opened up new possibilities for enzyme discovery and production, and in the early 1990s, gut fungal cDNA libraries expressed in *Escherichia coli* were successfully used to screen for and identify novel cellulolytic enzymes (Reymond *et al.*, 1991, 1992; Xue *et al.*, 1992). By the time of this work, around 100 gut fungal proteins have been successfully produced in a number of heterologous hosts, firmly establishing that these enzymes can be transferred to a wide variety of biotechnologically interesting organisms for downstream applications, including bacteria (*E. coli, Butyrivibrio fibrisolvens, Bacillus subtilis, Lactobacillus reuteri, Clostridium beijerinckii*) (Gilbert *et al.*, 1992; Xue, Gobius and Orpin, 1992; Lee *et al.*, 1993; Xue *et al.*, 1997; Elliott *et al.*, 1999; Smidt *et al.*, 2001; Liu *et al.*, 2005); yeasts (*Saccharomyces cerevisiae, Pichia pastoris, Kluveromyces lactis, Hansenula polymorpha*) (van der Giezen *et al.*, 1998; Durand, Rascle and Fèvre, 1999; Harhangi *et al.*, 2002; Li *et al.*, 2004; O'Malley, Theodorou and Kaiser, 2012a); filamentous

fungi (*Penicillium roqueforti, Trichoderma reesei, Hypocrea japonica*) (Durand, Rascle and Fèvre, 1999; Li *et al.*, 2007; Poidevin *et al.*, 2009); and plants (*Brassica napus, Nicotiana tabacum*) (Liu *et al.*, 1997; Obembe *et al.*, 2007).

1.10 Identification and characterization of fungal genes and pathways is critical to unleashing their potential

Fungi have played an important role in human activities for millennia, and continue to lend themselves well to diverse applications. Great effort has been made to identify and characterize novel fungal species and enzymatic activities. However, although ~1,200 new fungal species are described each year, it has been estimated that the ~100,000 described species represent a mere 2-10% of the total number (Blackwell, 2011; Tedersoo *et al.*, 2014), and so it seems that we have barely begun to unlock the hidden biotechnological potential of this extensive kingdom.

Among the early discoveries were pectinases from saprophytic fungi such as *Aspergillus sp.*, *Penicillium sp.*, and *Rhizopus sp.* that clarify fruit juices (Zoltan and John, 1933; Mantovani, Geimba and Brandelli, 2005), rennet from *Aspergillus sp.* and *Mucor sp.* that clot milk for cheese production (Arima, Iwasaki and Tamura, 1967), manganese peroxidases from *Phanerochaete sp.* that decolor paper pulp effluents and degrade xenobiotics (Glenn and Gold, 1983; Bajpai, Mehna and Bajpai, 1993; Field *et al.*, 1993), and cellulases from *Trichoderma sp.* (Allen *et al.*, 2009) and anaerobic fungi (Orpin, 1975; Youssef *et al.*, 2013; Solomon *et al.*, 2016) that degrade plant biomass. While screening, biochemical purification, and activity evaluation are central components of fungal enzyme discovery, they can be slow to identify specific fungal enzymes responsible for a chemical transformation and restrict discovery to

easily assayed enzymatic activities. On the other hand, molecular biology methods coupled with modern sequencing approaches have greatly expanded fungal enzyme discovery through a number of approaches.

The plummeting cost of DNA-sequencing and -synthesis, and the development of sophisticated molecular tools have greatly facilitated the discovery of enzymes and the transfer of these enzymes into amenable microorganisms such as bacteria and yeasts. In a way, this abolishes the need to cultivate the original host organism at all: as long as we have the sequence, from *e.g.* a metagenomic library, we can transfer the gene to a workhorse microbe to detect expression and activity. In other words, researchers can select enzymes and biosynthetic pathways from various sources and put them together in a cell factory host of choice, as shown in Figure 1.4.

1.11 Modern omics-based approaches speed up the rate of gene and pathway discovery

The past decade has seen extraordinary advances in sequencing technology that have rapidly sequenced the genomes of many previously uncharacterized fungal systems, including elusive anaerobic fungi (Dean *et al.*, 2005; Nierman *et al.*, 2005; Kämper *et al.*, 2006; Martinez *et al.*, 2008; Youssef *et al.*, 2013; Solomon *et al.*, 2016; Haitjema *et al.*, 2017a). These sequences are annotated and curated in fungal databases such as the Saccharomyces Genome Database (Cherry *et al.*, 2012), the Aspergillus Genome Database (Arnaud *et al.*, 2010), and the JGI Mycocosm, which is a repository for the 1000 fungal genome project covering all branches of fungi (Grigoriev *et al.*, 2014). The pace of sequencing has greatly overtaken discovery efforts, resulting in a growing catalog of unexplored fungal enzymes.

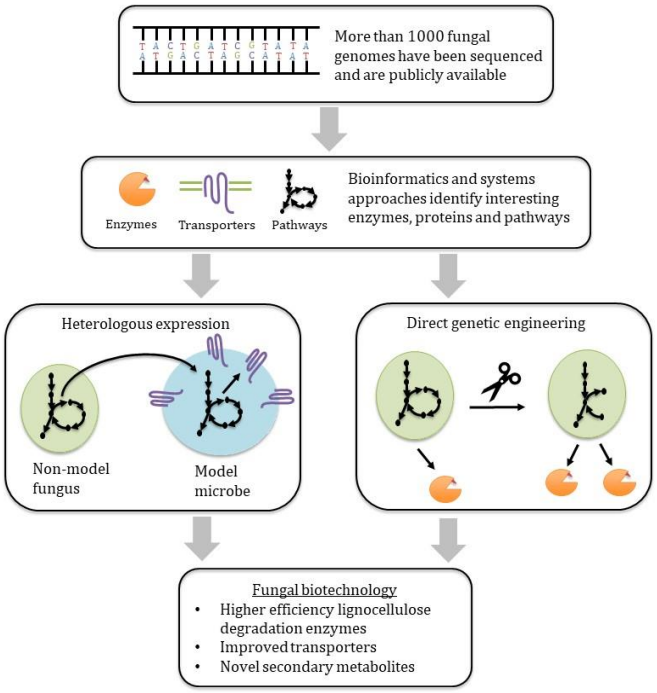

Figure 1.4: Vast amounts of sequencing data have been collected, predominately from non-model fungi. Systems and bioinformatic tools are used to identify enzymes and pathways of biotechnological relevance. *In silico* analyses and predictions must be coupled with experimental techniques to validate putative gene annotation. Figure reproduced from (Wilken *et al.*, 2019).

However, bioinformatic tools and approaches readily analyze these growing datasets to provide new insight. Classical bioinformatics approaches include homology and Hidden Markov Model (HMM) searches that analyze new sequences for conserved signatures found in well-documented proteins (Karplus, Barrett and Hughey, 1998; McGinnis and Madden, 2004). These approaches were recently applied to both genomic and transcriptomic sequences to discover hundreds of new cell wall degrading enzymes in gut fungi *P. finnis* and *Orpinomyces sp. C1A* (Youssef et al., 2013; Solomon et al., 2016). These genomics approaches are increasingly complemented with proteomics methods that map individual

proteins back to the genes that produce them to identify new enzymes from fungal secretions and lysates (Doyle, 2011; Solomon et al., 2016).

A particularly interesting recent study highlights how 'omics' based approaches are a powerful tool in enzyme discovery is the transcriptomics-guided construction of an enzyme cocktail composed of various biomass-degrading enzymes from *Orpinomyces sp. C1A* (Morrison, Elshahed and Youssef, 2016). The enzymes were produced and purified from *E. coli*, mixed in different ratios, and the activity of the cocktail on various substrates was assayed, resulting in the identification of several promising enzyme candidates for lignocellulosic bioprocessing.

Transcriptomic analysis has also been used to identify several novel lignocellulose active enzymes in anaerobic gut fungal transcriptomes (Henske, Gilmore, *et al.*, 2018). Some of these transcripts were found to co-regulate with transcripts encoding for previously known carbohydrate active enzymes (CAZymes), suggesting that they may code for currently unidentified lignocellulose modifying proteins. Furthermore, the recent discovery of anaerobic microorganisms that consume lignin as their sole carbon source suggest that these transcripts may actually code for anaerobic lignin-active enzymes from early branching fungi (Woo *et al.*, 2014; Billings *et al.*, 2015).

Apart from biomass degrading enzymes, fungi are likely to possess a wide repertoire of membrane proteins with importance to biotechnology, such as receptors and transporters for biohydrolysates (reviewed in (Boyarskiy and Tullman-Ercek, 2015; Jones, Hernández Lozada and Pfleger, 2015; Kell *et al.*, 2015)). Biohydrolysate transporters may allow simultaneous saccharification and fermentation of biomass, alleviate the need for complete hydrolysis, and remedy inhibitory effects of glucose on fungal cellulases (Kim et al., 2014). Owing to their

hydrophobic amino acid compositions, it is relatively straightforward to identify proteins with secretion signal peptides and transmembrane segments from primary sequence data (Krogh et al., 2001; Petersen et al., 2011; Tsirigos et al., 2015). Recently, a transcriptomic analysis of anaerobic gut fungi revealed a large number of transporters that have great biotechnological potential, including receptors and transporters that had previously not been found in the fungal kingdom (Seppälä et al., 2016).

Given the breadth and depth of high-quality omics data available to study the gut fungi, it is becoming increasingly important to collate this information in a systematic way to guide future studies. Systems biology, and specifically genome-scale models, is just the tool for this purpose (Mih and Palsson, 2019).

1.12 Genome-scale models organize omics data to facilitate the interrogation and utilization of omics datasets

A genome-scale model (GSM) is a detailed mathematical model of cellular metabolism and physiology. This representation of the metabolism of a microorganism has found widespread use in the metabolic engineering community because it makes it possible to accurately simulate the metabolic fluxes of a microorganism *in silico* (Orth, Thiele and Palsson, 2010). This enables systematically directed modifications of a strain for improved performance (King *et al.*, 2015). Briefly, a GSM is constructed by creating a catalog of all the reactions a cell is capable of, based on its annotated genome. In this way, each metabolic reaction may be associated with an enzyme, which is linked to a specific gene.

Consider Figure 1.5 as an example of a minimal cell importing metabolite A and exporting metabolite D. The cell's metabolism is shown in Figure 1.5.A, here each gene, G_i (for i =

1,…,6), is associated with a corresponding metabolic reaction, r_i. The full kinetic model of this cell is shown in Figure 1.5.B. By associating each reaction with a metabolite, it is possible to simulate the effect of genetic knockouts on metabolism. However, parameterizing even this simple first order kinetic model is technically challenging for all but reduced models of very well studied organisms, like *E. coli* (Khodayari *et al.*, 2014).

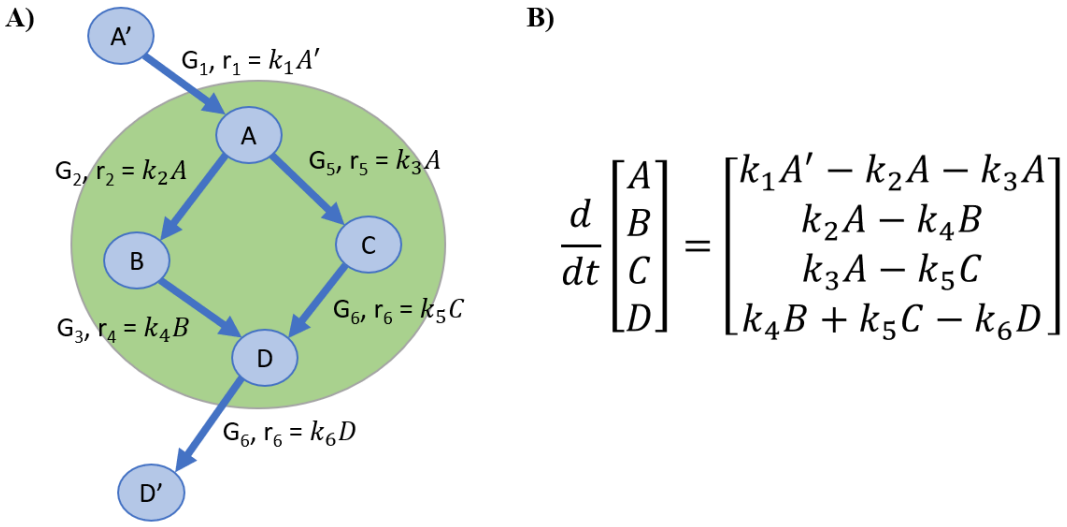

Figure 1.5: An example of a minimal cell with simplified metabolism. A) The metabolism of the cell. Metabolites are denoted by nodes and reactions (enzymes) are denoted by edges connecting metabolites. B) The associated full kinetic model of the cell assuming simple first order kinetics.

To perform simulations, it is necessary to simplify the model. First, steady state is assumed which reduces the set of differential equations into a set of algebraic equations, as shown in Equation (1.1).

$$\frac{d}{dt}\begin{bmatrix}A\\B\\C\\D\end{bmatrix} = \begin{bmatrix}k_1A' - k_2A - k_3A\\k_2A - k_4B\\k_3A - k_5C\\k_4B + k_5C - k_6D\end{bmatrix} = 0 \qquad (1.1)$$

Next, the model is rewritten to separate the reaction fluxes (**v**) from their stoichiometry, as shown in Equation (1.2).

$$\begin{bmatrix} k_1A' - k_2A - k_3A \\ k_2A - k_4B \\ k_3A - k_5C \\ k_4B + k_5C - k_6D \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{bmatrix} = \boldsymbol{Sv} = \mathbf{0} \tag{1.2}$$

Additionally, flux constraints (denoted, $\mathbf{v}_{min}$ and $\mathbf{v}_{max}$) are imposed on the model to ensure that the solved for fluxes are biologically reasonable. While the model could theoretically be solved now, there would be significant degeneracy in the flux solutions, **v**, due to underdetermined nature of Equation (1.2). An additional assumption is required to further reduce the solution space of **v** to better align with the physiology of the cell. Typically, this is achieved by tying the cell growth rate to the flux through the metabolic model by an empirical function called the biomass objective function, often denoted $\mu(\mathbf{v})$ (Feist and Palsson, 2010).

The biomass objective function is typically constructed by attempting to account for each biological building block a cell requires for growth. For a simple model this would include all the amino acids, nucleotides, lipids and carbohydrates that a cell synthesizes during growth. An added benefit of using the biomass objective function is that it allows the model to predict the growth rate of the organism under different conditions (nutrient, genetic knockouts etc.), which can be valuable for strain engineering purposes.

Finally, by assuming that the cell wishes to maximize its growth rate we can cast the aforementioned assumptions into a linear program, as shown in Equation (1.3). This

formulation is called flux balance analysis and forms the cornerstone of computational techniques that are used to interrogate genome-scale models.

$$\max_{\boldsymbol{v}} \mu(\boldsymbol{v})$$
$$\text{s.t. } \boldsymbol{Sv} = \boldsymbol{b} \tag{1.3}$$
$$\boldsymbol{v_{min}} \leq \boldsymbol{v} \leq \boldsymbol{v_{max}}$$

By solving Equation (1.3) a unique maximal growth rate is attained, as well as intracellular metabolic fluxes, which are typically found to match experimental data well (Schuetz, Kuepfer and Sauer, 2007). Furthermore, the mechanistic foundation of this approach makes analyzing the metabolic consequences of genetic modifications of a strain straightforward. This obviates the need to perform time consuming experiments when trying to engineering the metabolism of a cell to improve performance (Simeonidis and Price, 2015). Moreover, GSMs can also be used as a succinct way to collate various omics and experimental data to guide future characterization efforts (King *et al.*, 2015).

In addition to guiding metabolic engineering strategies, GSMs may also be used to improve gene annotation databases. For example, the model-enabled gene search (MEGS) algorithm uses discrepancies between model predictions and experimental culture data to identify missing or mis-annotated genes (Pan *et al.*, 2017). While this technique has only been applied successfully to bacteria, it is a promising approach that can be used to identify genes in other microbes with missing annotations, like non-model fungi.

1.13 High quality genome-scale models of fungi are hard to construct, and don't exist for Neocallimastigomycota fungi yet

Rapid improvements in sequencing technology have greatly increased the number of fully sequenced fungal genomes available for study and further exploitation. Figure 1.6 shows the number of sequenced genomes in each clade of the fungal tree of life, taken from the Joint Genome Institute (JGI) Mycocosm database (Grigoriev *et al.*, 2014).

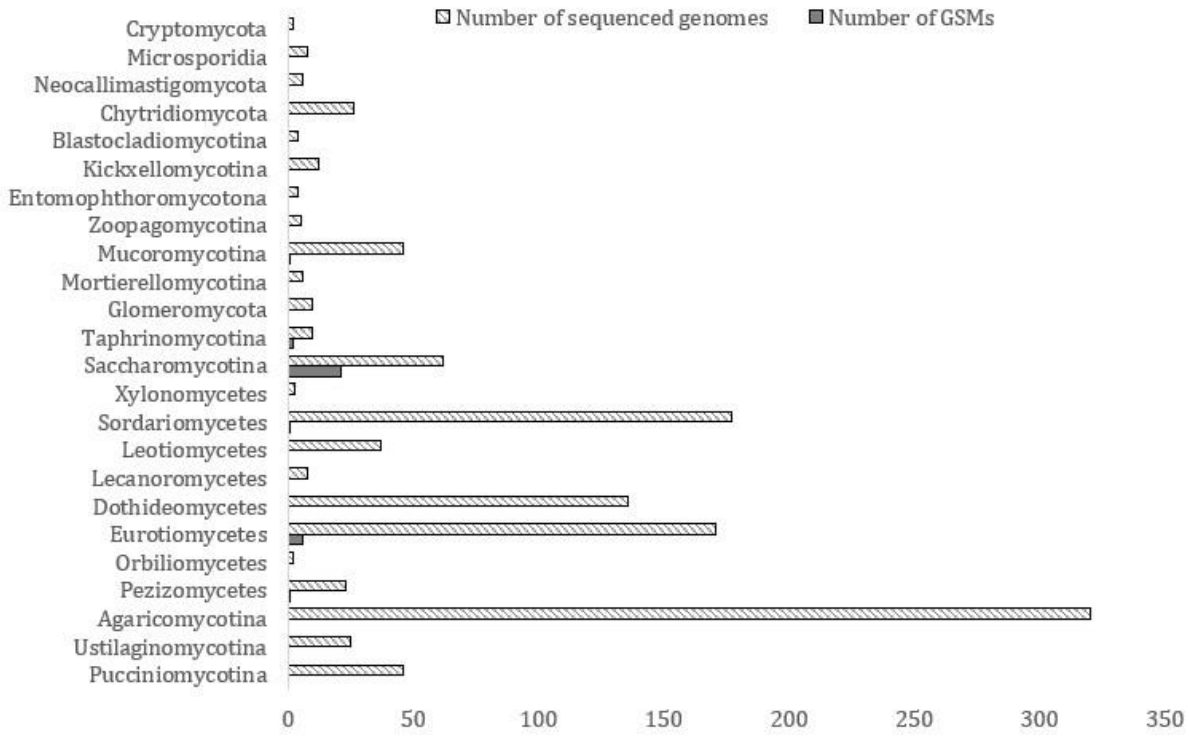

Figure 1.6: The number of sequenced genomes and genome-scale models (GSMs) available for each clade in the fungal tree of life. The generation of sequenced genomes (currently exceeding 1000), far out paces the generation of GSMs. Notably, most reconstructions have been done for biotechnologically important fungi, such as *Aspergillus niger* (Eurotiomycetes), and the yeasts *Saccharomyces cerevisiae* and *Pichia pastoris* (Saccharomycotina). Figure reproduced from (Wilken *et al.*, 2019).

The increasing availability of fungal genomes, coupled with complementary "omics" datasets (transcriptomics, proteomics and metabolomics) holds great promise for future discoveries. However, very few genome-scale models have been reconstructed from fungal omics data per clade. The significant difference between the number of sequenced genomes and available GSMs highlights the long-standing knowledge gap in our systems level understanding of fungi. To effectively utilize the enzymes and pathways of the fungal world, a multifaceted approach that combines sequence informatics, systems modeling, and traditional metabolic engineering approaches is required.

Gene annotation databases, e.g. KEGG, Pfam, etc., are important references to understand cellular metabolism and physiology (Akiva *et al.*, 2014), and are used to assign function to genes, which is a prerequisite for cellular modeling. The quality of the metabolic reconstructions greatly influences the accuracy of the predictions, and as such it is imperative to ensure that the models are of high quality. Unfortunately, this task is invariably time-intensive, as it requires significant manual oversight (Thiele and Palsson, 2010). Various automatic reconstruction tools have been developed, but historically they have focused on bacterial reconstructions (Henry *et al.*, 2010; Machado *et al.*, 2018). Due to their increased genome size and compartmentalization, eukaryotic models are more challenging to develop than prokaryotic models. The prevalence of mis- or un-annotated genes in non-model fungi exacerbates these difficulties. Recently, a few tools were developed that specifically aid in fungal reconstructions. The RAVEN toolbox was used to reconstruct a GSM of *Penicillium chrysogenum* (Wang *et al.*, 2018) and KBase recently released a tool that focuses specifically on fungal reconstructions (Arkin *et al.*, 2018). However, manually curation is an unavoidable element of GSM construction and is the primary reason for the slow rate of model creation

(Thiele and Palsson, 2010). Despite this obstacle, Table 1.2 shows a list of some recently completed fungal GSMs.

Table 1.2: Recent GSM reconstructions of fungi span only three clades in the fungal tree of life, namely Eurotiomycetes, Saccharomycotina, Mucoromycotina.

Modeled fungus	Main results
Aspergillus niger (Lu *et al.*, 2017)	Updated GSM includes 1764 reactions and 1210 open reading frames. Flux balance analysis (FBA) predictions were validated using ^{13}C labeling experiments. Investigated the effect of cofactor utilization on glucoamylase production.
S. cerevisiae (Aung, Henry and Walker, 2013)	The latest yeast consensus model focused on triglyceride production, but it is being continuously updated, see https://github.com/SysBioChalmers/yeast-GEM for the current state of the model.
Pichia pastoris (Tomàs-Gamisans, Ferrer and Albiol, 2016)	A consensus GSM from three other models updated the central metabolic pathways and accounts for 1026 genes, 1689 metabolites and 2035 reactions.
Mucor circinelloides (Vongsangnak *et al.*, 2016)	The reconstruction accounts for 1213 genes, 1413 metabolites and 1326 reactions. FBA accurately predicts nutrient usage and requirements. Comparative analysis to other oleaginous fungi revealed expanded central metabolic genes could explain nutrient utilization differences.

1.14 Coupling experimental validation to omics data and modeling is necessary to unlock the biotechnological promise of anaerobic gut fungi

The integration of computational techniques and experimental validation will pave the way for realizing the biotechnological potential of non-model fungi. Currently, direct experimental validation and exploitation lags behind the rate at which novel enzymes and biosynthetic compounds are identified *in silico* (Stewart *et al.*, 2018). Synthesizing this knowledge in a framework that lends itself to systemization, like GSMs, could help direct focus to areas of non-model biotechnology that need the most attention. The wealth of bioinformatically identified enzymes, pathways, and proteins, that are being coupled with nascent systems biology approaches designed to mine these datasets, suggests that rapid advances in fungal-based bioprocesses are imminent (Wilkinson *et al.*, 2018).

II. Sculpting gut microbial communities alters fermentation products and methane release

This chapter is based upon work that is under revision for publication in Nature Microbiology by Xuefeng Peng, St. Elmo Wilken, Thomas S. Lankiewicz, Sean P. Gilmore, Jennifer L. Brown, John K. Henske, Candice L. Swift, Asaf Salamov, Kerrie Barry, Igor V. Grigoriev, Michael K. Theodorou, David L. Valentine, and Michelle A. O'Malley, entitled "*Sculpting gut microbial communities alters fermentation products and methane release*". See the upcoming publication for more detailed information regarding the results and methods.

2.1 Introduction

Anaerobic consortia native to digestive tracts of herbivores have co-evolved with their hosts for millions of years to utilize lignocellulosic hydrolysates (Groussin *et al.*, 2017). Gut microbes span all three domains of life (archaea, bacteria, fungi, protozoa), and the plethora of biomass-degrading genes and genomes discovered from gut microbiomes indicate that the gut microbiome should be an ideal source for down-selecting lignocellulolytic consortia (Seshadri *et al.*, 2018).

While earlier studies focused on the chemical performance of anaerobic consortia (Adney *et al.*, 1991), advances in sequencing technology over the past decade have spurred a wave of studies investigating the community composition of anaerobic consortia (Stewart *et al.*, 2018). These recent sequencing efforts have revealed microbial members common to successful anaerobic consortia, but they are limited by the lack of mechanistic insight into the functional diversity displayed by anaerobic consortia. Moreover, anaerobic consortia studied before were almost exclusively prokaryotic. Cross-domain partnerships in anaerobic consortia has not

received as much attention. Interestingly, anaerobic gut fungi, which are known to possess a wide range of biomass-degrading enzymes (Solomon et al. 2016), are central to the lignocellulolytic ability of herbivorous animals (Gruninger *et al.*, 2014). This inspired us to investigate down-selected anaerobic consortia featuring anaerobic gut fungi as the primary lignocellulose degrader, using both broad marker gene survey and in-depth metagenomic analysis.

To elucidate these relationships, we performed several parallel enrichment experiments to enrich biomass-degrading consortia from goat feces and identify microbes that drive the activity and stability of these cultures. Fecal samples were enriched subject to different media conditions, as shown in Figure 2.1. Ten billion metagenomic reads spread across 396 enrichment cultures tracked biological diversity as the cultures converged to a minimal set of microorganisms that were stable after more than ten culture generations. Over 1.5 Tbp (10^{12} base pairs) of metagenome sequencing enabled the recovery of 719 high-quality metagenome-assembled genomes unique at the species level, 96% of which were previously uncultured novel microbes within the herbivore digestive tract. Nearly 165,257 carbohydrate active enzymes (CAZyme) domains were identified from the fecal samples alone, constituting over 10% of the known CAZymes in existence. Surprisingly, consortia dominated by anaerobic fungi generated more than twice the amount of methane (CH_4) compared to prokaryotic consortia, suggesting that fungi play a key role in CH_4 release in ruminant herbivores.

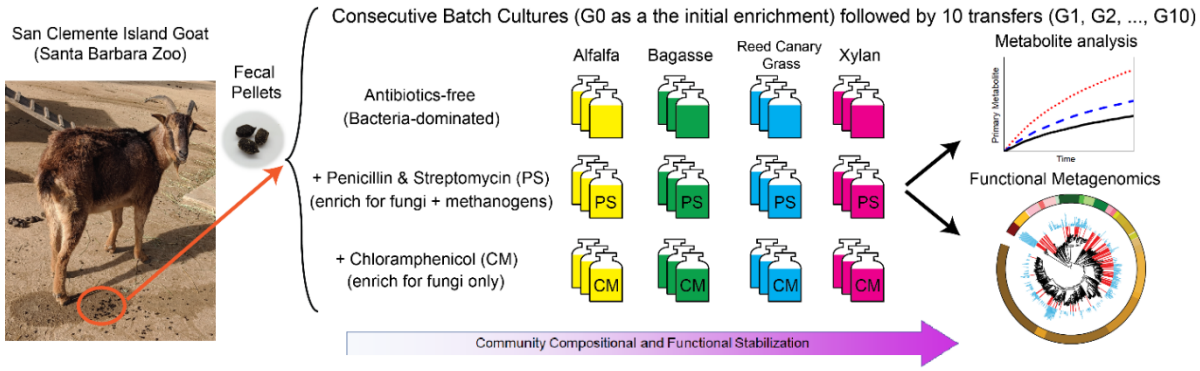

Figure 2.1: Overview of the parallel enrichment strategy to select and analyze cross-domain anaerobic lignocellulolytic consortia. Freshly produced fecal pellets from a San Clemente Island Goat served as the source microbiome for 396 parallel microbial enrichment experiments. Enrichment cultures were initiated by challenging the fecal consortia with four types of substrates and two types of antibiotics to bias survival of different microbial communities, with triplicate cultures for each condition. Penicillin and streptomycin (PS) were used to inhibit bacterial growth. Chloramphenicol (CM) was used to inhibit both bacterial and archaeal growth. Membership within the parallel enrichments were tracked via metabarcoding and whole metagenome assemblies (for G0, G5, and G10), and metabolomic analyses of headspace and liquid cultures were monitored at each generation.

The most active microbial consortia were comprised of cross-domain partnerships between anaerobic fungi from the genus Neocallimastix, methanogenic archaea from the genus Methanobrevibacter, and bacteria from the phylum *Firmicutes*, which produced high yields of CH_4, and are capable of cryopreservation and revival. Metabolic product profile comparison of consortia treated with and without antibiotics revealed that the level of CH_4 production was the highest when fermentation products are restricted to H_2, formate, and acetate (as in the consortia dominated by anaerobic fungi), as opposed to being diverted to the production of butyrate and propionate (as in the antibiotics-free consortia). Overall, this analysis points to natural compartmentalization between anaerobes as a means to degrade

crude biomass, which can be exploited to harness nature's microbes for industrial bioprocessing.

2.2 Results and discussion

2.2.1 Metagenomic reconstruction of the goat fecal microbiome reveals novel cultured taxa

Over 1.5 Tbp (10^{12} base pairs) of metagenome sequencing enabled the recovery of 2452 high-quality prokaryotic metagenome-assembled genomes (MAGs) from goat feces, all of which are >80% complete with <10% contamination evaluated by CheckM (Parks *et al.*, 2015). Of these, 719 are unique at the species level based on the recently proposed criteria of species definitions (Varghese *et al.*, 2015; Olm *et al.*, 2017), as shown in Figure 2.2. This collection of MAGs is among the largest and the highest quality to date (Campanaro *et al.*, 2016; Güllert *et al.*, 2016; Vanwonterghem *et al.*, 2016; Mukherjee *et al.*, 2017; Svartström *et al.*, 2017; Gharechahi and Salekdeh, 2018; Seshadri *et al.*, 2018; Solden *et al.*, 2018; Stewart *et al.*, 2018, 2019) for anaerobic microbiomes (with 91.8% mean completeness and 1.4% mean contamination). A comparative analysis was performed based on the genome-wide average nucleotide identity of open reading frames to quantify the increase in phylogenetic diversity contributed by MAGs assembled in this study. Compared to 9089 genomes from three of the largest rumen collections (Seshadri *et al.*, 2018; Stewart *et al.*, 2018, 2019), the Genomic Encyclopedia of Bacteria and Archaea (GEBA) collection (Mukherjee *et al.*, 2017), a recent human gut bacteria collection (Zou *et al.*, 2019), and 221 additional reference genomes from NCBI RefSeq (O'Leary *et al.*, 2016), 686 of the 719 MAGs (95%) in this dataset were novel at the species level. The MAG collection contributed by this study

underscores the vast untapped metabolic potential in herbivores that perform foregut and hindgut fermentations.

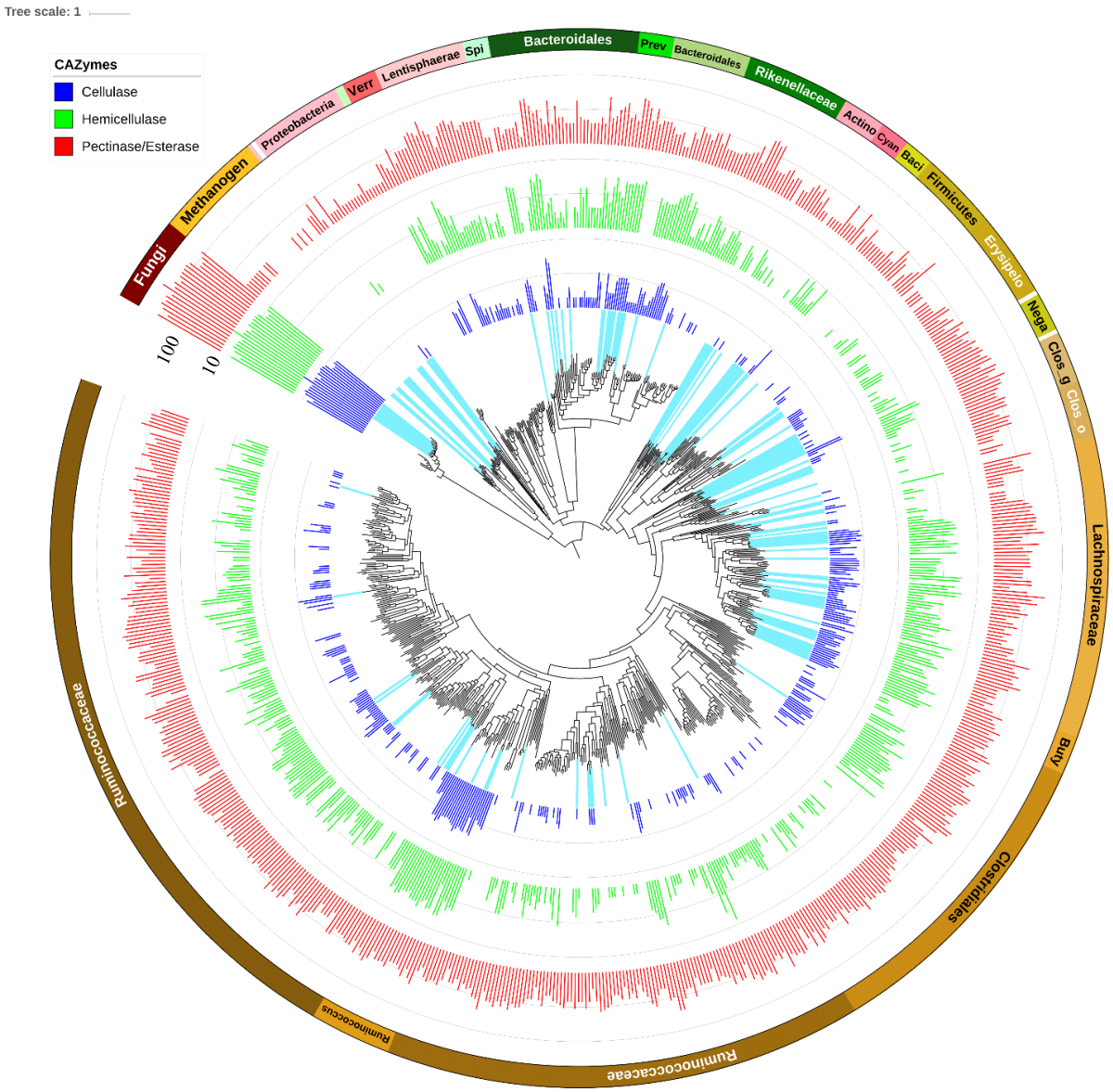

Figure 2.2. Microbial tree of prokaryotic (719 MAGs) and eukaryotic (20 eukMAGs) life reconstructed from goat fecal metagenomes with taxonomic annotations. The phylogeny was constructed from 400 broadly conserved proteins using PhyloPhlAn (Segata *et al.*, 2013). Prokaryotic MAGs are divided into 51 groups (colored in the outer ring) by phylogeny determined from a tree including 3237 reference genomes. Colored bars plotted on a logarithmic scale represent the number of carbohydrate-

active enzymes (CAZymes including cellulase, hemicellulase, and pectinase/esterase) found in each MAG/eukMAG. The cyan bars radiating from the tip of tree leaves indicate MAGs that were present in at least one of the enrichment cultures at the end of the experiment (G10). Cyan: Cyanobacteria; Actino: Actinobacteria; Verr: Verrucomicrobia; Spi: Spirochaetes; Prev: Prevotella; Baci: Bacilli; Erysipelo: Erysipelotrichaceae; Nega: Negativicutes; Clos_g: Clostridium; Clos_o: Clostridiales; Buty: Butyrivibrio.

While a number of recent studies have used metagenomics to assess the metabolic potential and interactions within the herbivore rumen (Hess *et al.*, 2011; Solden *et al.*, 2018; Stewart *et al.*, 2018), less attention has been paid to the hindgut of herbivores where microbes, many of which originate and were active in the rumen, encounter recalcitrant plant material that is not completely processed in the foregut. Additionally, gut microbes cultured directly from the rumen (via fistulated animals) have proven extremely difficult to stabilize in culture (Seshadri *et al.*, 2018), possibly due to strict nutritional or mechanical requirements that are difficult to mimic outside of the rumen. Therefore, it was hypothesized that gut microbial enrichment cultures from feces would be more robust and resilient than from the rumen, because the part of the fecal microbiome capable of surviving on recalcitrant lignocellulosic residues in the hindgut undergo a broad range of biological, chemical, and physical conditions. Three quarters (531) of the assembled MAGs were *Firmicutes* and more than half of them belong to the family *Ruminococcaceae*, most of which were novel and not enriched in culture except for the genus *Ruminococcus*, as shown in Figure 2.2. The second most abundant phylum (12%) among the MAGs was *Bacteroidetes* (85), of which the most abundant group belongs to the family *Rikenellaceae*. Twenty-five archaeal MAGs were also recovered, of which *Methanobrevibacter* was the most abundant. Additionally, three *Methanosphaera*

stadtmanae MAGs and seven *Thermoplasmata* MAGs with the potential to generate CH_4 using methanol and acetate were recovered. The rest of the prokaryotic MAGs (14%) were from the phyla *Proteobacteria* (23), *Lentisphaerae* (21), *Actinobacteria* (10), *Verrumicrobia* (8), *Cyanobacteria* (7), *Spirochaetes* (6), *Planctomycetes* (2), and *Elusimicrobia* (1).

Twenty MAGs larger than 40 Mbp in size were reconstructed from the enrichment consortia, which were termed "eukMAGs" because over 80% of the genes in the eukMAGs were classified as "Eukaryota" by BLAST+ (Camacho *et al.*, 2009). It is particularly challenging to recover eukaryotic MAGs, especially fungal MAGs, as their genomes are >10 Mbp and they are often characterized by long and frequent repeat regions of low GC content. All eukMAGs were classified to the fungal subphylum *Neocallimastigomycota*, commonly known as the anaerobic gut fungi. The eukMAGs in the enrichment consortia belong to the genus *Neocallimastix* and are closely related to the strain *Neocallimastix californiae*, which was previously isolated from the feces of a goat (Solomon *et al.*, 2016; Haitjema *et al.*, 2017b). Benchmarking the BUSCO-estimated completeness of our eukMAGs to the draft genome of *Neocallimastix californiae* showed that they are 81.1% complete on average; the most complete eukMAG was estimated to be 96.9% complete. To date, no previous MAG datasets have identified anaerobic fungi, as they are typically low-abundance members of the digestive tract microbiome (Stewart *et al.*, 2019), yet they have been recently found to contain a wealth of biomass-degrading enzymes (Solomon *et al.*, 2016) and multi-enzyme cellulosomes (Haitjema *et al.*, 2017b). This collection of prokaryotic MAGs and eukMAGs from the goat fecal microbiome serves as a rich resource for metagenomic studies of gut microbiomes, covering microbial taxa some of which are not included in published collections from the rumen microbiome (Seshadri *et al.*, 2018; Stewart *et al.*, 2018).

2.2.2 Gut microbes are rich in biomass-degrading enzymes

The annotated prokaryotic MAGs and eukMAGs were analyzed for their content and diversity of carbohydrate-active enzymes (CAZymes) that fall into the functional categories of cellulase, hemicellulase, and pectinase/esterase according to the CAZy database (Lombard *et al.*, 2014). It is important to note that glycoside hydrolase (GH) families 5, 8, 44, and 51 are versatile in function and can hydrolyze both cellulose and hemicellulose depending on the specific subfamily (Lombard *et al.*, 2014). Anaerobic fungi from the genus *Neocallimastix* represented by eukMAGs contained up to several hundred of each type of CAZymes per genome (see Figure 2.2) and were only enriched in antibiotics-treated (penicillin & streptomycin, "PS", or chloramphenicol, "CM") cultured consortia. Among prokaryotic MAGs, the taxa containing the largest number of CAZymes included the anaerobic bacteria *Ruminococcus*, *Butyrivibrio*, *Prevotella*, *Paenibacillaceae*, *Bacteroidaceae*, and *Ruminococcaceae* (all typically found in the rumen). These taxa were enriched in antibiotics-free consortia grown on lignocellulose, and generally include more than 5 cellulases and more than 10 hemicellulases and pectinases/esterases per strain.

In the 719 high-quality prokaryotic MAGs, there were predicted 1365 cellulases, 3686 hemicellulases, and 3433 pectinases/esterases, and in the 20 eukMAGs there were comparable number of CAZymes (1887 cellulases, 2597 hemicellulases, and 2662 pectinesteraes), indicating the vast hydrolytic potential of *Neocallimastigomycota*. Major bacterial cellulases include GH5 and GH9; one GH44 was unique to bacterial MAGs. Major cellulases sourced from fungi include GH5, GH6, GH9, GH45, and GH48, of which GH6 and GH45 were only found in eukMAGs. CAZymes, GH48 and GH6 are well-known abundant proteins in fungal cellulosomes (Haitjema *et al.*, 2017b). Major hemicellulases common to both MAGs and

eukMAGs include GH5, GH10, GH26, and GH43; GH62 and GH98 were less commonly found CAZymes, which were restricted to bacterial MAGs. CAZyme GH62 act on xylose moieties in xylan and arabinose moieties in arabinan (Wilkens *et al.*, 2017), and GH98 are endo-β-galactosidases (Rigden, 2005). Therefore, there is evidence to support some functional complementarity of the CAZymes contributed by anaerobic fungi and bacteria. Notably, *Neocallimastix* eukMAGs and MAGs from the anaerobic bacteria *Ruminococcus* and *Clostridium* also contained dockerin-associated CAZymes, indicating the potential to produce cellulosomes. Cellulosomes in anaerobic environments are multi-enzyme complexes deployed by all known anaerobic fungi and a small group of anaerobic bacteria, which are suspected to assist in synergistic breakdown of lignocellulose through enzyme tethering and rearrangement on a flexible protein scaffold (Resch *et al.*, 2013; Artzi, Bayer and Moraïs, 2017).

2.2.3 Metabolic potential for biomass deconstruction resolved at the species level

The reconstruction and annotation of 719 MAGs and 20 eukMAGs enabled estimation of metabolic potential of the microbial community members at the species level. This helped decipher the functional compartmentalization and redundancy among consortia members during biomass breakdown and fermentation. Quantification of major metabolic products provided validation of reconstructed metabolism and benchmarked the performance of each enriched consortium. MAGs-based analysis indicated that consortia membership is heavily shaped by the substrate used during enrichment. For example, in the antibiotics-free consortium grown on the most lignin-rich substrate, bagasse, the most abundant MAGs (*Lachnospiraceae sp. G11* and *Pseudobutyrivibrio sp. AR14*) were not enriched in antibiotics-

free consortium grown on alfalfa or reed canary grass. Conversely, a different MAG from the *Lachnospiraceae* family (Lac1) was abundant in antibiotics-free consortia grown on alfalfa and reed canary grass but absent in the antibiotics-free consortium grown on bagasse. Methanol-utilizing archaea *Thermoplasmata* and *Methanosphaera stadtmanae* were only enriched in antibiotics-free consortia grown on alfalfa which has the highest pectin content among all four substrates and one of the degradation products of pectin is methanol. In summary, substrate type selects for a suite of species equipped with metabolism to break down the corresponding carbon substrate and utilize the breakdown products for fermentation and methanogenesis.

2.2.4 Methane production is elevated in fungus-dominated consortia

Metabolic characterization based on assembled MAGs demonstrates different potential biomass degradation strategies and putative carbon flow between bacteria-dominated and fungus-dominated consortia, as shown in Figure 2.3. Some of the rare MAGs (<1% in relative abundance) harbor metabolic potentials that are not redundant compared to the more abundant members in the consortium. In antibiotics-free consortia, there was a high degree of functional redundancy among cellulolytic and fermentative bacteria from different phyla, whereas in antibiotics-treated consortia, anaerobic fungi dominated consortia membership. Methanogenic archaea were enriched to one of the most abundant prokaryotic members in PS consortia wherein carbon was not diverted by bacteria to produce propionate and butyrate and as a result, PS consortia produced the highest amount of CH_4, as shown in Figure 2.4.

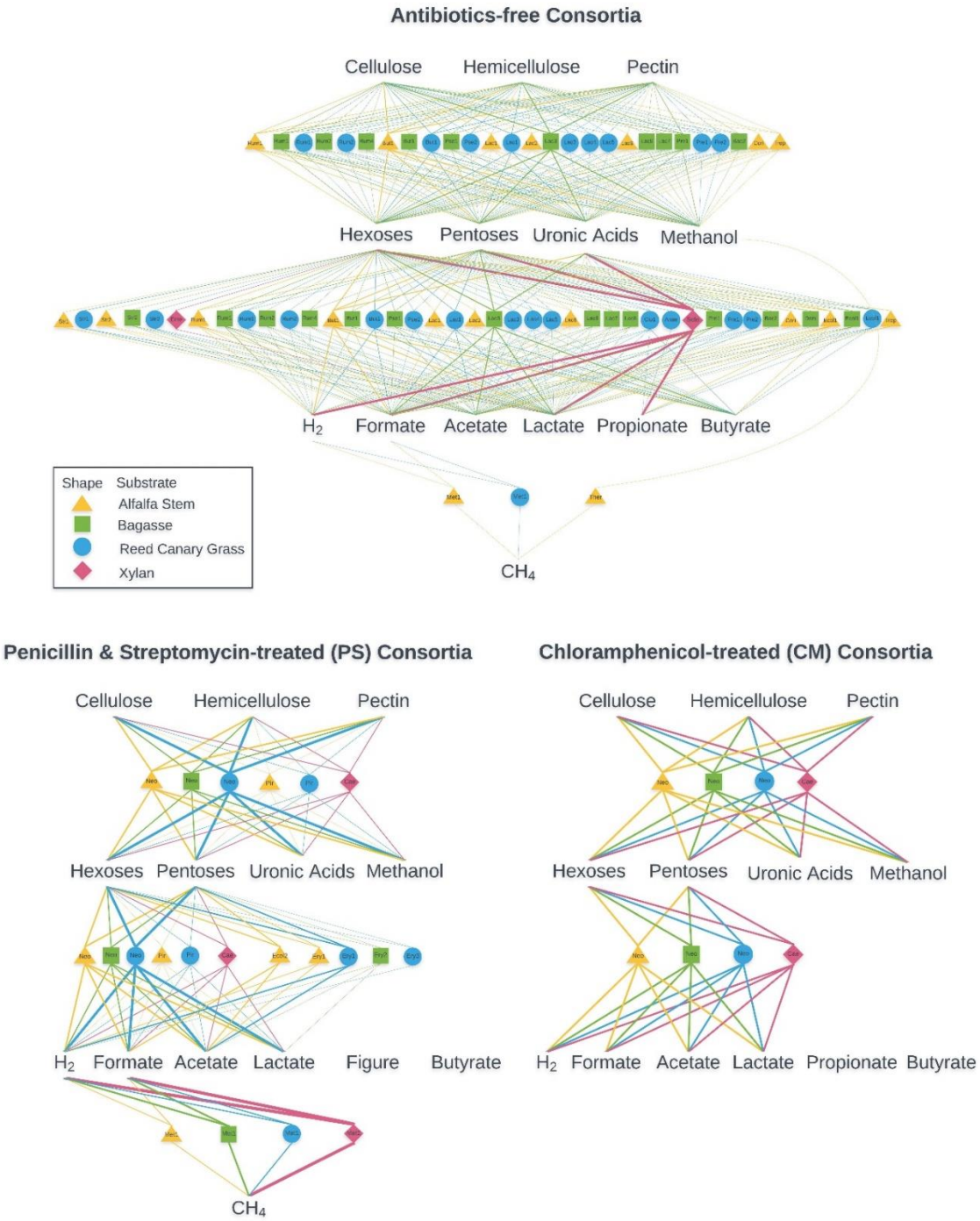

Figure 2.3. Carbon cross-feeding between microorganisms in enriched anaerobic consortia. Each shape contains a three-to-five-letter acronym representing a metagenome-assembled genome (MAG) contributing to > 1% to total community in each indicated consortium at the end of the tenth generation

(G10). Each shape corresponds to a type of carbon substrate. Triangles: Alfalfa stem, Squares: Bagasse, Circles: Reed Canary Grass, Diamonds: Xylan. The thickness of the lines is scaled with the relative microbial abundance of the connected MAG in the corresponding consortium. A line is connected between a MAG and a metabolite if the pathway responsible for the utilization/production of the metabolite is at least 75% complete in the reconstructed MAG. The taxonomic acronyms for each MAG used in the figure are: Anae: Anaerovibrio sp., Bac2: Bacteroidales, But1: Butyrivibrio sp., Clo1: Clostridium cochlearium, Cori: Coriobacteriaceae sp., Ecol1, Ecol2: Escherichia coli, Ente: Enterococcus faecium, Ery1, Ery2, Ery3: Erysipelotrichaceae, Lac1, Lac2, Lac4, Lac5, Lac8: Lachnospiraceae, Lac3: Lachnospiraceae sp. G11, Lac6: Lachnoclostridium clostridioforme, Lac7: Lachnospiraceae sp. JC7, Met1: Methanobrevibacter thaueri, Met2: Methanobrevibacter millerae, Pre1, Pre2: Prevotella ruminicola, Pse1: Pseudobutyrivibrio sp. AR14, Pse2: Pseudobutyrivibrio ruminis, Rum1: Ruminococcus albus, Rum2: Ruminococcus flaveflaciens, Rum4: Ruminococcaceae, Sele: Selenomonas ruminantium, Str1: Streptococcus equinus, Str2: Streptococcus gallolyticus, Ther: Thermoplasmata, Trep: Treponema sp., Neo: Neocallimastix sp., Pir: Piromyces sp., Cae: Caecomyces sp.

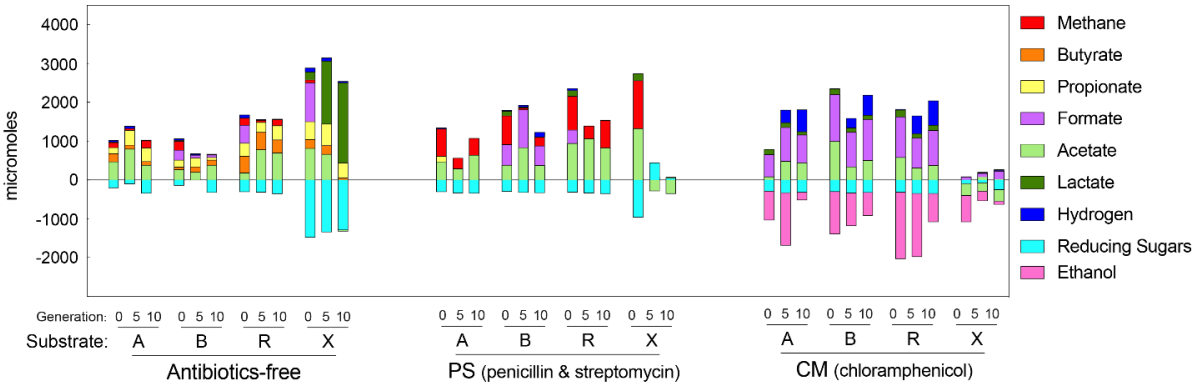

Figure 2.4. Net change in primary metabolic products from day 0 to day 3 in anaerobic consortia during the course of parallel enrichments. Measurements are grouped by antibiotic treatment and the type of carbon substrate used to drive enrichment. The three bars for each subgroup represent measurements made at generation zero G0 (0), generation 5 G5 (5), and generation 10 G10 (10). Substrates include

alfalfa stems (A), bagasse (B), reed canary grass (R), and xylan (X). Each bar represents the average of three biological replicates. Negative values indicate a net decrease of the metabolic product from day 0 to day 3.

When grown on alfalfa, PS consortia produced nearly twice as much methane as the antibiotics-free consortia, as shown in Figure 2.5. Very little H_2 accumulation was observed in the PS consortia, whereas a small amount of H_2 build-up occurred in antibiotics-free consortia, suggesting a more efficient metabolic product exchange in PS consortia than in antibiotics-free consortia. As expected, Chloramphenicol-treated (CM) consortia did not produce CH_4 but produced H_2 due to the presence of anaerobic fungi and the absence of methanogens.

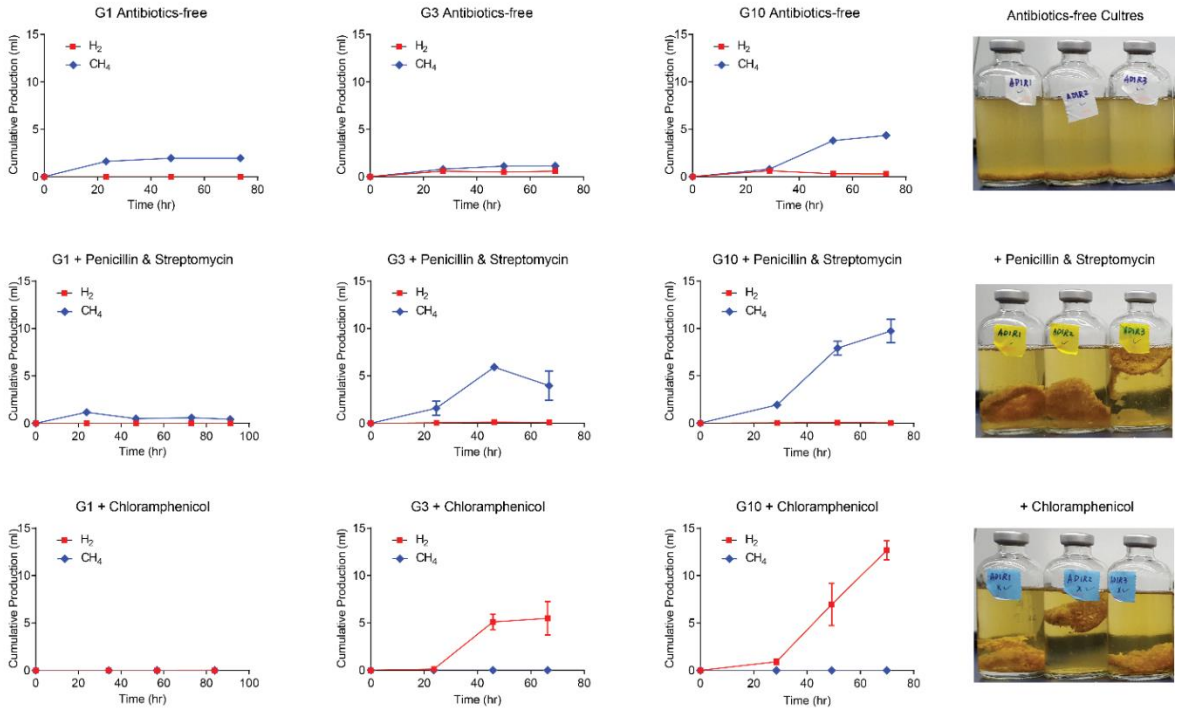

Figure 2.5. Cumulative production (ml) of hydrogen (H_2) and methane (CH_4) by enrichment cultures grown on alfalfa stems at generations 1 (G1), 3 (G3), and 10 (G10). Error bars represent standard deviations of three replicates. Photographs of the cultures from generation 1 (G1) are shown on the

right. Results for antibiotics-free cultures are shown in the top row; the turbid liquid media are characteristic of bacterial growth. Results for penicillin and streptomycin-treated cultures are shown in the middle row, and results for chloramphenicol-treated cultures are shown in the bottom row. The clear liquid media in antibiotics-treated cultures indicate low prokaryotic abundance. The clumped alfalfa stems floating in the liquid media are characteristic of anaerobic fungal growth, as fungi associate directly with the substrate.

2.2.5 A high degree of functional redundancy is seen in antibiotics-free consortia

In antibiotics-free consortia, most enriched bacteria likely occupy a mixed trophic level with the dual capability of degrading plant cell walls and fermenting simple sugars, resulting a high degree of functional redundancy with up to 44 bacteria capable of hydrolysis and 78 bacteria capable of fermentation. Abundant cellulolytic and hemicellulolytic bacteria belonging to the genera *Ruminococcus*, *Prevotella*, *Butyrivibrio*, *Pseudobutyrivibrio*, and a few other *Lachnospiraceae* bacteria were enriched. Typical of gut microbial communities, most of these bacteria can produce formate, acetate, and lactate (Wolin, 1981). Although less than half of the microbial community can produce butyrate and less than 20% of the community can produce propionate, the potential for butyrate and propionate production is redundantly spread across four bacterial phyla including *Proteobacteria*, *Actinobacteria*, *Firmicutes*, and *Bacteroidetes*.

The antibiotics-free consortium grown on xylan was dominated (72%) by *Selenomonas ruminantium*, with the presence of another enriched bacteria, *Enterococcus faecium*. There were large amounts of reducing sugars available in xylan culture media relative to complex fibers, which likely contributed to the very low microbial diversity observed. The moderate level of CH_4 production and the high production of propionate and butyrate clearly indicate

that carbon flow in antibiotics-free consortia was diverted towards short-chain fatty acids (SCFAs) instead of CH_4 compared to PS consortia.

2.2.6 Narrowed fermentation products in PS consortia leads to higher CH_4 production

In the antibiotics-treated consortia, anaerobic fungi use a large suite of biomass-degrading enzymes to depolymerize lignocellulose and ferment mono- and oligo-saccharides into H_2, formate, acetate, lactate, and ethanol (Lowe, Theodorou and Trinci, 1987b; Solomon *et al.*, 2016). The overall fermentation product profiles in CM consortia grown on lignocellulose were similar to those of fungal monocultures of *Neocallimastix* (Lowe, Theodorou and Trinci, 1987b), with significant accumulation of formate and H_2 and small amounts of lactate and no accumulation of reducing sugars, as shown in Figure 2.4. In PS consortia, archaea from the genus *Methanobrevibacer* were enriched to be one of the most abundant prokaryotes as most bacterial growth was repressed by penicillin and streptomycin. Methanogens use H_2 and produce CH_4, accounting for the different metabolic product profile in PS consortia than in CM consortia. Seven *Firmicutes* MAGs were recovered in PS consortia, and *Erysipelotrichaceae* was the most abundant among all of them. These PS-resistant bacteria can utilize hexose sugars and produce formate, acetate, and lactate. Hence they might contribute to preventing catabolite repression of anaerobic fungi by maintaining consistent but low levels of simple sugars (Solomon *et al.*, 2016).

2.2.7 Fungus-methanogen partnership demonstrates accelerated cellulose degradation potential

A key motivation in characterizing the herbivore microbiota is to understand how to control lignocellulosic degradation, as well as to regulate the release of methane. Given the

high levels of CH_4 production observed in PS-treated consortia compared to antibiotics-free consortia (see Figure 2.4), the cellulolytic performance of these enriched consortia was compared. When grown on cellulose paper (Whatman), the PS consortium degraded nearly twice as much substrate as the antibiotics-free consortium after seven days of growth, as shown in Figure 2.6A. Excess reducing sugars were released from the PS consortium but not from the antibiotics-free consortium when grown on cellulose paper, as shown in Figure 2.6B. By contrast, the PS consortium and the antibiotics-free consortium degraded comparable amounts of reed canary grass after seven days of growth (see Figure 2.6A). This suggests that despite the advantage of PS over antibiotics-free consortia in degrading cellulose paper, complex substrates like plant cell walls limit the performance of anaerobic fungi in depolymerization.

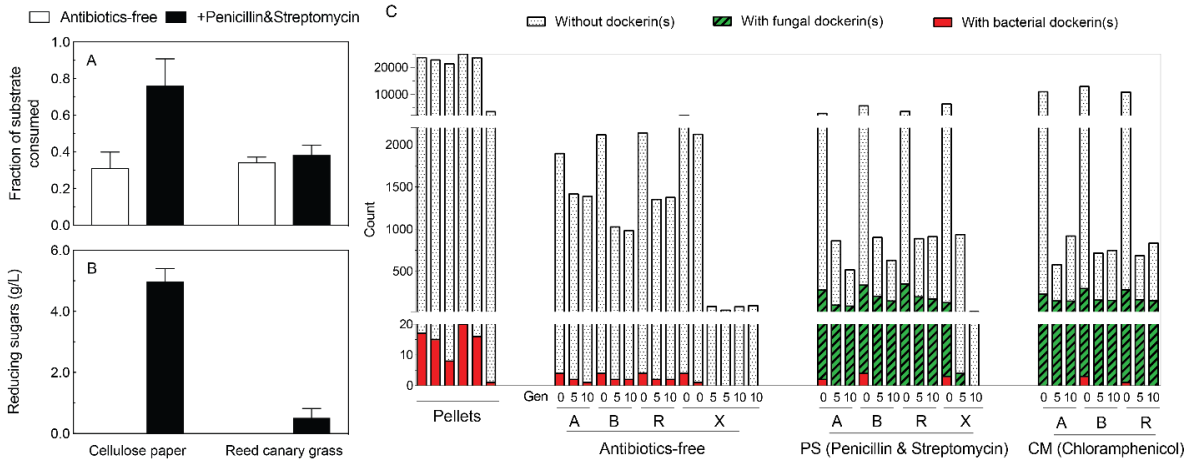

Figure 2.6. Fungal CAZymes and cellulosomes drive lignocellulosic efficiency of consortia dominated by anaerobic fungi and methanogenic archaea. The fraction of substrate consumed (A) and the release of reducing sugars (B) by enrichment cultures grown on cellulose paper and reed canary grass. Each bar in panels A and B represents the average of three replicates and the error bars represent the standard deviation. Panel C shows number of carbohydrate-active enzymes (CAZymes) classified as cellulase, hemicellulase, pectinase, or esterase in enrichment cultures, distinguished by the presence or absence

of bacterial or fungal dockerin domains fused to CAZymes. Note that the y-axis was broken into three different scales. Red boxes represent bacterial CAZymes associated with dockerin(s), green boxes with black slanted stripes represent fungal CAZymes associated with dockerin(s), and black dotted patterns represent CAZymes unassociated with dockerin(s).

The enzymatic strategies for hydrolyzing lignocellulose deployed by PS and antibiotics-free consortia were compared by enumerating the number of cellulase, hemicellulase, pectinase, and esterases with and without dockerin associated domains in each consortium. We found that the number of cellulosomal CAZymes in PS consortia grown on lignocellulosic substrates was more than two orders of magnitude higher than that found in antibiotics-free consortia, as shown in Figure 2.6C. By comparison, the total number of cellulase, hemicellulase, pectinase, and esterase in antibiotics-free consortia grown on lignocellulosic substrates (983-1417 in G5 and G10) was higher than that in PS consortia (518-909 in G5 and G10). The larger number of CAZymes observed in antibiotics-free consortia compared to PS consortia is attributed to the large numbers of GH5, GH8, GH9, GH16, GH26, GH30, GH43, GH28, CE4, and CE8. Nevertheless, GH6 and GH45 (cellulases) were only enriched in PS consortia, and there were a larger number of GH48, GH11, and PL3 in PS consortia than antibiotics-free consortia.

Additional experiments were performed to determine the long-term stability of the best-performing consortium (PS consortium grown on alfalfa) in CH_4 production. This consortium was maintained in lab for over three years while consistently producing methane. The eukaryotic member of the consortium was an anaerobic fungus from the genus *Neocallimastix*, and the dominant bacterial members were bacteria from the family *Erysipelotrichaceae*. *Methanobrevibacter* were the archaeal member of the consortium responsible for methane

generation. *Ruminococcus* were present as a rare bacterial member of the community. Furthermore, both the prokaryotic and eukaryotic parts of this consortium were stable after cryo-preservation at -80 °C for over a year. Importantly, this consortium produced methane after reviving from cryo-preservation.

2.3 Conclusions

Here we have presented an in-depth analysis of the goat fecal microbiome and a large-scale enrichment experiment that suggests design rules of microbial consortia pairings for biomass conversion into value-added products. The vast majority (96%) of the 719 high-quality MAGs we recovered are novel at the species level compared to previously published MAGs from similar microbiomes (rumen, feces, and anaerobic digesters), and about 20% of them were enriched in our experiments, along with anaerobic fungi from the phylum *Neocallimastigomycota*. The metabolic product profile comparison of consortia treated with and without antibiotics revealed that the level of CH_4 production was the highest when fermentation products are restricted to H_2, formate, and acetate, as opposed to being diverted to the production of butyrate and propionate. Comparing the resultant communities by challenging the source microbiome with four different types of carbon sources indicate that substrate composition plays a critical role in determining the community composition. Specifically, simple reducing sugars enriched for a consortium featuring a small number of specialists that dominated the community, whereas complex plant material substrates enriched for many functionally redundant lignocellulolytic members. A consortium featuring cross-domain partnership between anaerobic fungi, methanogenic archaea, and bacteria demonstrated the highest cellulose degradation potential, and this high performance is

attributed to the combination of physical breakdown of plant substrate by fungal rhizomycelia and the cellulosomal CAZymes. The lessons learned from this large-scale enrichment experiment will be valuable for the design of lignocellulolytic consortia for the bioconversion of lignocellulose into value added chemicals.

2.4 Methods

Freshly voided fecal pellets were collected from a San Clemente Island Goat at the Santa Barbara Zoo, which served as source material for 396 parallel anaerobic enrichment experiments (see Figure 2.1). This was initiated with the inoculation of 36 cultures with the source material, which were supported on four different types of carbon substrates in complex anaerobic culture medium (MC-) with a carbon dioxide headspace (Peng *et al.*, 2018): alfalfa stems (A), bagasse (B), reed canary grass (R), and xylan (X). A, B and R were crude plant material ground to 4 mm in size, and X was manufactured from corn core in powder form. Chloramphenicol (CM) was applied to one group of cultures to bias selection for anaerobic fungi; penicillin and streptomycin (PS) were applied to a second group of cultures to bias selection of anaerobic fungi and methanogenic archaea; and an antibiotics-free group received no antibiotic selection (see Figure 2.1). In this manuscript, the initial generation of enrichment cultures are referred to as “G0”, and the subsequent generations were referred to by their consecutive batch culture numbers (“G1”, “G2”, etc. to “G10”). The initial enrichment cultures were sub-cultured (10% v/v) without pooling into fresh media and appropriate carbon substrate every three days with the exceptions of G0 and G1, which were allowed to grow for five and four days, respectively, before sub-culturing to enable maximum development of the community (Theodorou, Gascoyne and Beever, 1984). Culture growth was monitored daily

by sampling the headspace of the anaerobically sealed bottles to measure the production of total pressure, hydrogen (H_2) and methane (CH_4) concentrations. Concurrently, 1 ml of the liquid media was also sampled daily to measure the production of metabolites (short-chain fatty acids and reducing sugars). After sub-culturing into fresh media at the end of three days of growth, the remainder of the cultures were harvested for nucleic acids extraction. High-resolution marker gene (16S, 18S, and ITS) analysis was performed on an Illumina MiSeq sequencer (300 bp x 2) for six of the 11 generations (G0, G1, G3, G5, G8, and G10) to track enrichment of the community under different selective pressures. Deep metagenome sequencing was performed on an Illumina HiSeq sequencer (150 bp x 2) for the initial (G0), middle (G5), and final (G10) enrichments for all substrates. The total number of reads summed to over 1.5 Tbp (10^{12} base pairs). Raw reads were quality filtered using Trimmomatic (Bolger, Lohse and Usadel, 2014), assembled using SPAdes (Nurk *et al.*, 2017), and the contigs were binned using Metabat 2 (Kang *et al.*, 2019) and were annotated using IMG/M (Chen *et al.*, 2017). A custom bioinformatics pipeline that combines marker gene and metagenome sequencing was developed to analyze the metabolic potential of each metagenome assembly, including the prevalence of Carbohydrate Active Enzymes (CAZymes) and other metabolic capabilities related to carbohydrate uptake. Metabolic pathways consisting of multiple genes are considered present if it is at least 75% complete in each MAG and eukMAG. Additional experiments were performed to compare the hydrolytic potential of microbial consortia untreated and treated with antibiotics.

2.5 Data Availability

The metagenome sequencing reads can be accessed at the Joint Genome Institute. Contigs for each MAG are available at NCBI's Whole Genome Shotgun database under accession numbers SAMN11294286 - SAMN11295004 and project number PRJNA530070.

2.6 Acknowledgements

The authors are grateful for funding support from the National Science Foundation (NSF) (MCB-1553721), the Office of Science (BER) the US Department of Energy (DOE) (DE-SC0010352), the Institute for Collaborative Biotechnologies through grants W911NF-09–D-0001 and W911NF-19-2-0026 from the US Army Research Office, and the Camille Dreyfus Teacher-Scholar Awards Program. Authors also acknowledge support from the California NanoSystems Institute (CNSI) Challenge Grant Program, supported by the University of California, Santa Barbara and the University of California, Office of the President. This work was part of the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the Office of Biological and Environmental Research of the DOE Office of Science through contract DE-AC02–05CH11231 between Lawrence Berkeley National Laboratory and the DOE. The sequencing conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). Additional

supercomputing resources were provided by Center for Scientific Computing at UCSB, which is supported by NSF Grant CNS-0960316.

III. Genomic and proteomic biases inform metabolic engineering strategies for anaerobic fungi

This chapter is based upon an article that was published in Metabolic Engineering Communications, Volume 10, 2019, by St. Elmo Wilken, Susanna Seppälä, Thomas S. Lankiewicz, Mohan Saxena, John K. Henske, Asaf A. Salamov, Igor V. Grigoriev, and Michelle A. O'Malley, entitled "*Genomic and proteomic biases inform metabolic engineering strategies for anaerobic fungi"*, Copyright Elsevier. For more information regarding the methods and results, please see the published paper.

3.1 Introduction

Metabolic engineering strives to streamline and sculpt microorganisms for the optimal production of valuable fuels and chemicals. To date, most metabolic engineering efforts have targeted well-characterized microorganisms such as *E. coli* and *S. cerevisiae*, but it is well recognized that non-model microorganisms hold tremendous biotechnological potential (Bonugli-Santos *et al.*, 2015; Coker, 2016; Susanna Seppälä *et al.*, 2017; Podolsky *et al.*, 2019). In this regard, the anaerobic fungi in the phylum Neocallimastigomycota possess an unparalleled collection of carbohydrate active enzymes (CAZymes) that can be leveraged to convert plant biomass into value-added commodity and fine chemicals (Youssef *et al.*, 2013; Morrison, Elshahed and Youssef, 2016; Solomon *et al.*, 2016; Haitjema *et al.*, 2017a). The Neocallimastigomycota fungi are primarily found in the digestive tracts of herbivorous animals where they break down ingested lignocellulosic plant biomass (Orpin, 1975; Theodorou *et al.*, 1996; Liggenstoffer *et al.*, 2010) and although their importance for animal

welfare is well established, anaerobic gut fungi have not yet been adapted for metabolic engineering or bioprocessing applications.

To fully exploit the biotechnological potential of anaerobic fungi, it is first necessary to understand the functional properties of their proteins, especially their diverse set of biotechnologically important CAZymes (S Seppälä *et al.*, 2017; Podolsky *et al.*, 2019). To achieve this goal, there is a critical need to (1) develop strategies to transfer gut fungal genes to heterologous hosts, and (2) develop molecular tools to modify the genomic content of the gut fungi. The recently acquired high-resolution transcriptomes and genomes of several gut fungal strains aid in this regard as they not only reveal the enzymatic and proteomic potential of these fungi, but also the genomic guanine-cytosine (GC)/adenine-thymine (AT) nucleotide content, apparent codon-usage patterns, and the amino acid composition of encoded proteins (Youssef *et al.*, 2013; Solomon *et al.*, 2016; Haitjema *et al.*, 2017a; Henske, Wilken, *et al.*, 2018). In particular, the GC content of any genome often dictates genetic engineering strategies, whether the aim is to transfer genes to a more easily manipulated organism or to engineer the genome of the non-model organism directly.

While a handful of gut fungal proteins have already been produced in model microorganisms (reviewed in (S Seppälä *et al.*, 2017))(O'Malley, Theodorou and Kaiser, 2012b; Dollhofer *et al.*, 2019), the vast majority remains uncharacterized, and recent reports suggest that at least some gut fungal genes must be codon optimized for the successful expression in heterologous hosts like yeast (Solomon *et al.*, 2016; Seppälä *et al.*, 2019). Likewise, production of non-native proteins (e.g. reporter proteins) and exogenous metabolic pathways in the anaerobic fungi is likely aided if the codon composition of the exogeneous gene is matched to the apparent codon preference of the host, as has been demonstrated in

other fungi (Wang *et al.*, 2019). Moreover, the genomic nucleotide composition may affect how efficiently a genome can be engineered using endonucleases that recognize specific, often G-rich, nucleic acid motifs, such as the increasingly popular Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-CRISPR Associated protein (Cas) system.

Here, we have analyzed the genomes and proteomes of anaerobic fungi to establish a framework for metabolic engineering in these non-model organisms. The nucleotide composition of all published anaerobic fungal genomes was used to determine gut fungal codon preferences, amino acid distributions, and identify low-complexity regions in the predicted proteomes with emphasis on their rich repertoire of biotechnologically interesting CAZymes. Currently, five species of anaerobic fungi have published nuclear genomes: *Orpinomyces sp.* C1A (later re-classified as *Pecoramyces ruminantium*) (Youssef *et al.*, 2013; Hanafy *et al.*, 2017), *Piromyces sp.* E2, and the high-quality genomes of *Neocallimastix californiae*, *Anaeromyces robustus* and *Piromyces finnis* (Haitjema *et al.*, 2017a). Using the Joint Genome Institute's (JGI) Mycocosm fungal genomes repository, we compared these anaerobic fungal genomes to 438 other sequenced fungi, spanning the fungal tree of life (Grigoriev *et al.*, 2014). Overall, we find that the coding genomes of anaerobic fungi are exceptionally GC depleted, which significantly impacts codon and amino acid usage in anaerobic gut fungi and limits the application of certain CRISPR variants. Based on these native biases, we introduce a codon optimization table for use in expressing non-native genes in the gut fungi. Analysis of the genomes also reveals genetic machinery implicated in sexual reproduction, and shows that gut fungal CAZymes are highly enriched in repetitive sequences that are linked to glycosylation motifs. Overall, this comparative analysis will aid in the

development of metabolic engineering strategies by identifying common pitfalls and suggesting possible solutions to genetically manipulate Neocallimastigomycota fungi.

3.2 Results and discussion

3.2.1 Anaerobic gut fungi have the most GC depleted genomes in the fungal kingdom

Biased genomic GC content has significant implications for modern genome sequencing and engineering techniques. For example, it has been shown that regions with extreme nucleotide content hamper next-generation sequencing techniques owing to poor read coverage and difficulties in assembly (Oyola *et al.*, 2012). Moreover, the apparent preferred codon usage of an organism may affect how efficiently genes can be transferred between organisms, in particular those that exhibit extreme codon biases (Seppälä *et al.*, 2019). An analysis of 443 published fungal genomes, sourced from the JGI Mycocosm database and covering 278 genera from across the fungal kingdom, reveals a large variation in the GC content of fungal protein coding genomes, ranging from ~25% GC to ~69% GC as shown in Figure 3.1. Among these, the obligate anaerobic Neocallimastigomycota consistently have the most GC depleted coding genomes of all sequenced fungi, ranging from ~25% GC in *A. robustus* to ~29% GC in *Piromyces sp.* E2 (Youssef *et al.*, 2013; Haitjema *et al.*, 2017a). The GC content of the intergenic, non-coding regions in the anaerobic gut fungi is even lower (~16% on average): causing the whole-genome GC content of Neocallimastigomycota to range from ~16% to ~22%. This peculiar nucleotide composition of anaerobic fungi was suggested by thermal denaturation studies more than two decades ago, and is a contributing factor to why the first high-resolution genomes were only recently acquired via long-read

sequencing technologies (Brownlee, 1989; Nicholson, Theodorou and Brookman, 2005; Oyola *et al.*, 2012).

Figure 3.1 also illustrates that the GC content of the fungal protein coding genomes is not readily explained by phylogenetic relationships, as has also been noted for other kingdoms (Knight, Freeland and Landweber, 2001; Wu *et al.*, 2012; Reichenberger *et al.*, 2015). For example, while Neocallimastigomycota appears to be the most GC depleted fungal phylum, the phylogenetically related Chytridiomycetes is rather GC rich at ~56% based on 4 genomes from 4 genera. Possibly confounding this analysis is the number of sequenced genomes analyzed in each clade, as some clades are extremely under-sampled to date.

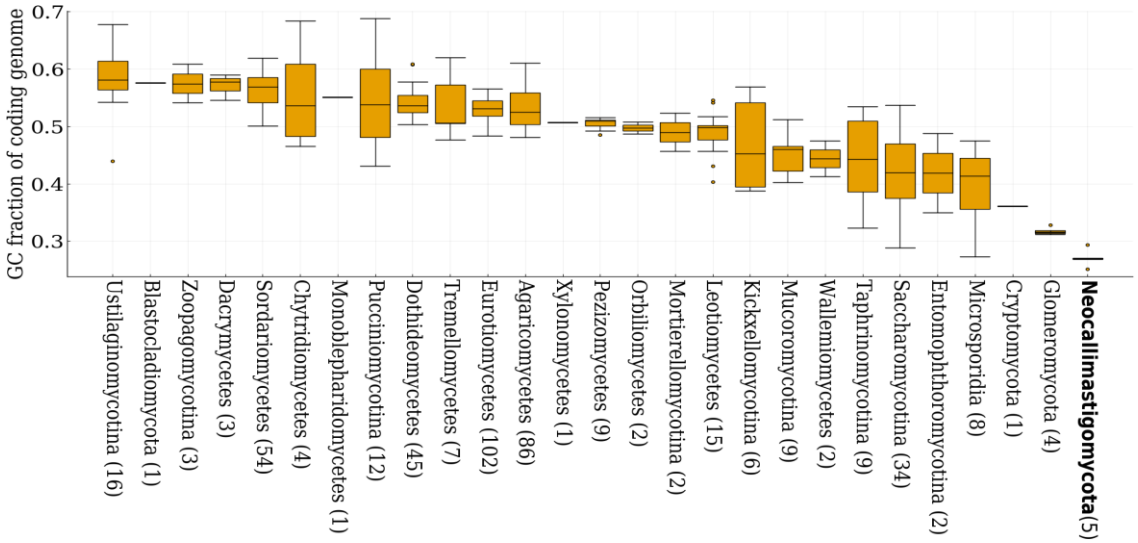

Figure 3.1: Neocallimastigomycota are characterized by extremely GC-depleted genomes and proteomes. GC content within the predicted proteome of 443 fungal genomes is plotted as a function of fungal clade and varies significantly across the fungal kingdom. The number of species analyzed per clade is indicated in brackets on the x-axis. The box-and-whisker plots show outliers as points, minima and maxima as whiskers, and the inter-quartile ranges inside the boxes. Figure taken from (Wilken *et al.*, 2020).

3.2.2 The extreme AT-richness of anaerobic fungal genomes limits CRISPR and other genetic engineering strategies

The GC content of a genome can be affected by mutation rates, recombination as well as by selection (Birdsell, 2002; Duret and Galtier, 2009; Hershberg and Petrov, 2010; Hildebrand, Meyer and Eyre-Walker, 2010). While it remains unclear how the anaerobic gut fungal genomes became GC depleted, their AT-richness has several implications for genetic engineering approaches. For example, plasmid-based expression systems are hampered by difficulties associated with identifying promotors and regulatory elements in these genomes, and AT-rich sequences are associated with non-specific binding affecting both primer design and homologous recombination approaches.

Moreover, most genome editing approaches use specific DNA sequence motifs to guide nucleases to the genome. For example, technologies using transcription-factor-like-endonucleases (TALENs) (Arazoe *et al.*, 2015) and zinc finger nucleases (ZFN) (Boch *et al.*, 2009) depend on DNA-binding protein domains, and the CRISPR-Cas9 system depends on a guide-RNA that brings the nuclease to the desired site (Gasiunas *et al.*, 2012). Genomes that contain regions with extreme nucleotide content may cause poor or nonspecific targeting. For example, the canonical CRISPR-Cas9 system makes use of a G-rich (NGG) protospacer-adjacent motif (PAMs) to target genes for editing (Jiang *et al.*, 2013), however recent engineering efforts have broadened the diversity of PAM sites that can be targeted to include: TTN (Zetsche *et al.*, 2015); NGG, NGA, NGAG, and NGCG (Benjamin P. Kleinstiver *et al.*, 2015); as well as NNNRRT (Benjamin P Kleinstiver *et al.*, 2015) among others. Table 3.1 shows that the frequency of observing GC-rich PAMs increases accordingly with genomic GC content. Conversely, PAMs that are richer in AT bases are much more abundant in

genomes with lower GC content. The relative paucity of GC-rich PAM sites in anaerobic fungal genomes is likely to limit the ability of certain endonucleases to target specific positions of interest, and suggests that AT-rich PAM targeting Cas enzymes may be the most appropriate choice for CRISPR engineering efforts.

Table 3.1: Increasing GC content of fungal genomes increases the number of PAM sequences with higher GC content. The number of PAMs (PAM sequences ordered in decreasing GC content from left to right) found per mega base pair in the genomes (coding and non-coding regions) of fungi in Neocallimastigomycota, as well as *Saccharomyces cerevisiae*, *Trichoderma reesei* and *Rhodotorula graminis*, ordered in increasing GC content (%). The color scale shows the abundance of PAM sequences found in the genome of each fungus, with darker cells corresponding to more PAM sequences identified.

Fungus	PAM/GC	NGG	NGCG	NGAG	NGA	NNNRRT	TTN
A. robustus	16.3	9924	433	2700	32869	73134	111343
P. ruminantium	17	11265	578	3067	34046	74168	110050
N. californiae	18.2	11957	754	3667	35703	72761	108990
P. finnis	21.2	14530	863	4189	40390	72228	104066
P. sp. E2	21.8	142327	99825	101766	157469	109692	61015
S. cerevisiae	38.2	33865	6749	11565	59981	60548	72583
T. reesei	52.8	52350	15407	18713	63539	44932	44227
R. graminis	68.9	68959	40830	34522	76667	29472	19288

On the other hand, one of the most GC-depleted non-fungal eukaryotic microbes are the Apicomplexan *Plasmodium spp.*, which include avian and human malaria parasites *P. gallinaceum* (~21% GC) and *P. falciparum* (~24% GC) (Videvall, 2018). Recently, it was

suggested that the genome of *P. falciparum* undergoes an extremely high rate of mutations, associated with sequences that have an extreme GC or AT bias, and that this phenomenon may contribute to adaptive evolution (Hamilton *et al.*, 2017). Although the mutation rate of Neocallimastigomycota is unknown, it is tempting to speculate that similar mechanisms are in place for anaerobic fungi, possibly facilitating horizontal gene transfer of enzymes from ruminal bacteria to the anaerobic fungi (Haitjema *et al.*, 2017a; Duarte and Huynen, 2019; Murphy *et al.*, 2019) that could be harnessed for genome engineering. Nevertheless, the possibility of high mutation rates could negatively impact the efficacy of the highly specific edits made by CRISPR-Cas like systems.

3.2.3 Anaerobic fungal genomes contain genes used in sexual reproduction

Sexual reproduction is often leveraged for engineering industrial fungal strains, thus identification of a putative mating pathway in anaerobic gut fungi could inform future approaches to generate genetic variants (Steensels, Meersman, *et al.*, 2014; Steensels, Snoek, *et al.*, 2014; Mertens *et al.*, 2015; Solieri *et al.*, 2015). Inducing breeding events in yeast strains, and other biotechnologically relevant fungi, rapidly increases the diversity of mutant libraries through naturally occurring homologous recombination. This mode of diversity generation has advantages over direct genetic engineering; it is straightforward, rapid, and generates genetic variants that are not considered as genetically modified organisms by regulatory frameworks (Steensels, Snoek, *et al.*, 2014).

Sexual reproduction is associated with several genomic signatures, including the presence of genes required for mating events and GC content enrichment in genomes and genomic regions that are prone to homologous recombination (Hull, Raisner and Johnson, 2000;

Galtier, 2001; Magee, 2002; Meunier and Duret, 2004; Glémin, 2015; Ropars *et al.*, 2016; Kiktev *et al.*, 2018; Liu *et al.*, 2018). These genomic signatures have successfully been leveraged to interrogate the existence of sexual reproduction in other fungi (Hull, Raisner and Johnson, 2000). Further, the positive relationship between GC content and the different rates of outcrossing among fungi could help rationalize why GC content within the fungal kingdom is not readily explained by phylogeny, see Figure 3.1 (Hartfield, 2016; Nieuwenhuis and James, 2016). While many variables likely influence GC content in fungal genomes, the gut fungi standout as being particularly GC depleted, possibly suggesting very infrequent outcrossing.

However, some organisms with GC content near that of Neocallimastigomycota were until recently erroneously thought to be asexual, bolstering the idea that the anerobic gut fungi might be able to outcross. For example, sexual reproduction was demonstrated in the opportunistic human pathogen *Candida albicans* (35% GC) (Hull, Raisner and Johnson, 2000; Magee, 2002). Likewise, fungi in the phylum Glomeromycota (~32% GC) were also thought to be asexual, yet recent sequencing of several genomes revealed genes encoding the molecular machinery required for sexual reproduction (Ropars *et al.*, 2016). These findings seem to support the hypothesis that sexual reproduction is an ability shared by all fungi, even those that infrequently outcross.

While experimental evidence has thus far failed to confirm a sexual cycle in anaerobic gut fungi, we find genes with high homology to sex-implicated proteins in every high quality anaerobic fungal genome sequenced to date. Using well characterized proteins from *Saccharomyces cerevisiae*, we are able to identify homologs to kinases and accessory proteins heavily implicated in sexual reproduction (STE20, STE6, GPA1), as well as several meiosis

specific genes such as the meiotic recombinase DMC1 (see Table S2 in the supplementary materials of the paper upon which this chapter is based). Notably absent in the Neocallimastigomycota genomes are genes homologous to those coding for peptide mating factors deployed by *S. cerevisiae*, but regions of homology to the *N. crasssa* mating type "a" pheromone are detected in each genome.

The presence of genes implicated in sexual reproduction indicate that Neocallimastigomycota can, or at one point in evolutionary time were able to, sexually reproduce. However, low GC content in anaerobic fungal genomes could imply that these organisms outcross with extreme discretion. Further experimentation is needed to determine whether sexual reproduction will be a useful tool to generate anaerobic fungal variants. Specifically, elucidation of viable signaling pathways that lead to induction of mating type cells should be tested to determine how such a mating event could be induced and used for metabolic engineering applications (Magee, 2002).

3.2.4 Codon usage preferences of Neocallimastigomycota are a recipe for genetic engineering and expression optimization

Although the fungi in Neocallimastigomycota are not yet genetically tractable, efforts are being made to develop genetic transformation methodologies (Durand *et al.*, 1997; Calkins *et al.*, 2018). The introduction of novel genes encoding reporter proteins and selection markers into anaerobic gut fungi will likely require codon optimization such that the gut fungal machinery properly maintains and decodes the material. Codon optimization will likely be an important tool to aid in this regard, as has been shown for other fungal clades (Camiolo *et al.*, 2019; Wang *et al.*, 2019). Analysis of the preferred codon usage in highly expressed genes in

Neocallimastigomycota suggests that the anaerobic fungi have a strong preference for AT-rich codons, see Table 3.2, consistent with their GC-depleted genomes. Table S3 in the supplementary material of the paper upon which this chapter is based, shows the individual codon usage for highly expressed transcripts, as well as the predicted tRNA counts, for both the fungi with transcriptomic expression level data available (Henske, Wilken, *et al.*, 2018).

While codon optimization may be necessary to express exogenous genes in the gut fungi, their very strong codon bias has implications for heterologous expression of gut fungal genes in model microorganisms. For example, codon optimization to increase the GC content of genes was shown to be a prerequisite for the expression of gut fungal genes in some hosts (Li *et al.*, 2007; Seppälä *et al.*, 2019). However, this does not seem to be a universal constraint as other genes from gut fungi have been expressed without codon optimization (Kuyper *et al.*, 2003; Wang *et al.*, 2011). Nevertheless, codon optimization may prove to be an important consideration for genetic exploitation of gut fungi because their genomes, and consequently their genes, are so extremely GC depleted. Interestingly, it is also observed that the most abundant codon does not always correspond to the most abundant associated tRNA. For example, across both anaerobic fungi analyzed, AAT is the most common asparagine codon, however only AAC (a synonymous asparagine codon) matching tRNAs (TTG anticodons) were identified on the genome. It is likely that tRNA wobbling in the third base position explains this phenomena, as had been noted in other filamentous fungi (W. Chen *et al.*, 2012).

3.2.5 Amino acids coded by AT rich codons are favored by Neocallimastigomycota

Amino acid composition is important for protein production, stability and post-translational modifications, which has implications for heterologous production. As shown in

Figure 3.2, there is a clear correlation between GC content and predicted amino acid distribution in the fungi. The GC depleted fungi, including the Neocallimastigomycota, appear to be enriched in amino acids that are encoded for by AT rich codons (lysine, isoleucine and asparagine) and depleted in amino acids that are encoded for by GC rich codons (alanine, glycine, arginine). Conversely, fungi with GC rich coding genomes have a higher proportion of amino acids that are encoded for by GC rich codons. This is consistent with previous cross-kingdom analyses suggesting that the relative abundances of amino acids in a proteome is largely determined by the GC content of the genome (Knight, Freeland and Landweber, 2001).

Table 3.2: Codon optimization table for Neocallimastigomycota. Fraction of the proteome encoded for by each codon in highly expressed transcripts of *N. californiae* and *A. robustus* averaged, with standard deviation noted. The most AT rich codon for each amino acid is shown in red font, while the most abundant codon within the transcriptome is shown in blue font. AT-rich codons are invariably preferred in anaerobic fungi, in line with the predicted low GC content of the clade.

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	F [TTT]: 0.60 ± 0.03	S [TCT]: 0.28 ± 0.00	Y [TAT]: 0.65 ± 0.03	C [TGT]: 0.62 ± 0.04	U
	F [TTC]: 0.40 ± 0.03	S [TCC]: 0.15 ± 0.02	Y [TAC]: 0.35 ± 0.03	C [TGC]: 0.38 ± 0.04	C
	L [TTA]: 0.34 ± 0.01	S [TCA]: 0.22 ± 0.01	STOP [TAA]: 0.56 ± 0.01	STOP [TGA]: 0.31 ± 0.04	A
	L [TTG]: 0.31 ± 0.0	S [TCG]: 0.08 ± 0.01	STOP [TAG]: 0.14 ± 0.02	W [TGG]: 1.0 ± 0.0	G
C	L [CTT]: 0.19 ± 0.02	P [CCT]: 0.17 ± 0.03	H [CAT]: 0.58 ± 0.03	R [CGT]: 0.20 ± 0.02	U
	L [CTC]: 0.08 ± 0.01	P [CCC]: 0.09 ± 0.02	H [CAC]: 0.27 ± 0.17	R [CGC]: 0.04 ± 0.01	C
	L [CTA]: 0.12 ± 0.01	P [CCA]: 0.45 ± 0.01	Q [CAA]: 0.78 ± 0.0	R [CGA]: 0.07 ± 0.02	A
	L [CTG]: 0.11 ± 0.01	P [CCG]: 0.08 ± 0.01	Q [CAG]: 0.22 ± 0.0	R [CGG]: 0.04 ± 0.02	G
A	I [ATT]: 0.49 ± 0.01	T [ACT]: 0.65 ± 0.06	N [AAT]: 0.69 ± 0.0	S [AGT]: 0.19 ± 0.03	U
	I [ATC]: 0.18 ± 0.01	T [ACC]: 0.15 ± 0.01	N [AAC]: 0.17 ± 0.0	S [AGC]: 0.10 ± 0.01	C
	I [ATA]: 0.36 ± 0.03	T [ACA]: 0.27 ± 0.02	K [AAA]: 0.67 ± 0.02	R [AGA]: 0.45 ± 0.06	A
	M [ATG]: 1.0 ± 0.0	T [ACG]: 0.10 ± 0.01	K [AAG]: 0.33 ± 0.02	R [AGG]: 0.19 ± 0.0	G
G	V [GTT]: 0.42 ± 0.0	A [GCT]: 0.67 ± 0.05	D [GAT]: 0.79 ± 0.03	G [GGT]: 0.62 ± 0.04	U
	V[GTC]: 0.15 ± 0.03	A [GCC]: 0.17 ± 0.03	D [GAC]: 0.21 ± 0.03	G [GGC]: 0.07 ± 0.02	C
	V[GTA]: 0.25 ± 0.02	A [GCA]: 0.28 ± 0.22	E [GAA]: 0.89 ± 0.01	G [GGA]: 0.23 ± 0.01	A
	V[GTG]: 0.19 ± 0.01	A [GCG]: 0.03 ± 0.01	E [GAG]: 0.11 ± 0.01	G [GGG]: 0.08 ± 0.01	G

Unique glycosylation patterns, influenced by the amino acid composition of a protein, are often difficult to mimic in heterologous hosts and may affect function (Gerngross, 2004). The high abundance of serine, threonine and asparagine in the gut fungal proteomes suggest that glycosylation could be an important component in protein production, and perhaps activity and stability. Additionally, amino acid composition of enzymes has been shown to correlate with, amongst other properties, thermal stability. The gut fungi grow optimally at 39°C, suggesting that their enzymes, and specifically their biotechnologically relevant CAZymes, are tailored for this temperature (Haitjema *et al.*, 2014). In contrast, *T. reesei* grows optimally at 28°C, but significant protein engineering efforts have improved the thermal stability of its cellulases to function in the range of 50 – 70°C (Chokhawala *et al.*, 2015). It is tempting to speculate that gut fungal enzymes may also be amenable to such engineering efforts and, combined with their natively very diverse cellulolytic enzyme repertoire, might increase the efficiency of high temperature biomass conversion processes even further. Finally, the nitrogen content of media has been shown to influence the growth rates of gut fungi, possibly due to amino acid biosynthesis (in particular lysine) bottlenecks (Atasoglu and Wallace, 2002). Since lysine is one of the most abundantly used amino acids in the gut fungal proteome,

it suggests that media supplementation strategies could be beneficial for protein production in the gut fungi.

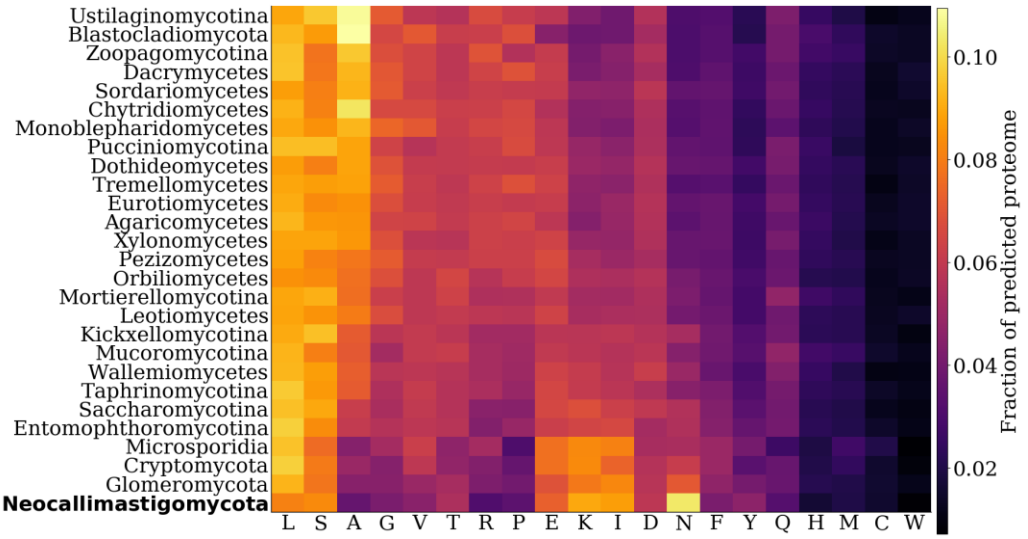

Figure 3.2: GC-depleted fungal proteomes are enriched in lysine, isoleucine and asparagine. Average predicted amino acid abundance per clade, ordered in decreasing GC content, is shown across the fungal kingdom. GC-rich fungal phyla are enriched in alanine, glycine, arginine, proline and valine. Asparagine is particularly enriched in Neocallimastigomycota, similar to *P. falciparum*, another extremely GC depleted organism. Figure taken from (Wilken *et al.*, 2020).

3.2.6 Anaerobic fungal CAZymes are enriched in homopolymeric amino acid runs

Homopolymeric runs of five or more consecutive identical amino acids are common in eukaryotic proteins (Albà, Tompa and Veitia, 2007). While their evolutionary origin is debated, it has been suggested that these low-complexity regions provide eukaryotes with a major source of phenotypic variation (Fondon III and Garner, 2004) and are associated with functionally important intrinsically disordered regions (Wright and Dyson, 2015). All fungal clades we investigated here have proteins with runs, where the average fraction of the proteome with runs ranges from 3% in Microsporidia (based on 8 genomes, 5 genera), to 30% in Neocallimastigomycota (based on 5 genomes, 4 genera), and finally to 37% in

Ustilaginomycotina (based on 16 genomes, 14 genera) for each clade. However, these runs are not evenly distributed across all the amino acids. Figure 3.3 shows that runs with leucine, valine, isoleucine, arginine, phenylalanine, tyrosine, methionine, cysteine and tryptophan are largely absent. The absence of proteins with bulky aromatic or hydrophobic amino acids implies that there is likely a cost associated with having long stretches of these residues. In the case of hydrophobic amino acids (valine, leucine, methionine and isoleucine) protein aggregation likely plays a role in preventing such runs from occurring. Smaller amino acids, like glycine, serine and alanine are more frequently found in runs, along with most of the polar amino acids. Cysteine is an exception to this, likely due to its reactive sulphur side chain.

Despite the likely complex evolutionary origin of these runs in proteins, analysis of all the CAZymes found in Neocallimastigomycota revealed that more than a quarter of all CAZymes contained a run motif. Interestingly, a large variation in the fraction of CAZymes with runs were found throughout the fungal kingdom, as shown in Figure 3.4. Neocallimastigomycota (28% with 5 genomes), Orbiliomycetes (27% with 2 genomes), and Monoblepharidomycetes (25% for a single genome) had the highest fractions of CAZymes with runs; all the other phyla had less than 20% on average. Given that only the simplest repetitive structure was searched for, this is likely an underestimate of the CAZymes that contain such low-complexity regions.

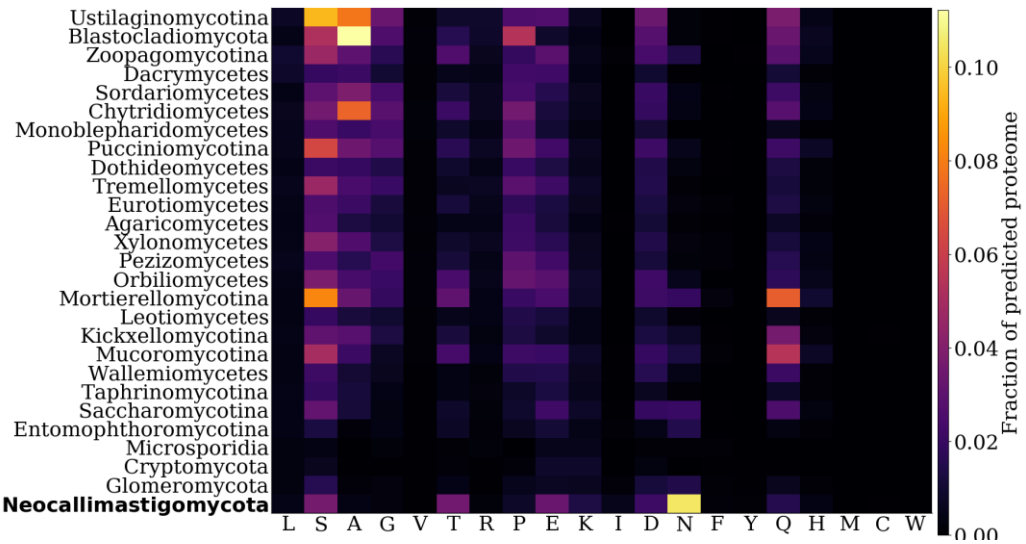

Figure 3.3: Proteins with asparagine runs constitute an unusually large fraction of the Neocallimastigomycota proteome. Average amino acid run (five or more of the same amino acid consecutively in a protein) fraction per clade, ordered in decreasing GC content, in the fungal kingdom. Hydrophobic (valine, leucine, methionine and isoleucine) and bulky (phenylalanine, tyrosine and tryptophan) amino acids are noticeably absent in runs, while smaller (alanine) uncharged, polar (serine, threonine, proline, glutamine) amino acids are frequently found in runs. Figure taken from (Wilken et al., 2020).

Furthermore, using transcriptomic data for *N. californiae* and *A. robustus* (Solomon *et al.*, 2016; Henske, Wilken, *et al.*, 2018) it was found that there are no significant differences in expression levels between CAZymes with and without run motifs (using the two sample Kolmogorov-Smirnov test). However, there is a significant difference (using the two-sample unequal variance t-test, $P<0.01$) in the ratio of CAZymes with runs versus total number of CAZymes between the fungi in Neocallimastigomycota and the genera *Trichoderma* and *Aspergillus*, which contain biotechnologically relevant organisms, (mean ratio of 0.28, 0.11, and 0.14 CAZymes with runs versus the total number CAZymes for each group respectively). Given the wide spread usage of *Trichoderma reesei* and *Aspergillus niger* in cellulase

production (Sukumaran *et al.*, 2009), it raises the question of the function of these runs, and if they impart some benefit to enzyme effectiveness.

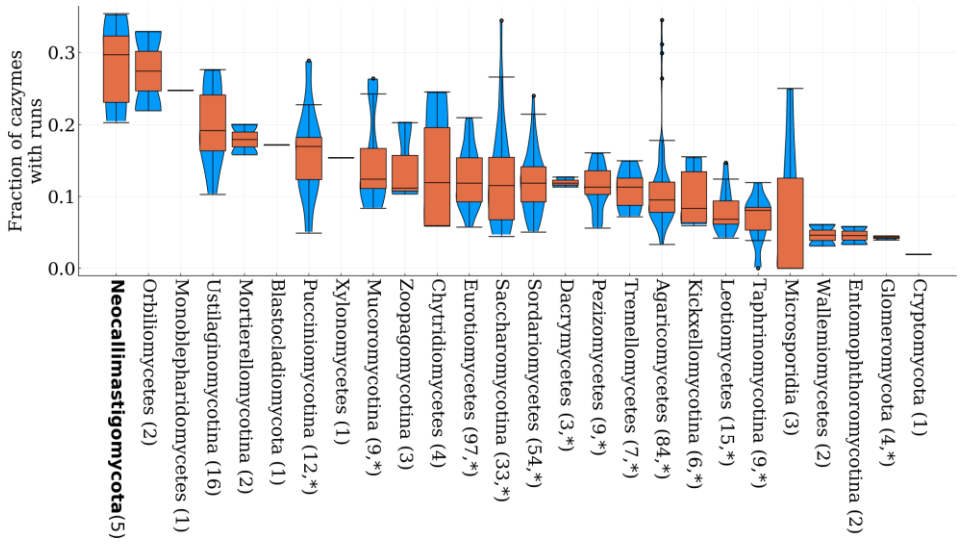

Figure 3.4: Neocallimastigomycota have significantly more CAZymes with amino acid repeat "runs" than other fungal clades. Distribution of the fraction of CAZymes with runs relative to all the CAZymes in each fungus, grouped by clade (the bracketed number is the number of fungi included in each clade). Statistically significant differences in the distributions between Neocallimastigomycota and all the other clades are indicated by * using the two sample Kolmogorov-Smirnoff test ($P<0.05$). The distribution of the fraction of CAZymes with runs in each clade is shown in the blue violin plots overlaid by orange box-and-whisker plots where outliers are shown as points, minima and maxima as whiskers, and the inter-quartile ranges inside the boxes. Figure taken from (Wilken et al., 2020).

3.2.7 Homopolymeric amino acid runs in CAZymes are enriched in threonine and serine, suggesting these enzymes are heavily glycosylated

While the organisms belonging to Neocallimastigomycota are still genetically intractable, heterologous expression of its CAZymes is likely the most expedient route to unlocking its biotechnological promise. However, expressing CAZymes heterologously is not

straightforward, in part due to glycosylation patterns that are difficult to mimic outside of the native host (Greene *et al.*, 2015). Moreover, recent work highlighting the role of processive enzymes attached to the cellulosome produced by members of Neocallimastigomycota showed that its CAZymes are heavily glycosylated (Haitjema *et al.*, 2017a). Indeed, genomic data indicate (see Table S5 in the supplementary material of the paper upon which this chapter is based) that the machinery for both N- and O-linked glycosylation is present in each sequenced genome of Neocallimastigomycota. Furthermore, by scanning the linker regions of all the CAZymes found in Neocallimastigomycota, *T. reesei*, and *A. niger*, as shown in Figure 3.5.A, for the canonical N-X-(S or T) (where X is any amino acid except proline) N-glycosylation motif, it becomes apparent that the motif is more abundant in the anaerobic gut fungi than in the latter two organisms. Only ~22% of the CAZymes found in the high-quality genomes of *N. californiae*, *A. robustus*, and *P. finnis* lack N-glycosylation motifs, compared to ~49% and ~35% for *T. reesei* and *A. niger*, respectively.

While N-linked glyosylation sites are straightforward to predict, no such recognition site has yet been identified for O-linked glycosylation. However, threonine and serine rich regions in the linker region of cellulase proteins are likely candidates for O-glycosylation (Beckham *et al.*, 2010; Sammond *et al.*, 2012). Figure 3.5.B shows the amino acid abundance in CAZymes split into domains and the inter-domain (linker) regions, which are further separated into linker regions of proteins with and without runs. It is clear that asparagine, serine, and especially threonine, are significantly enriched in the linker regions of Neocallimastigomycota. The threonine enrichment is even more pronounced in the linker regions of proteins that have runs, reflecting the disproportionate abundance of threonine runs in CAZymes.

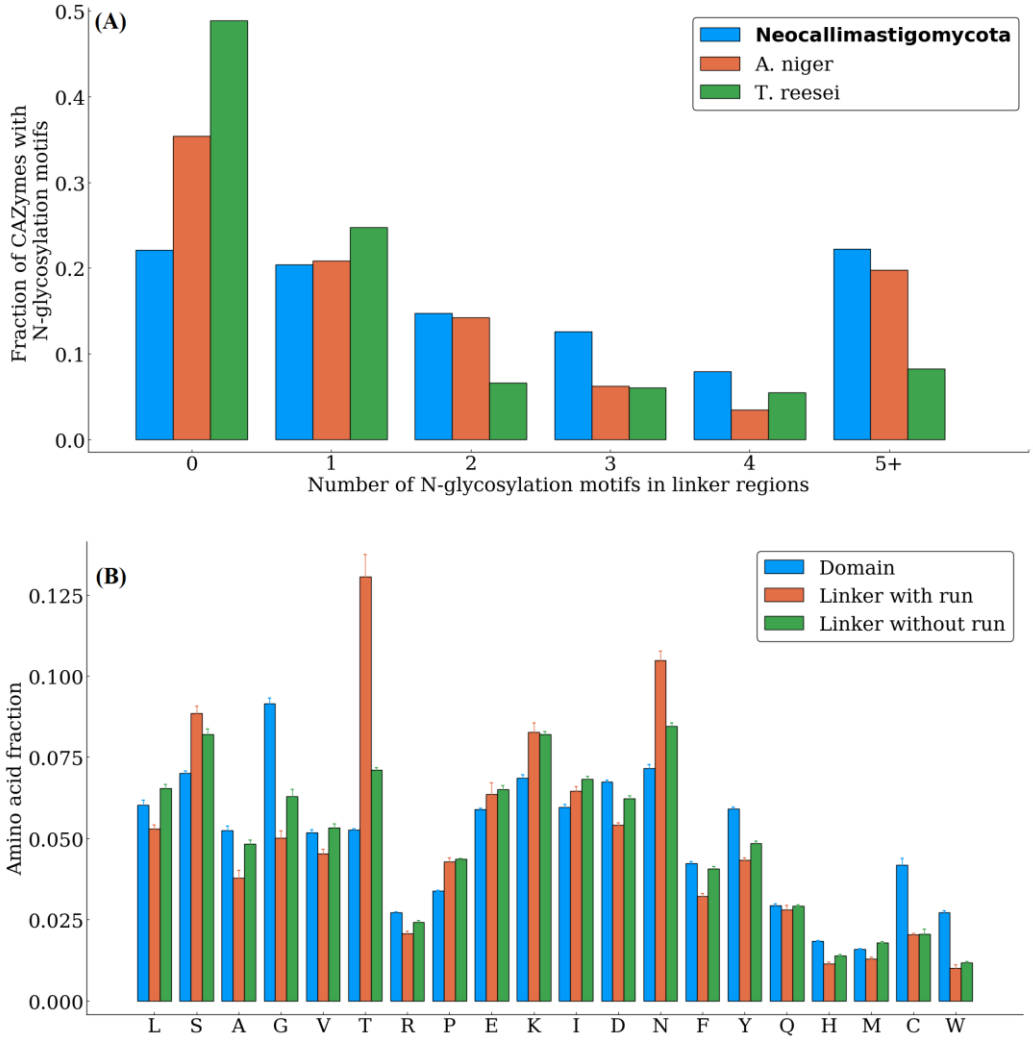

Figure 3.5: (A) CAZymes in Neocallimastigomycota have more N-glycosylation motifs relative to other industrially important cellulolytic fungi. The fraction of CAZymes with a specified number of N-glycosylation motifs (N-X-S/T where X is not proline) on the x-axis in *N. californiae*, *A. robustus* and *P. finnis* (grouped as Neocallimastigomycota here, the other members are not shown due to their lower quality genomes), *T. reesei*, and *A. niger*. Linker regions are defined as the inter-domain regions of proteins. Neocallimastigomycota has a higher proportion of CAZymes with 2 or more N-glycosyaltion motifs than either *T. reesei* or *A. niger*. (B) Threonine is disproportionately abundant in the linker region of CAZymes in Neocallimastigomycota, suggesting O-glycosylation sites may be abundant. Amino acid fraction in all proteins with at least one CAZyme (carbohydrate active enzyme)

domain divided into three groups: domains, linker regions of proteins with runs (five or more of the same amino acid consecutively in a protein), and linker regions of proteins without runs. Linker regions are defined as the inter-domain regions. Serines, and especially threonines, are highly enriched in the inter-domain regions of CAZymes with runs and without runs. Figure taken from (Wilken et al., 2020).

Given that glycosylation is a mechanism used by cells to protect CAZymes from proteolytic cleavage, and that the rumen of herbivores is heavily populated with proteases (Bach, Calsamiglia and Stern, 2005), is reasonable to hypothesize that these regions are indeed glycosylated *in vivo*. Prior work has shown that CAZymes within the cellulosomes of Neocallimastigomycota are indeed heavily glycosylated (Haitjema *et al.*, 2017a). Additionally, marked increases in CAZyme activity have been observed when the expression host is changed to an organism capable of glycosylating its enzymes (Ximenes *et al.*, 2005; Cheng *et al.*, 2014, 2015). Finally, the importance of linker regions in cellulase function (Sonan *et al.*, 2007) reinforces the idea that metabolic engineering strategies should take these features into account to optimally leverage the CAZyme machinery of Neocallimastigomycota.

3.3 Conclusions

While the underlying reasons for GC depletion in Neocallimastigomycota remain unclear, the consequences of this AT-richness for metabolic engineering are numerous. The possibility that the anaerobic gut fungi have high mutational rates due to their GC depletion has interesting implications for strain evolution, engineering, and stability. Understanding how, and at what rate, their genomes evolve will provide an improved roadmap to engineer these organisms (Sekowska *et al.*, 2016; Nørholm, 2019). While functional genetic tools to modify the anaerobic fungi are in development, the codon optimization strategy presented here may

attenuate the current difficulties associated with expressing non-native genes in these hosts. The GC depleted genomes likely also limit the use of G-rich PAM targeting Cas enzymes in the CRISPR system, suggesting that Cas enzymes engineered to target T-rich PAM sites should be prioritized for engineering anaerobic fungi. Comparative genomic analyses have shown that homopolymeric runs of amino acids are unusually common in anaerobic fungi, especially in their CAZyme machinery. These motifs likely serve important functions, e.g. glycosylation sites that prevent proteolytic cleavage, suggesting the importance of understanding their role if gut fungal CAZymes are heterologously produced in a model organism.

3.4 Acknowledgements

The authors acknowledge funding support from the National Science Foundation (NSF) (MCB-1553721), the Office of Science (BER) the US Department of Energy (DOE) (DE-SC0010352), the Institute for Collaborative Biotechnologies through grants W911NF-19–D-0001 and W911NF-19-2-0026 from the US Army Research Office, and the Camille Dreyfus Teacher-Scholar Awards Program. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was part of the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the Office of Biological and Environmental Research of the DOE Office of Science through contract DE-AC02–05CH11231 between Lawrence Berkeley National Laboratory and the DOE.

3.5 Materials and methods

3.5.1 Collection and processing of genomic data

The Mycocosm database, curated by the Joint Genome Institute (JGI), was used to download 443 sequenced fungal genomes (listed in Table S1 in the supplementary material of the paper upon which this chapter is based, (Grigoriev *et al.*, 2014)), as well as their predicted protein coding genes, predicted proteomes and associated PFAM annotations. The whole genomes (protein coding and non-coding regions) were also downloaded for all the sequenced fungi in Neocallimastigomycota, as well as *Saccharomyces cerevisiae*, *Trichoderma reesei* and *Rhodotorula graminis* (accessed November 2018) (Grigoriev *et al.*, 2014). The *de novo* assembled transcriptomes of *A. robustus* and *N. californiae* (Solomon *et al.*, 2016) and the associated differential transcriptomic datasets for each these fungi grown in isolation on reed canary grass (Henske, Wilken, *et al.*, 2018) were also used as described below. Scripts using the Julia programming language, and the associated BioSequences and HypothesisTests packages, were used to process and analyze the data (Bezanson *et al.*, 2017). Code is available on Mendeley Data, DOI: 10.17632/26vywtfkrz.1

3.5.2 Nucleotide content, genome analysis and construction of codon tables

Using genomic data, the nucleotide content of the protein coding genes for all 443 fungi, as well as the unmasked whole fungal genomes for the fungi in Neocallimastigomycota, were calculated by counting each nucleotide base (G, C, T, A) and ignoring gaps and indeterminate bases (N). The GC fraction was then calculated by dividing the total number of G and C bases by the total number of G, C, T, and A bases ((G+C)/(G+C+A+T)). Similarly, the amino acid abundances were calculated by counting the number of each amino acid found in the predicted

proteome, which is the translated protein coding gene, relative to the total number of amino acids in the same predicted proteome for each organism. Phylogenetic classifications were based on the taxonomic assignments as defined by the JGI. The number of protospacer adjacent motif (PAM) sites per genome was counted by parsing through each scaffold on the whole genomes of all the sequenced fungi in Neocallimastigomycota, as well as *Saccharomyces cerevisiae*, *Trichoderma reesei* and *Rhodotorula graminis,* and counting the number of matches to a particular motif, e.g. TTN would match to TTA, TTG, TTT and TTC, using the Julia BioSequences package. The number of hits was then divided by the total number of bases in each fungal genome.

Codon optimization tables for anaerobic fungi were calculated by first identifying the top 1000 most expressed genes in the fungal isolates *N. californiae* and *A. robustus* using the differential transcriptomic data for these fungi grown on reed canary grass from Henske, et al. (2018) (Henske, Wilken, *et al.*, 2018). BLASTn was then used to align these transcripts to their predicted genes (Camacho *et al.*, 2009). Only genes with coverage greater than 90% and e-values less than 10^{-60} were decomposed into codons. The frequency of each codon was then calculated by counting the number of times it appears relative to all the other synonymous codons. The tRNA gene counts in the genomes of *A. robustus* and *N. californiae* were found by tRNAscan-SE using the eukaryotic specific parameters (Chan and Lowe, 2019).

3.5.3 Identification of homopolymeric amino acid runs and glycosylation motifs in fungi

Using the Julia BioSequences package, the predicted proteomes from downloaded Mycocosm fungal genomes were searched for homopolymeric runs of 5 or more consecutive amino acids of the same type (Karlin *et al.*, 2002) through the regular expression "X{5,}"

where X is the amino acid query. This bioinformatic search returns the longest uninterrupted hit of 5 of more amino acids (X) in succession within the proteome. For example, the hypothetical protein, "MGK**TTTTT**L**TTTTTTT**F", has two threonine runs of length five and six. The canonical N-glycosylation motif, N-X-(S or T) (where N is asparagine, X is any amino acid except proline, S is serine and T is threonine) (Deshpande *et al.*, 2008), was found by searching each protein using the regular expression "N[^P](S|T)".

3.5.4 CAZyme identification and transcriptomic expression analysis

CAZymes were identified by matching the predicted protein family annotations from the PFAM annotation files in the 443 sequenced fungal genomic datasets to a list of CAZyme family domains. See Table S6 in the supplementary material upon which this chapter is based for the PFAM to CAZyme family domain association table (Carlson *et al.*, 2019). A protein was designated as a CAZyme if at least one annotated PFAM domain was found in the CAZyme family domain list. For each fungus, this filtered list of CAZymes was used to search for amino acid runs as described above, to determine the amino acid composition of the CAZymes and to find predicted N-glycosylation motifs. Furthermore, only predicted CAZymes that had a coverage greater than 90% and an e-value less than 10^{-40} (using BLASTn (Camacho *et al.*, 2009) to match the associated gene against the transcriptomes in (Solomon *et al.*, 2016)) were included in the CAZyme expression analysis using the reed canary grass condition data of (Henske, Wilken, *et al.*, 2018).

3.5.5 Annotation of sexual reproduction and glycosylation genomic machinery in fungal genomes

To evaluate the potential for sexual reproduction by phylum Neocallimastigomycota, we searched member genomes for a subset of proteins required by other fungi for sexual reproduction (Hull, Raisner and Johnson, 2000). Using *S. cerevisiae* peptide sequences obtained from "The Saccharomyces Genome Database (SGD)" (www.yeastgenome.org) as queries we searched for mating factors, proteins involved in sexual reproduction, and proteins involved in meiosis. Specific *Saccharomyces* genes queried included sex-implicated kinases (STE20) (Leberer *et al.*, 1996), sex-signal transduction proteins (STE6, GPA1) (Sadhu *et al.*, 1992; Raymond *et al.*, 1998), meiosis specific recombinases (DMC1) (Diener and Fink, 1996), and mating factors (MATa/MATα) (Hull, Raisner and Johnson, 2000). Additionally, peptide mating factors of *N. crassa* (MATA/MATa), and pheromone receptor domain containing proteins from cryptically sexual fungi were queried against anaerobic fungal genomes (Glass, Grotelueschen and Metzenberg, 1990; Staben and Yanofsky, 1990; Ropars *et al.*, 2016). The tBLASTn algorithm with a BLOSUM62 substitution matrix was used to score peptide alignments against genomes using an expected e-value of 10^{-25} (Camacho *et al.*, 2009).

The glycosylation machinery in fungi is highly conserved, as such *S. cerevisiae*'s canonical genes were used as benchmarks for the identification of putative glycosylation pathways (Deshpande *et al.*, 2008). The predicted proteins of the gut fungi were compared to benchmark proteins found in *S. cerevisiae* (downloaded from Uniprot ('UniProt: a worldwide hub of protein knowledge', 2019)) using BLASTp (Camacho *et al.*, 2009). A gene was deemed present if the coverage was greater than 50% and the e-value less than 10^{-20}. O-

glycosylation was deemed possible if all PMT1-4 genes were found (Gentzsch and Tanner, 1996) and N-glycosylation was deemed possible if at least three of DPM1, ALG3, ALG9, ALG12, OST1, OST3 and STT3 were found (Knauer and Lehle, 1999; Deshpande *et al.*, 2008), see Table S5 in the supplementary material of the paper upon which this chapter is based for a collation of the blast results.

IV. Experimentally validated reconstruction and analysis of a genome-scale metabolic model of an anaerobic Neocallimastigomycota fungus

This chapter is based upon work that is in preparation for publication in mSystems Journal by St. Elmo Wilken, Jonathan M. Monk, Patrick A. Leggieri, Christopher Lawson, Thomas S. Lankiewicz, Susanna Seppälä, Stephen J. Mondo, Kerrie W. Barry, Igor V. Grigoriev, John K. Henske, Michael K. Theodorou, Bernhard O. Palsson, Linda R. Petzold, and Michelle A. O'Malley, entitled "*Experimentally validated reconstruction and analysis of a genome-scale metabolic model of an anaerobic Neocallimastigomycota fungus*". See the upcoming publication for more detailed information regarding the datasets used to construct the model, the supplement, as well as the model itself.

4.1 Introduction

Anaerobic gut fungi, in the early-branching phylum Neocallimastigomycota, are found in the digestive tracts of herbivores where they play an integral role in the lignocellulolytic microbiome of their host (Gruninger *et al.*, 2014). Recent transcriptomic and genomic analyses have revealed that these fungi harbor an incredible diversity of carbohydrate active enzymes (CAZymes) that are tailored to excel at decomposing lignocellulosic plant biomass (Resch *et al.*, 2013; Youssef *et al.*, 2013; Solomon *et al.*, 2016; Haitjema *et al.*, 2017b; S Seppälä *et al.*, 2017). Given their ability to metabolize raw lignocellulose, anaerobic gut fungi are an appealing biotechnological platform to drive the conversion of lignocellulose into hydrolyzed sugars, and ultimately into renewable chemicals via biofermentation (Himmel *et al.*, 2007; Sarkar *et al.*, 2012; Liao *et al.*, 2016).

While the lignocellulolytic abilities of anaerobic gut fungi motivate their biotechnological interest, they are temperature sensitive, obligately anaerobic, relatively slow growing, typically require specialized media, and their genomes are extremely AT and repeat rich (Youssef *et al.*, 2013; Haitjema *et al.*, 2014; Wilken *et al.*, 2020). Furthermore, no robust genetic engineering tools have been developed for this class of fungi, hampering classic molecular biology techniques that can be used to investigate, understand and engineer their metabolism. Despite these challenges, experimental and 'omic' datasets have emerged to elucidate some aspects of their metabolism (Marvin-Sikkema *et al.*, 1994a; Akhmanova *et al.*, 1999; Boxma *et al.*, 2004; Youssef *et al.*, 2013; Henske, Gilmore, *et al.*, 2018). Still lacking, however, is a framework to synthesize this data, to clarify lingering uncertainties regarding their unique physiology, and to provide a systematic way to engineer anaerobic fungi for biotechnology. In particular, their hydrogenosomal metabolism is unresolved, with no clear consensus on the pathways used in this mitochondrion like organelle. Current hypotheses either suggest an energetically unfavorable pathway involving pyruvate formate lyase is used to produce H_2, or a pathway involving pyruvate ferredoxin oxidoreductase that is not supported by extra-cellular metabolite measurements (Marvin-Sikkema *et al.*, 1994a; Boxma *et al.*, 2004).

Genome-scale models (GSMs) can be used to address these shortcomings, as they are well suited to act as knowledge base platforms for integrating multi-omic datasets and have been successfully used to drive the engineering of both pro- and eukaryotes (Blazeck and Alper, 2010; Aung, Henry and Walker, 2013; Simeonidis and Price, 2015). Moreover, by experimentally testing the predictions of a GSM we can systematically refine our

understanding of the metabolism of an organism. This is particularly appealing in the context of non-model microbes, like the anaerobic gut fungi, that are relatively understudied.

Here, we introduce a high-quality, PacBio sequenced genome (200 Mbps, 62x sequencing depth) of the anaerobic gut fungus *Neocallimastix lanati*. Comparative genomic analyses revealed that *N. lanati* is metabolically similar to the other sequenced isolates, suggesting that insights gained from understanding its genome may be generalizable to the other species in the clade. Moreover, the genome of *N. lanati* encodes for many CAZymes (~1788 CAZymes with 585 associated with the fungal cellulosome), as found in other sequenced Neocallimastigomycota fungi, reinforcing its biotechnological promise. Based on the genome, we constructed the first genome-scale metabolic model of an anaerobic gut fungus. This fungus is well suited to act as a platform to investigate the metabolism of anaerobic gut fungi because it can grow in completely defined (M2) media, is relatively fast growing among gut fungal strains ($\mu = 0.045 \pm 0.003$ 1/h in M2 media), and can be cryopreserved. The 3-compartment (extracellular, cytosolic and hydrogenosomal compartments) model introduced here, named iSW587, is composed of 587 genes, 1014 reactions, 815 metabolites and models the primary metabolism of *N. lanati*. The model is stoichiometrically consistent, as well as mass and charge balanced. Experimental, genomic, transcriptomic and metabolic flux analysis data were used to build and validate the model. The model recapitulates extracellular metabolite production rates and accurately models the observed growth rate. Furthermore, the model refines and expands on previous hypotheses regarding the metabolism of the gut fungal hydrogenosome. In particular, both the model and experimental data suggest that pyruvate formate lyase (PFL) is significantly more active than pyruvate ferredoxin oxidoreductase (PFO) in the hydrogenosome, but that hydrogen formation can only occur via the latter

pathway. Going forward, this fungus and its associated model can be used to guide efforts to elucidate aspects of gut fungal metabolism that remain unclear and direct future metabolic engineering strategies. Indeed, model based analysis could be invaluable in designing stable consortia between anaerobic gut fungi and other industrially utilized organisms – something that has not yet been fully realized (Ranganathan *et al.*, 2017; Henske, Wilken, *et al.*, 2018; Gilmore *et al.*, 2019).

4.2 Results and discussion

4.2.1 The genome of *N. lanati* is rich in carbohydrate active enzymes (CAZymes) and metabolically similar to other anaerobic gut fungi

Given the large genomic size and repeat-richness inherent to anaerobic fungi (Wilken *et al.*, 2020), PacBio sequencing was used to obtain a high-quality genome of the isolate *N. lanati* (see Figure S1 in the supplement for its phylogeny; the Index Fungorum identification number is IF557810), which was sourced from a fecal pellet of a sheep. While the genome of this fungus is large, as shown in Table 4.1, it is the second least fragmented of all 5 of the published gut fungal genomes, as shown in Table S1 in the supplement. The *N. lanati* genome encodes for a rich array of carbohydrate active enzymes (CAZymes) in similar numbers to those reported from other gut fungal genomes (Youssef *et al.*, 2013; Solomon *et al.*, 2016; Haitjema *et al.*, 2017b). In total, 1788 CAZymes were identified in the genome, of which 1253 were expressed in the transcriptome (see the transcriptome supplied in the supplemental dataset). Like other anaerobic gut fungi, *N. lanati* deploys both complexed (cellulosomes) and uncomplexed CAZymes through its rhizoidal network (see Table 4.1 and Figure 4.1), which mechanically disrupts lignocellulose, to aid in its decomposition. Figure S2 in the supplement

shows the breakdown of CAZyme domains identified in the genome of *N. lanati*. This, in combination with its relatively high growth rate on defined M2 media, suggests that *N. lanati* is a good model anaerobic gut fungus. The genome is available online at https://mycocosm.jgi.doe.gov/Neolan1/Neolan1.info.html.

Table 4.1: A summary of the features of the genome of *N. lanati*. CAZyme = carbohydrate active enzyme, GH = glycoside hydrolase. Metabolic genes are defined as genes that have an enzyme commission (EC) number assigned to them. GH genes that have a dockerin domain are likely present in cellulosomal complexes (Haitjema *et al.*, 2017b).

Feature	Value
Genome size (Mbp)	200.97
Number of scaffolds	970
Sequencing read coverage depth	62.05x
Number of predicted genes	27677
Number of CAZymes	1788
Number of GH genes	678
Number of GH genes containing a dockerin domain	271
Number of metabolic genes	2761
Number of transporters	1754

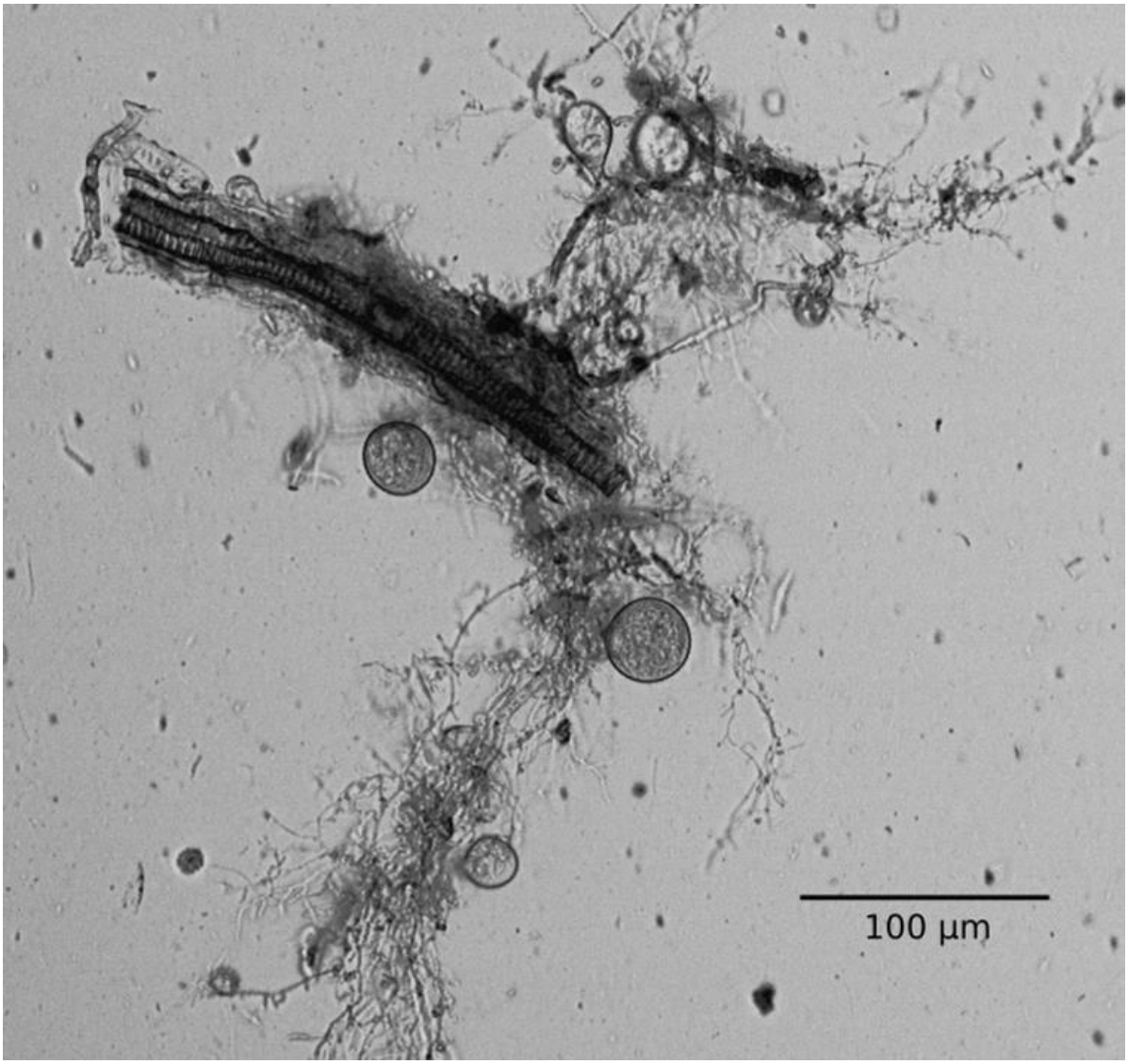

Figure 4.1: The morphology of *Neocallimastix lanati* aids in the decomposition of unpretreated lignocellulose by disrupting the lignocellulosic plant biomass to increase the surface area available for enzymatic attack. A micrograph of a mature *N. lanati* sporangium growing on corn stover in M2 media after 3 days of growth at 39°C. The filamentous rhizoidal network is used to increase the surface area for its lignocellulolytic enzymes that decompose the lignocellulosic corn stover into its fermentable sugar constituents.

Despite advances in sequencing and annotation, a large number of putative gut fungal genes remain unannotated (~48% of the 27,677 predicted genes of *N. lanati*) (as shown in

Table S1 in the supplement), which is consistent with previous genomic annotations in this clade. These unannotated genes contribute to gaps found in the reconstructed metabolism of *N. lanati*. A comparative genomic analysis within the primary metabolism across all high quality publicly available gut fungal genomes (*Anaeromyces robustus, Neocallimastix californiae, Pecaromyces ruminantium, Piromyces finnis*) revealed that the gut fungi are metabolically similar. Of the 1023 unique EC numbers identified across these 5 genomes, less than ~3% are unique to each isolate (as shown in Figure 4.2). This suggests that gut fungi share a similar primary metabolism. Thus, metabolic gaps can potentially be filled by searching for genes in *N. lanati* that are homologous to those encoded for in the genomes of the other gut fungal isolates. In this way, key enzymes in the biosynthesis pathways of arginine, asparagine, biotin, riboflavin, lipids and fatty acids were identified and included in the metabolic reconstruction of *N. lanati*. In total, 53 gaps in the primary metabolic pathways were identified and annotated in this manner, as noted in the confidence score and homologous gene annotation fields in the model.

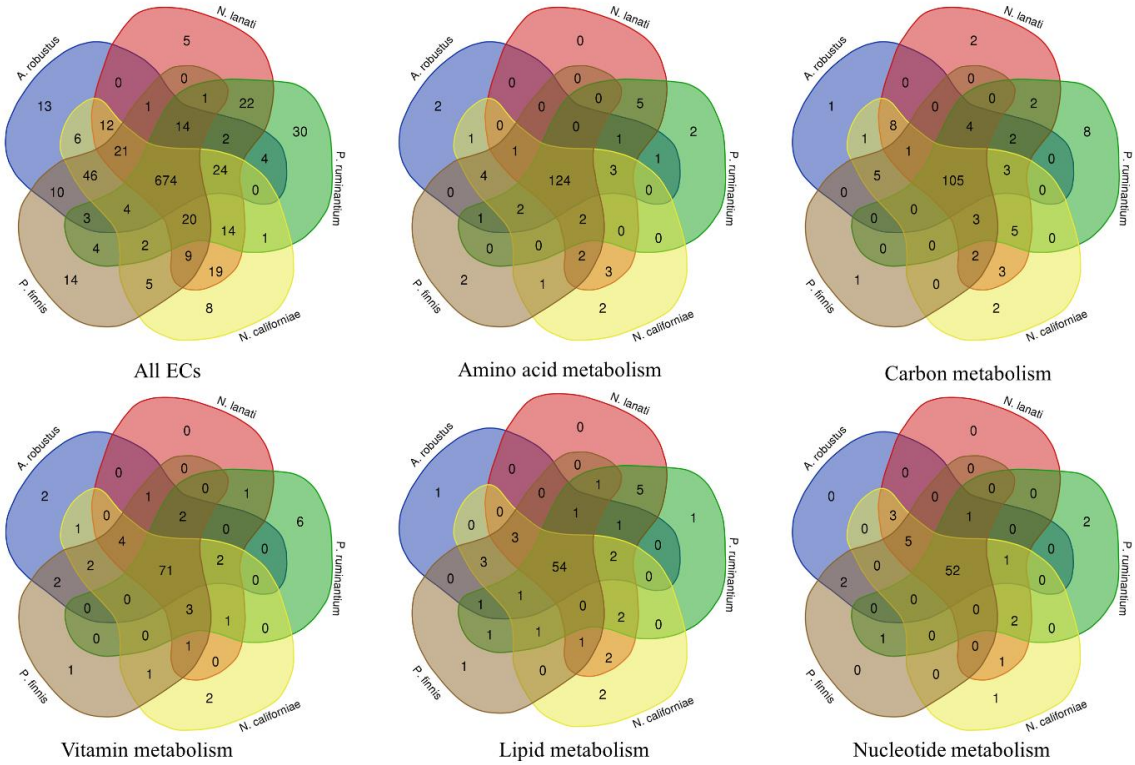

Figure 4.2: Anaerobic gut fungi have very similar metabolic potential, suggesting that metabolic gaps can be filled by looking for homologous genes found in the other sequenced isolates. Each Venn diagram was generated by inspecting the intersection of the annotated EC numbers contained in the genome of each fungus for each metabolic module. Overlapping regions imply that those isolates share the EC assignments contained in each of the metabolic modules. The EC numbers contained in each module are based the KEGG database (Kanehisa *et al.*, 2016) (see the supplement for the list of modules encompassing each Venn diagram), while the EC assignments for each fungus are based on the JGI and bidirectional annotation data as described in the methods section.

4.2.2 The curated metabolic model of *N. lanati* captures the carbon, amino acid, vitamin, fatty acid, nucleotide and lipid metabolism

Based on the metabolic reconstruction of *N. lanati*, a manually curated genome-scale model of *N. lanati* was built (iSW587) by following an established protocol for generating high-quality reconstructions (Thiele and Palsson, 2010). The model contains 1014 reactions, 815 metabolites, and 587 genes distributed across 3 compartments (hydrogenosome,

cytoplasm, and extracellular space). Where possible, experimental data was used to curate the model. The methods section details specifics on the curation process, as well as experiments used to construct the biomass objective function of the model. Briefly, Table 4.2 shows the experimentally measured macromolecular components of *N. lanati* that were used to construct the biomass objective function for the genome-scale model. Further simplifying assumptions were made to construct the specific biomass objective function used in iSW587. The carbohydrate component of the biomass was assumed to be solely chitin and the amino acid composition of the protein component of the biomass was assumed to follow the amino acid distribution of the predicted genes (i.e. the predicted proteome). Similarly, the nucleotide composition was assumed to follow the composition of the genome (for the DNA nucleotides) and the transcriptome (for the RNA nucleotides). The lipid component was assumed to be composed of myristic, palmitic and stearic acid, which were found to be the major fatty acid components of the lipid fraction of *N. lanati*, as shown in Figure S3 in the supplement. The growth associated and non-growth associated maintenance (GAM and NGAM, respectively) functions were estimated using experimental data, see Figure S4 in the supplement and Table 4.2.

Table 4.2: The experimentally measured macromolecular constituents of *N. lanati* that were used to construct the biomass objective function for the genome-scale model. Experimental data were used to estimate the biomass objective function. See the methods section for more details.

Biomass component	Mass fraction [g/g_{dw} %]
Carbohydrate	32.4 ± 1.6
Protein	43.7 ± 1.2
Lipids	4.9 ± 0.2
DNA	0.2 ± 0.1

RNA	0.6 ± 0.1
Sum	81.8 ± 3.2
GAM	76 mmol ATP/g_{DW}/h
NGAM	2.3 mmol ATP/g_{DW}

Table 4.3 provides a brief summary of the main features of the model, while the full model is included in the supplement. Table S2 in the supplement explains the confidence rating assigned to each reaction in the model. In the energy generating pathways, particular attention was paid to modeling the hydrogenosome (a mitochondrion like organelle that functions completely anaerobically), which is discussed in greater detail in the following sections. More generally, the Embden–Meyerhof–Parnas variant of glycolysis is present in *N. lanati*, as well as pathways for mixed acid fermentation (succinate, acetate, lactate, formate and ethanol), which are typically found in anaerobic gut bacteria (Flint *et al.*, 2008). Interestingly, it was found that *N. lanati* possesses both the NAD^+ and $NADP^+$ variants of glyceraldehyde-3-phosphate dehydrogenase in glycolysis, with the latter used to conserve energy as NADPH instead of ATP. The pentose phosphate pathway of *N. lanati* is incomplete, with glucose-6-phosphate dehydrogenase and 6-phosphogluconate missing. These reactions regenerate NADPH and possibly explain the presence of the $NADP^+$ variant of glyceraldehyde-3-phosphate dehydrogenase as a compensating mechanism (Martínez *et al.*, 2008). The xylose isomerase pathway is also present in *N. lanati*, as has been found in other sequenced gut fungi (Henske, Wilken, *et al.*, 2018).

Table 4.3: Summary of iSW587 features. The model is also consistent, as well as completely mass and charge balanced.

Features	Number

Total reactions	1014
Total metabolites	815
Total genes	587
Number of compartments	3 (Extracellular, cytosolic, hydrogenosomal)
Carbohydrate reactions	138
Nucleotide reactions	133
Transporters	162
Amino acid reactions	145
Vitamin reactions	133
Lipid reactions	162

The major components (amino acids, nucleotides, vitamins, fatty acids and lipids) of the anabolic metabolism of *N. lanati* were found to be present, in agreement with its ability to grow in sparse M2 media. Specifically, the complete biosynthesis pathways for all the proteogenic amino acids and the modeled fatty acids were found. Most of the canonical vitamin and co-factor (vitamin B5, vitamin B6, riboflavin and thiamin) biosynthesis pathways were also found to be complete, with the exception of folate where no synthesis mechanism of 4-aminobenzoate was found. However, the heme and biotin biosynthesis pathways were found to be incomplete. Since *N. lanati* can grow in completely defined M2 media, gaps in the model due to nutritional requirements were relatively easy to fill. Finally, the model recapitulates the experimentally observed growth rate in defined media using only the measured flux of glucose (1.5 mmol/gDW/h) as an input constraint (flux balance analysis predicted $\mu = 0.047$ 1/h vs. an experimentally measured $\mu = 0.045 \pm 0.003$ 1/h).

4.2.3 iSW587 includes an expanded model of the hydrogenosomal metabolism

Anaerobic gut fungi possess a variant of the hydrogenosome, with the core set of enzymes that catalyze the conversion of malate and pyruvate to acetate, H_2 and formate already identified, as shown in Figure 4.3 (Yarlett *et al.*, 1986; Marvin-Sikkema *et al.*, 1994a; Boxma *et al.*, 2004; Hackstein *et al.*, 2019). However, the metabolic pathway leading to H_2 production is not resolved with literature suggesting either pyruvate ferredoxin oxidoreductase (PFO) or pyruvate formate lyae (PFL) are possible routes. Both enzymes were identified in the genome and transcriptome and are thus included in the model of the hydrogenosome.

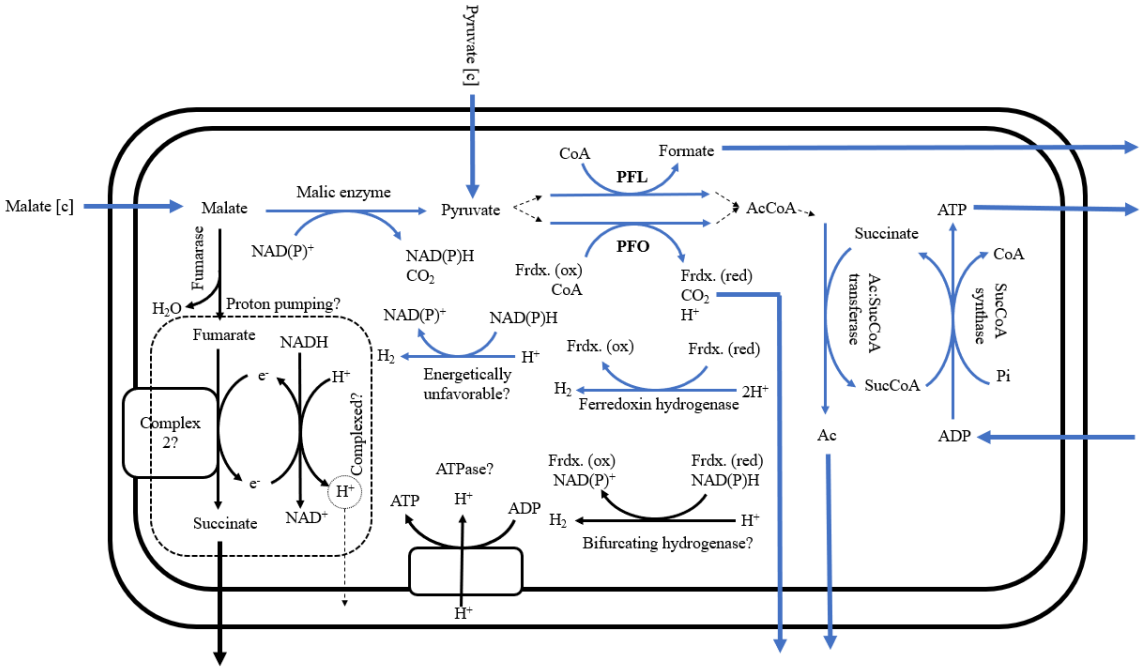

Figure 4.3: An expanded model of the hydrogenosome is included in the model based on genomic annotation, literature and predicted localization data (Marvin-Sikkema *et al.*, 1994a; Akhmanova *et al.*, 1999; Boxma *et al.*, 2004). Core hydrogenosome enzymes are colored in blue, while speculative enzymes are shown in black. PFL = pyruvate formate lyase, PFO = pyruvate ferredoxin oxidoreductase, Ac = Acetate, SucCoA = Succinyl Coenzyme A, CoA = Coenzyme A, AcCoA = Acetyl Coenzyme A, Frdx. = ferredoxin.

By combining literature sources, gene annotation, transcriptomic expression and subcellular localization data, we have included additional pathways in the model of the

hydrogenosome for *N. lanati*, as shown in Table 4.4 and Figure 4.3. Of note is the inclusion of an ATP synthase, which has previously been speculated to be present in other anaerobic gut fungal isolates (Marvin-Sikkema *et al.*, 1994b; Youssef *et al.*, 2013). Additionally, we also found evidence that complex 2 of the mitochondrial electron transport chain is present: homologs to all four subunits were found to be expressed and localized to the hydrogenosome, see Table 4.4 (complex 2: sub. A, B, C, D). We could not find any homologs of the membrane bound subunits of complex 1 or the ATP synthase in the *N. lanati* genome, as has also been reported previously for other anaerobic gut fungi (Seppälä *et al.*, 2016). It is perplexing that no homologs of the membrane bound subunits of complex 1 were found, since these are used to shuttle electrons between the two complexes in the inner membrane of the mitochondria, and complex 2 cannot function without them. However, homologs of the soluble subunits of complex 1, nuoF and nuoE, are highly expressed relative to the other core enzymes of the hydrogenosome, see Table 4.4. The presence of the soluble subunits, coupled with the absence of the membrane associated subunits of complex 1, has also been observed in the hydrogenosomes of, amongst others, *Trichomonas vaginalis* (Hrdy *et al.*, 2004; Schneider *et al.*, 2011) and *N. ovalis* (Boxma *et al.*, 2007). This raises two possibilities. First, that *N. lanati* possesses a proton pumping mechanism. While the membrane bound subunits of complex 1 would be critical for this function to work, we do find preliminary evidence of a pH gradient inside the hydrogenosome, as shown in Figure S4 in the supplement. Second, it is possible that its hydrogenase associates with the identified nuoF-like subunit of complex 1 to form a bifurcating hydrogenase, as has been speculated to occur in *T. vaginalis* (Hrdy *et al.*, 2004; Muller *et al.*, 2012). Indeed, we find high homology sequences in the *N. lanati* genome to all

three of the bifurcating hydrogenase subunits characterized in *Thermotoga maritima* (Gerrit J Schut and Adams, 2009), as shown in Table S3 in the supplement.

Table 4.4: Enzymes included in the model of the hydrogenosome metabolism. Mitochondrial localization is probably hydrogenosomal due to their evolutionary relationship (Youssef *et al.*, 2013). Transcriptomic expression count data is derived from the M2 cellobiose expression dataset and represent the mean of a triplicate for each enzyme. Localization was predicted using DeepLoc (Almagro Armenteros *et al.*, 2017). PFL = pyruvate formate lyase, PFO = pyruvate ferredoxin oxidoreductase, Ac = Acetate, SucCoA = Succinyl Coenzyme A, syn = synthase, trans = transferase, sub = subunit.

Enzyme	Gene (Protein ID)	Mean expression [TPM]	Localization	Number of other gut fungi where this gene was found
PFL 1	981064	1967	Cytoplasm	5
PFL 2	1027775	182	Cytoplasm	5
PFO	623223	17	Mitochondrion	4*
Ac:SucCoA trans.	1731457	217	Cytoplasm	5
Ac:SucCoA trans.	1316948	217	Cytoplasm	5
SucCoA syn. sub. A	1636158	1048	Mitochondrion	5
SucCoA syn. sub. B	1276456	1544	Mitochondrion	5
Hydrogenase 1	1341048	219	Mitochondrion	5
Hydrogenase 2	1718044	17	Cytoplasm	5
Complex 1: nuoF	1047445	339	Mitochondrion	5
Complex 1: nuoE	993995	519	Mitochondrion	5
Complex 2: sub. A	1702000	4	Mitochondrion	5
Complex 2: sub. B	1688149	13	Mitochondrion	5
Complex 2: sub. C	1286787	12	Mitochondrion	3
Complex 2: sub. D	1677752	8	Mitochondrion	2
Fumarase	985684	4	Cytoplasm	5
ATP syn.: sub. Alpha	1037070	1	Mitochondrion	5
ATP syn.: sub. Beta	1706307	8	Mitochondrion	5
ATP syn.: sub. Delta	1045818	26	Mitochondrion	5

ATP syn.: sub. Gamma	1061751	3	Mitochondrion	5

*Not identified in the genome of *N. californiae*, however a transcript with close homology to PFO was identified.

Taken together, our expanded model of the hydrogenosome includes the core enzymes previously reported in other fungal species as well as a speculative bifurcating hydrogenase, an ATP synthase and a proton pumping module composed of complex 1 and complex 2 enzymes identified in the *N. lanati* genome (similar to what has found in other H_2 producing mitochondria (Muller *et al.*, 2012)). Given the speculative nature of the proton pumping mechanism and the bifurcating hydrogenase, these reactions are constrained to carry zero flux in the working model. Additionally, it has previously been suggested that a hydrogen dehydrogenase ($NAD(P)^+ + H_2 \leftrightarrow H^+ + NAD(P)H$) operates in the reverse direction in the hydrogenosome (Akhmanova *et al.*, 1999; Boxma *et al.*, 2004; Youssef *et al.*, 2013). Consequently, this hydrogen dehydrogenase simultaneously produces H_2 and prevents the accumulation of NAD(P)H produced by the malic enzyme in the hydrogenosome. However, in this direction the reaction is energetically very unfavorable ($\Delta G \approx 34 \pm 5.9$ kJ/mol assuming physiologically realistic conditions). Therefore, the flux bounds of this reaction in the hydrogenosome were set to reflect the assumption that the hydrogen dehydrogenase only carries flux in the forward, energetically feasible, direction.

4.2.4 iSW587 accurately predicts substrate utilization and *in vivo* fluxes

The curated model was validated using a combination of growth curves, extracellular metabolite and metabolic flux analysis (MFA) data. Substrate utilization tests were done on 36 different carbon sources, focusing on metabolites that are present in the natural

environment of the anaerobic fungi, see Table 4.5. The qualitative prediction accuracy of the model for the substrate utilization and vitamin essentiality validation tests is 89%. Interestingly, despite the presence of a full xylose isomerase pathway, *N. lanati* did not grow using xylose as its sole carbon source, as has been found in other gut fungi (Henske, Wilken, *et al.*, 2018). In this case, the model's predictions were incorrect. Cellular regulation or co-factor imbalances might explain this discrepancy (Henske, Wilken, *et al.*, 2018). Vitamin essentiality tests were also conducted, as shown in Table 4.5. It was found that both heme and 4-aminobenzoate were essential for growth, in agreement with the model's predictions. In other gut fungi, heme has also been found to be essential (Orpin and Greenwood, 1986), suggesting that its *de novo* biosynthesis pathway may be absent across the clade. It was found that only cysteine could be used as a sulfur source. However, it is not clear if this is a nutritional requirement since every other reducing agent tested (Na_2S, 2-mercaptoethanol and dithiothreitol) appeared to be toxic to the fungus. Since cysteine was used to ensure anaerobicity of the media, we could not test nitrogen source utilization.

Table 4.5: Substrate utilization table suggests that the model accurately captures phenotypic behavior of *N. lanati*. The model accurately predicts phenotypic responses in 89% of the tested cases. See the methods section for details about the experiments that yielded these results. + indicates that the model predicted growth/there was experimentally observed growth, while – denotes the opposite.

	Substrate	Model prediction	Experimental observation
Carbon utilization	Glucose	+	+
	Cellobiose	+	+
	Sorbitol	+	-
	Fructose	+	+
	Galactose	+	-
	Maltose	+	+

Mannose	+	-
Sucrose	+	+
Xylose	+	-
Arabinose	-	-
Rhamnose	-	-
Pyruvate	-	-
Succinate	-	-
Citrate	-	-
Glycerol	-	-
Pectin	-	-
Cellulose	+	+
Lignocellulose	+	+
Acetate	-	-
Fumarate	-	-
N-acetyl-glucosamine	-	-
Lactate	-	-
Maltodextrin	+	+
Methanol	-	-
Oxaloacetate	-	-
Xylan	+	+
Ethanol	-	-
Malate	-	-
Formate	-	-
Raffinose	+	+
Phenylalanine	-	-
Arginine	-	-
Leucine	-	-

	Proline	-	-
	Serine	-	-
	Threonine	-	-
Vitamin essentiality	Pyridoxine	+	+
	p-aminobenzoic acid	-	-
	Biotin	-	+
	Cyanocobalamin	+	+
	Riboflavin	+	+
	Folic acid	+	+
	Pantothenate	+	+
	Nicotinic acid	+	+
	Thiamin	+	+
	Heme	-	-

Metabolic flux analysis was also used to experimentally verify the predicted intracellular fluxes of the GSM. A [1,2]-^{13}C labeled glucose tracer was used in conjunction with a carbon atom transition model built from the *N. lanati* metabolic reconstruction (see the supplement). For the MFA model, metabolic degeneracy caused by the ability of the hydrogenosome to metabolize both malate and pyruvate resulted in large bounds on the fluxes involving these metabolites. To circumvent this, the MFA model was constrained to only import pyruvate into the hydrogenosome, based on previous observations (Boxma *et al.*, 2004). Extracellular metabolic product measurements (ethanol, formate, H_2, acetate, succinate, lactate) were also used to constrain the MFA model. This resulted in accurate internal metabolic flux measurements based on a statistically significant fit between measured and simulated proteinogenic amino acid labelling patterns, as shown in Figure 4.4. These measured fluxes

were then compared to the fluxes predicted using the GSM under the same constraints. Parsimonious-FBA (pFBA) was then used to find unique flux predictions. Using these constraints, the coefficient of determination between the pFBA and MFA simulation was found to be 0.98, as shown in Figure S5 in the supplement. This suggests that the metabolic model accurately predicts the steady-state measured intracellular fluxes of *N. lanati*.

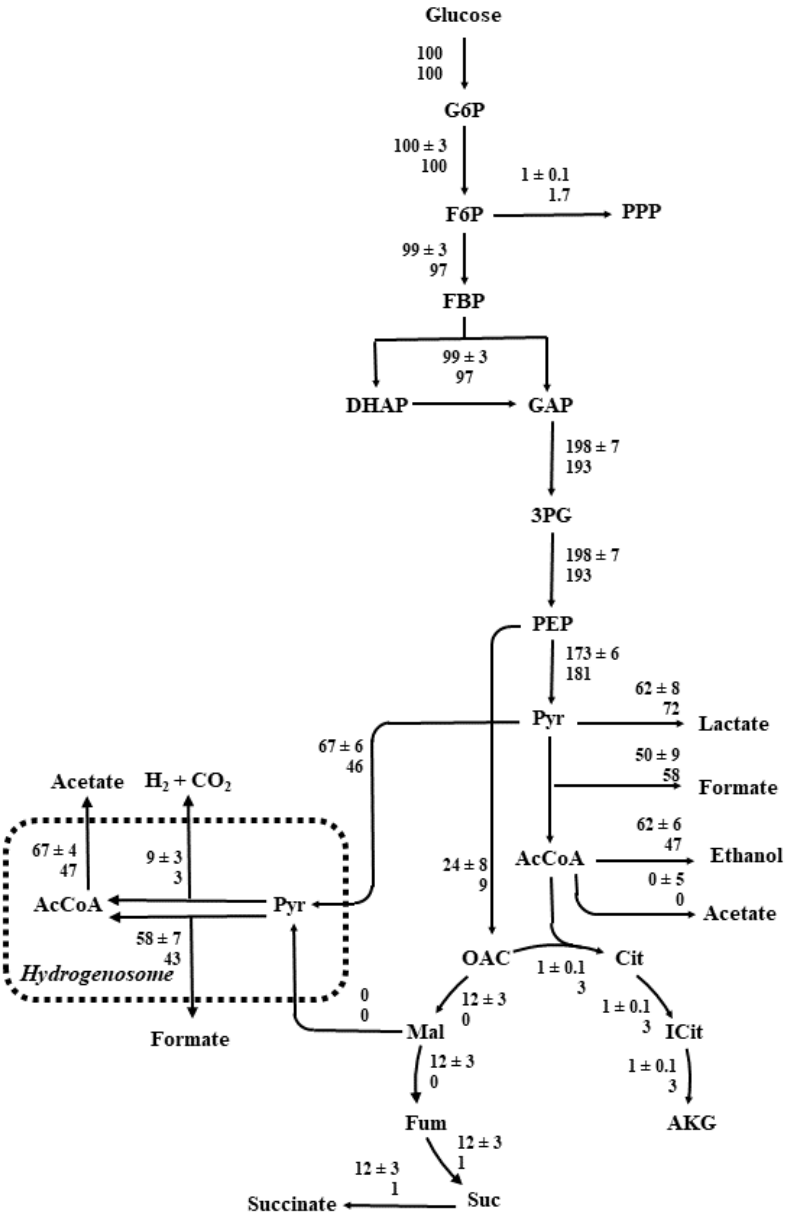

Figure 4.4: The genome-scale model accurately predicts the *in vivo* carbon metabolism of *N. lanati*. Experimentally determined MFA fluxes and predicted pFBA fluxes (top and bottom respectively) for glycolysis, the TCA cycle and the hydrogenosome of *N. lanati*. Error estimates denote one standard deviation from the reported mean for the MFA measurements. Three serially passaged [1,2]-^{13}C glucose tracer experiments, grown in M2 media at 39°C and harvested during exponential phase, were used to measure the *in vivo* fluxes (see the methods section for more details). The MFA model is included in the supplement.

4.2.5 The core hydrogenosome metabolism uses PFO to produce hydrogen but PFL carries the most flux

There remains uncertainty regarding the presence of pyruvate ferredoxin oxidoreductase (PFO), and its relative importance in the hydrogenosomal metabolism of anaerobic fungi. Earlier enzymatic characterization of hydrogenosomal proteins in Neocallimastigomycota suggested that PFO is the primary route for H_2 production through an associated ferredoxin hydrogenase, as found in the hydrogenosomes of other organisms (Yarlett *et al.*, 1986; Marvin-Sikkema *et al.*, 1993, 1994b; Muller *et al.*, 2012). However, more recent studies suggest that PFO is either absent, or of only marginal importance in the gut fungal hydrogenosomal metabolism (Akhmanova *et al.*, 1999; Boxma *et al.*, 2004). These later studies suggest that pyruvate formate lyase (PFL), which was likely acquired through horizontal gene transfer from bacteria (Stairs, Roger and Hampl, 2011), is significantly more active than PFO. It has been suggested that hydrogen evolution occurs through a hydrogen dehydrogenase working in an energetically infeasible reverse direction (Boxma *et al.*, 2004; Youssef *et al.*, 2013). Both PFO and PFL were identified in all published gut fungal genomes, as well as in *N. lanati*, as shown in Table 4.4. Therefore, we used the model to reconcile the

role and relative importance of these two enzymes to hydrogenosome function under steady-state growth conditions.

Due to the reaction stoichiometry of PFL, the molar ratio of formate to acetate and ethanol[3] produced is expected to approach unity if PFL is metabolically dominant (Akhmanova *et al.*, 1999). Since PFO only produces acetate, and not formate, the ratio of formate to acetate and ethanol will not be unity if PFO carries significant metabolic flux. Figure S6.A in the supplement shows that the experimentally measured molar ratio of formate to acetate and ethanol is not significantly different (using the unequal variance test, $p<0.05$) from unity for *N. lanati*, similar to earlier metabolite measurements for *P. sp. E2*, suggesting that PFL is dominant (Boxma *et al.*, 2004). Figure S6.B shows that the unconstrained model predicts a wide range of possible ratios, reflecting the metabolic degeneracy of the carbon metabolism of *N. lanati*. Since there is no energetic cost associated with using PFO versus PFL, the model predicts that both could be used. However, external metabolite flux measurements show only modest H_2 production (see Table 4.6), suggesting that cellular regulation may play a role in diverting flux to PFL instead of PFO. This can also be seen in the relative expression difference between PFL and PFO (an order of magnitude difference between them) in Table 4.4. By constraining the model's H_2 flux to the observed values, the range of possible ratios is reduced to those observed experimentally, as shown in Figure S6.B. Since PFO is the only (known) energetically feasible way to produce H_2, this result is not surprising. Using this constraint, the model suggests that PFL carries the most flux in the hydrogenosome, but that PFO is used to produce H_2.

[3] Ethanol is produced from PFL (via acetyl co-enzyme A) in the cytosol.

Table 4.6: Experimentally measured external fluxes of various metabolites produced by *N. lanati* growing on cellobiose in M2 media during exponential phase. See the methods section for details.

Metabolite	Mean flux [mmol/g_{DW}/h]	Standard deviation [mmol/g_{DW}/h]	Lower bound [mmol/g_{DW}/h]	Upper bound [mmol/g_{DW}/h]
Succinate	0.03	0.01	0.02143	0.045297
Lactate	0.87	0.14	0.716237	1.089
Ethanol	0.66	0.20	0.47280	1.0135
Formate	1.40	0.30	1.0875	1.7930
Acetate	0.56	0.12	0.42398	0.7118
H_2	0.10	0.06	0.0474	0.189

4.2.6 Electron bifurcation and proton pumping likely form part of the hydrogenosomal metabolism

Electron bifurcation is an energy conservation mechanism that can be used to drive thermodynamically unfavorable reactions by coupling endergonic and exergonic reactions through an enzyme complex (Müller, Chowdhury and Basen, 2018a). In the simplest case, this phenomenon is used by anaerobes to increase the yield of ATP through their carbon metabolism by using H_2 as an electron sink for the recycling of NADH to NAD^+ (Müller, Chowdhury and Basen, 2018b). In the case of the anaerobic gut fungi, the hydrogenosome can be used to generate 2 extra moles of ATP for every mole of glucose that enters glycolysis. However, not all the glycolytic flux can be diverted to the hydrogenosome because NAD^+ needs to be regenerated from the NADH that is produced by glycolysis to maintain cellular redox balance. As mentioned before, NAD^+ is unlikely to be produced by the hydrogen dehydrogenase since the redox potential of NADH/NAD^+ is too electropositive to reduce H^+

directly (Muller *et al.*, 2012). On the other hand, the ferredoxin-based hydrogenase included in the model only recycles the oxidized ferredoxin, produced by PFO, to reduced ferredoxin and does not impact the NADH/NAD^+ pools in the cell. However, sequence homology suggests that the *N. lanati* hydrogenosome could potentially house a bifurcating hydrogenase, see Table S3 in the supplement, which would couple the reduction of H^+ to the oxidation of NADH through the ferredoxin produced by PFO. This enzyme complex would allow more flux to be channeled into the hydrogenosome for energy production, since the hydrogenase would now generate NAD^+ as well as H_2 (see Figure 4.3).

Overall, these findings suggest that there is a significant energetic advantage associated with possessing a bifurcating hydrogenase. The model captures this benefit by predicting a 13% increase in growth rate associated with the use of the bifurcating hydrogenase, as opposed to the ferredoxin hydrogenase (μ=0.053 1/h vs. μ=0.047 1/h, respectively). Additionally, the model also dictates that the production of NAD^+ shifts from the cytosol to the hydrogenosome when the bifurcating hydrogenase is used, as shown in Tables 4.7 and 4.8. However, this metabolic necessitates flux being diverted from PFL to PFO in the hydrogenosome. Consequently, significantly more H_2 is predicted to be produced, which is not observed experimentally, as shown in Table 4.9. This discrepancy could be due to metabolic regulation that is unaccounted for in the GSM. Further experimental work needs to be done to investigate the potential presence of the bifurcating hydrogenase in the anaerobic gut fungal hydrogenosome.

Table 4.7: Relative predicted flux of NAD^+ producing reactions without the bifurcating hydrogenase using parsimonious flux balance analysis (μ=0.046 1/h). Acetaldehyde dehydrogenase = ACALD, ALCD2x = alcohol dehydrogenase, MDH = malate dehydrogenase.

NAD^+ producing reactions	% of total flux	Localization

ACALD	42	Cytosol
ALCD2x	42	Cytosol
MDH	16	Cytosol

Table 4.8: Relative flux of NAD^+ producing reactions when the model includes a bifurcating hydrogenase using parsimonious flux balance analysis (μ=0.053 1/h). Acetaldehyde dehydrogenase = ACALD, ALCD2x = alcohol dehydrogenase, HYDhbi = bifurcating hydrogenase, MDH = malate dehydrogenase. NAD^+ is transported out of the hydrogenosome by a putative aspartate-malate shuttle.

NAD^+ producing reactions	% of total flux	Localization
ACALD	15	Cytosol
ALCD2x	15	Cytosol
MDH	20	Cytosol
HYDhbi	50	Hydrogenosome

Table 4.9: H_2 flux predicted by the unconstrained model using the ferredoxin and bifurcating hydrogenases, as well as the experimentally measured values. Predicted flux bounds are based on 2000 samples of the relevant model, the biomass objective function was constrained to carry at least 90% of the flux of the optimal solution.

Hydrogenase type in model	Mean H_2 flux [mmol/g_{DW}/h]	St. D. H_2 flux [mmol/g_{DW}/h]
Unconstrained ferredoxin	0.73	0.42
Unconstrained bifurcating	4.28	0.41
Experimentally measured H_2 flux	0.10	0.06

Given that the hydrogenosome in anaerobic fungi is relatively understudied, yet related to the mitochondrion (Boxma *et al.*, 2005; Muller *et al.*, 2012), we assumed that their metabolite trafficking machinery is similar. Specifically, we assumed that the hydrogenosome has an inner membrane that is impermeable to H^+, like mitochondria (Zorova *et al.*, 2018). This

implies that H^+ can only enter and leave the hydrogenosome through the action of transporters or the speculative complex 1 and complex 2 proton pump introduced earlier. The H^+ balance in the hydrogenosome has a direct effect on the ability of the putative ATP synthase to produce ATP. However, with or without the proton pumping mechanism of complex 1 and 2, the impact of the ATP synthase on ATP production was found to be small, with it only supporting small fluxes (~2% of the glucose flux into the model) as shown in Figure S7 in the supplement. This suggests that the putative hydrogenosomal proton gradient, as shown in Figure S4 in the supplement, may not be very important for the generation of ATP, as is also suggested by the low expression of the ATP synthase complex subunits (see Table 4.4). It could also be that the ATP synthase is actually a V-type ATPase and not present in the hydrogenosome (Seppälä *et al.*, 2016). Furthermore, the mechanism by which the putative complex 2 operates remains to be answered without the membrane bound subunits of complex 1.

4.2.7 Metabolic degeneracy is related to the regeneration of NAD^+

The modeled gut fungal metabolism displays significant degeneracy, as shown by the high degree of flux variability in the unconstrained model, as shown in Table 4.10. The degeneracy is primarily due to the ability of *N. lanati* to regulate how NAD^+ is regenerated through its mixed acid fermentation pathways, i.e. through a combination of lactate dehydrogenase, acetaldehyde dehydrogenase and alcohol dehydrogenase. Interestingly, the relative mean error between the predicted flux distributions and experimental measurements of the fermentation products is much more sensitive to constraints placed on acetate production than any other single measured external metabolite flux, as shown in Figure 4.5. Likewise, constraints placed on lactic acid flux also narrow the deviation of the predicted flux distributions. This effect is

due to the different yields of NAD^+ that can be achieved per mole of pyruvate depending on which mixed acid fermentation pathway, or combination thereof, is constrained. Without the bifurcating hydrogenase, H_2 production does not significantly impact the overall redox balance of the cell. This possibly explains why its constraint has the smallest effect on the flux variability predicted by the model and may allow the cell to fine tune its metabolism to suit the environmental needs, e.g. sugar availability, by up- or down-regulating the flux channeled to the hydrogenosome (Hackstein *et al.*, 2019).

Table 4.10: Flux variability analysis indicates that there is significant variability in the production rates of metabolic products that all lead to near optimal growth rates (the optimal growth rate, $\mu = 0.0465$ 1/h, was lower bounded to 90% of its optimum).

Metabolite flux (produced)	Lower bound [mmol/g_{DW}/h]	Lower bound [mmol/g_{DW}/h]
Formate	0.89	2.94
Acetate	1.00	1.76
Ethanol	0.72	1.77
Lactate	0	1.05
H_2	0	1.76

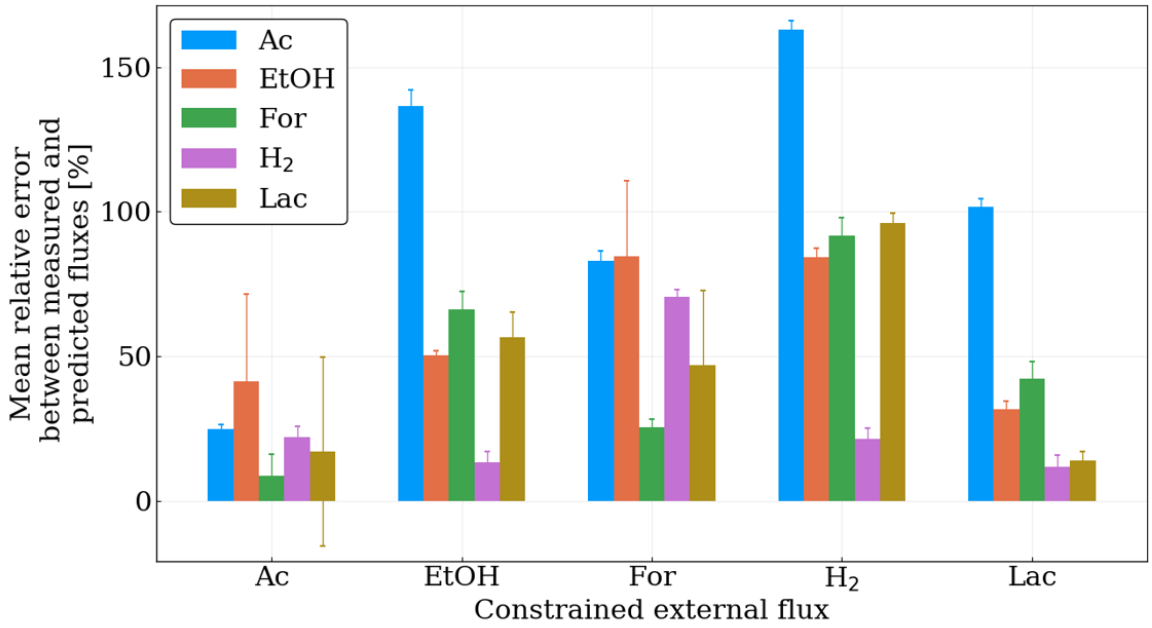

Figure 4.5: The absolute relative error between the model predictions and the experimentally measured values suggest that constraining the flux of acetate production has the biggest impact on the model's accuracy. The flux of acetate (Ac), ethanol (EtOH), formate (For), H_2 and lactate (Lac) was constrained individually to their observed ranges (variables on the x-axis). The resultant predicted fluxes of these metabolites (generated by sampling 2000 possible solutions where the biomass objective function was within 90% of its optimal value and subject to the respective additional constraints as shown in the figure) were then compared to the experimental observations as shown in the legend.

4.2.8 Model predicted fluxes suggest strategies to engineer *N. lanati* for bioprocessing

While anaerobic gut fungi specialize in lignocellulose decomposition, they are not well-developed bio-production platforms compared to model microbes *Escherichia coli* and *Saccharomyces cerevisiae* (Chubukov *et al.*, 2018). It has been suggested that pairing the gut fungi with one or more microbial cell factory organisms may offer an alternative approach that avoids the bottleneck associated with engineering and optimizing gut fungal metabolism for commodity chemical production (Ranganathan *et al.*, 2017; Henske, Wilken, *et al.*, 2018). Indeed, similar approaches have been used to pair *Trichoderma reesei*, a cellulose degrading specialist, with *E. coli* to produce isobutanol from pretreated corn stover (Minty *et al.*, 2013).

Furthermore, it has been suggested that such approaches are more likely to be successful if the microbes are engineered to co-operate, instead of compete, with each other for resources (Mee *et al.*, 2014).

To make use of the lignocellulolytic capacity of *N. lanati* in a bioprocessing context, it is important to understand its metabolic limitations, for example, the maximum metabolic fluxes that can be diverted through metabolic engineering without excessively compromising cellular viability. Figure 4.6 shows the theoretical maximum flux of an assortment of metabolites that can be siphoned away from the primary metabolism of *N. lanati* with the constraint that its growth rate does not fall below 90% of its maximum growth rate predicted by the model. These metabolites can be used to either pair another organism with *N. lanati* through nutritional auxotrophies or can be used to produce value added chemicals directly from the fungi. It is interesting to note that H_2 seems to be the best candidate metabolite to explore for bioprocessing because the model suggests that it has the highest available flux yield that can be used without compromising growth. This is in agreement with literature sources that indicate that stable fungal/H_2 consuming methanogen co-cultures are readily formed (Gilmore *et al.*, 2019; Li *et al.*, 2019). Furthermore, measured extracellular accumulation of amino acids, as shown in Figure S8 in the supplement, qualitatively agree with the predictions of Figure 4.6A. In particular, alanine, aspartate and glutamate each accumulate to the highest concentration (relative to the other amino acids) in the measured data, which corresponds to the predictions of Figure 4.6A that *N. lanati* can produce these amino acids in excess with the least effect on its growth rate. This suggests that amino acid auxotrophies may be another

route that can be used to ensure that a faster growing microbial partner does not outcompete *N. lanati*.

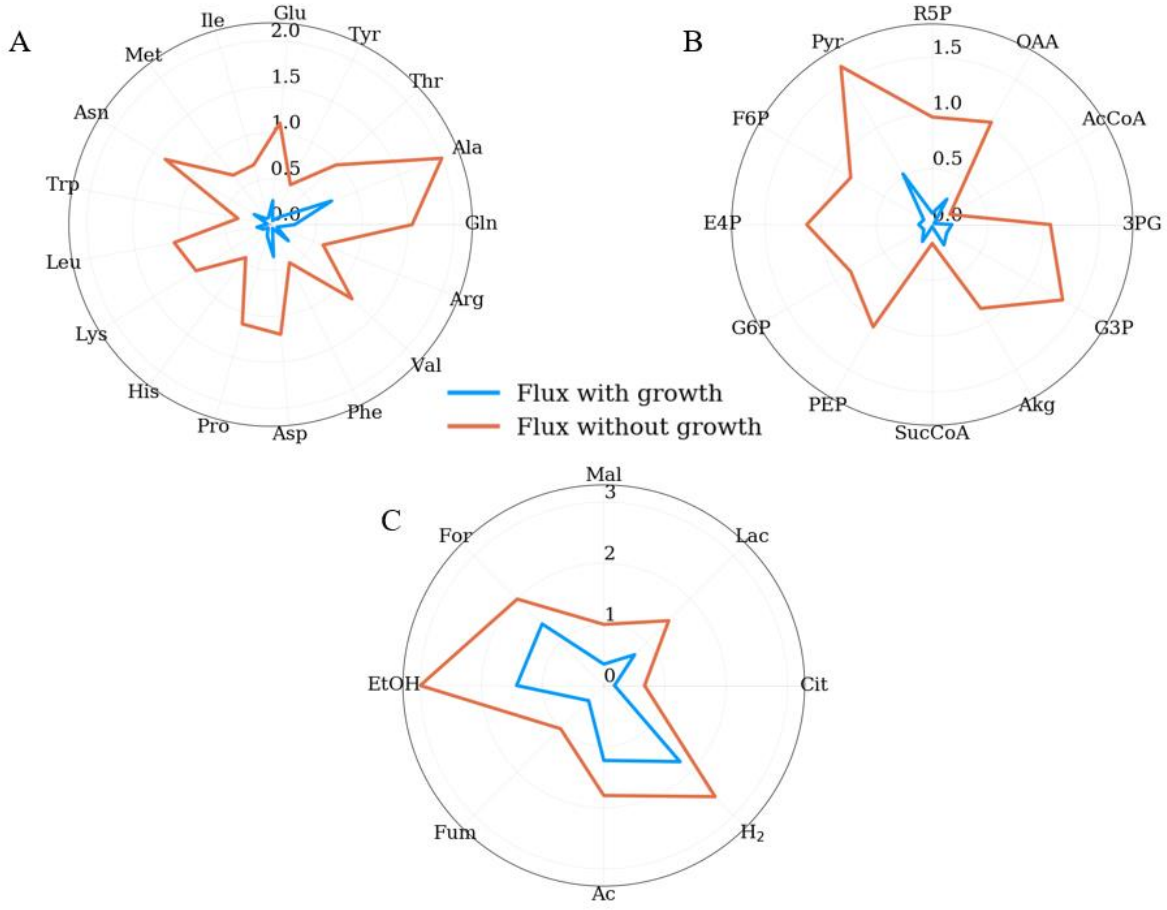

Figure 4.6: The model can be used to predict the maximum available flux of metabolites that can be re-directed without compromising the growth rate of *N. lanati*. The units of each plot in the radial direction are mmol/g_{DW}/h. (A) The flux of amino acids, (B) precursor metabolites, and (C) metabolic by-products of fermentation that can be re-directed without compromising the overall growth rate of *N. lanati* by more than 10%. The orange data are the theoretical maximum fluxes that can be diverted if *N. lanati* channels all metabolic flux towards the production of each of the metabolites (still subject to the mass balance of its genome), while the blue data correspond to the case where the growth rate is constrained to not fall below 90% of the optimum growth rate. For each metabolite a dummy sink reaction was added to allow the respective metabolite to be generated in excess in the model. Legend: All the standard 3 letter abbreviations for amino acids were used, as well as Glu = glucose, Pyr = pyruvate, OAA = oxaloacetate, AcCoA = acetyl-coenzyme A, 3PG = 3-phospho-D-glycerate, G3P = glyceraldehyde 3-phosphate, Akg = 2-oxoglutarate, SucCoA = succinyl-coenzyme A, PEP = phosphoenolpyruvate, G6P = D-glucose 6-phosphate, E4P = D-erythrose 4-phosphate, F6P = fructose-6-phosphate, R5P = ribose-5-phosphate, Mal = malate, Lac = lactate, Cit = citrate, Ac = acetate, Fum = fumarate, EtOH = ethanol, For = formate.

4.3 Conclusion

Here we have introduced a high-quality genome and transcriptome of a novel anaerobic gut fungus, *N. lanati*. While the genome is large, it is relatively unfragmented compared to the genomes of the other sequenced anaerobic gut fungi. Additionally, the genome encodes for a large number and diversity of CAZymes, most of which are expressed in the transcriptome. This genome was used to construct the first genome-scale metabolic model of an anaerobic gut fungus. The model, iSW587, accurately recapitulates the observed growth rate, *in vivo* fluxes and substrate consumption/requirement profiles. The model refines and expands on our understanding of gut fungal hydrogenosomal metabolism. We confirm previous findings that suggested that PFL carries more flux than PFO in the hydrogenosome, but an energetically favorable route to hydrogen production still requires the action of PFO. The possible presence of a bifurcating hydrogenase and/or a proton pumping mechanism suggests that anaerobic fungi may have evolved more complex energy conservation mechanisms that allow them to compete with faster growing rumen bacteria (Gilmore *et al.*, 2019). Experimental work, likely involving the isolation, purification, and enzymatic characterization (through assays and proteomic analysis) of the hydrogenosome is necessary to further refine our understanding of its metabolism. This model is well poised to serve as a platform to build a better understanding of these non-model organisms. Moreover, the model will serve as a valuable tool to systematically guide future engineering efforts of gut fungi for converting lignocellulose into value-added products.

4.4 Materials and methods

4.4.1 Metabolic reconstruction, visualization and simulation

All publicly available annotated genomes within the clade Neocallimastigomycota were downloaded from the Joint Genome Institute's (JGI) Mycocosm database (Grigoriev *et al.*, 2014). This includes the high quality PacBio sequenced genomes of *Anaeromyces robustus*, *Piromyces finnis*, and *Neocallimastix californiae* (Haitjema *et al.*, 2017a), as well as the novel isolate *Neocallimastix lanati* introduced here. The genomes of *Pecaromyces ruminantium*, also known as *Orpinomyces sp. C1A* (Youssef *et al.*, 2013; Hanafy *et al.*, 2017), and *Piromyces sp. E2* (Haitjema *et al.*, 2017a) were also included for completeness. The gene annotation data supplied by the JGI was combined with annotations derived from bi-directionally blasting (using BLASTp (Camacho *et al.*, 2009)) the predicted genes from the gut fungal genomes against the curated Swiss-Prot database from Uniprot ('UniProt: a worldwide hub of protein knowledge', 2019). Briefly, bi-directional blasting annotates a predicted gut fungal gene if 1) the top hit using the fungal genome as the query and the reference collection as the database is the same as when 2) the gut fungal genome is used as the database and the reference collection is used as the query. Furthermore, only matches with e-values smaller than $1e^{-20}$ were considered for assigning Enzyme Commission (EC) annotations to genes. This information was collated into a master metabolic table, see the Supplement, and subsequently used to construct the model and assign genes to reactions. Enzyme complexes were assigned by using the "Subunit structure" field in the Uniprot database. Protein localization was predicted using DeepLoc (Almagro Armenteros *et al.*, 2017). Reaction directions were primarily inferred from MetaCyc (Caspi *et al.*, 2018), and the

Gibbs free energy change of a reaction calculated using eQuilibrator (Flamholz *et al.*, 2012). Transcriptomic and expression experiments for *N. lanati* were conducted as part of this study. These omics datasets were used to assign a confidence score to each gene in the model of *N. lanati*. Gaps in the model of *N. lanati* were filled by inspecting the EC assignments found for each fungus using the approach described above and looking for homologous genes in the genome of *N. lanati*. The universal reactions and metabolites from the BiGG Models platform (King *et al.*, 2016) was used to construct the *in silico* model where possible; if a reaction did not exist in that database it was manually added. The KEGG and MetaCyc databases were used as references to reconstruct the metabolic model based on the EC assignments of the metabolic annotation data (Kanehisa *et al.*, 2016; Caspi *et al.*, 2018). The curated model for *N. lanati* was constructed by carefully following established genome-scale model construction protocols (Thiele and Palsson, 2010). Specifically, each reaction was inspected to ensure consistency, mass and charge balance where possible. Model quality was benchmarked by the Memote application (Lieven *et al.*, 2020). The curated *N. lanati* model can be found in the Supplement. An experimentally measured flux of 1.5 mmol/g_{DW}/h of glucose was used in all simulations. Flux balance analysis was used to simulate the genome-scale model of *N. lanati* using the COBRA Toolbox and Cobrapy (Ebrahim *et al.*, 2013; Heirendt *et al.*, 2019). Flux samples (N=2000) were generated by sampling from the model and constraining the objective function to be within 90% of the optimum found by FBA. Escher was used to visualize the metabolism (Rowe, Palsson and King, 2018). Example code used to run the computational experiments is supplied in the Supplement.

4.4.2 Culturing conditions used for experiments

Standard anaerobic gut fungal culturing techniques were used (Haitjema *et al.*, 2014) for all experiments. Briefly, *N. lanati* was grown at 39°C in sealed Hungate tubes (10 mL liquid volume) or 70 mL serum bottles (40 mL liquid volume) in both undefined complex medium C (MC) (Davies *et al.*, 1993), as well as completely defined medium 2 (M2) (Teunissen,' *et al.*, 1991), with 100% CO_2 headspace unless otherwise specified. Pressure accumulation was used as a proxy for growth and the fungus was serially passaged after 2-3 days of growth. The carbon source was cellobiose at 5 g/L unless otherwise noted. The cultures were not shaken.

4.4.3 Genome and transcriptome isolation, sequencing and analysis of *N. lanati*

N. lanati was isolated from a sheep located at the Santa Barbara Zoo, following an established protocol (Solomon *et al.*, 2016). Fungal cell pellets for gDNA isolation were grown by inoculating 20 mL from a serum bottle of fungi in exponential phase (2 - 3 days of growth given a 10% inoculation volume into the serum bottle) into a 1 L bottle of medium C using cellobiose as a carbon source. The serum bottle used to grow the inoculum was treated with Chloramphenicol to reduce the risk of contamination. After 4 days of growth the fungal cell mat was spun down and frozen at -80°C. Four of these frozen samples were subsequently shipped to the Arizona Genome Institute (University of Arizona, Tucson, AZ), where high quality gDNA was isolated using a modified Cetrimonium bromide (CTAB) protocol (Doyle and Doyle, 1987). Briefly, these fungal cell mats were ground to a fine powder in a frozen mortar with liquid N_2 followed by very gentle extraction in CTAB buffer, which included proteinase K, PVP-40 and 2-mercaptoethanol (Sigma, St. Louis, Missouri), for 1 hour at 50°C. After centrifugation, the supernatant was gently extracted twice with 24:1 chloroform:iso-

amyl alcohol. The upper phase was removed and adjusted to 1/10th volume with 3M KAc, gently mixed, and the gDNA precipitated with iso-propanol. Subsequently, the gDNA was collected by centrifugation, washed with 70% ethanol, air dried for 20 min and dissolved thoroughly in 1x TE buffer at room temperature. The purified gDNA was shipped to the JGI where it was sequenced.

RNA for transcriptome and expression analysis was isolated as previously described (Solomon *et al.*, 2016; Henske, Wilken, *et al.*, 2018) in the Biological Nanostructures Lab (University of California Santa Barbara, CA). For the transcriptome, the RNA was harvested from fungal cell pellets grown in serum bottles on a variety of substrates (cellobiose, filter paper, reed canary grass and corn stover, solids loading 1% w/v, in both medium C and medium 2) to capture as much transcript diversity as possible. For expression analysis triplicate serum bottles of fungus grown on medium C and medium 2, using cellobiose as the sole carbon source, was used. The RNA was isolated and purified using an RNEasy kit (Qiagen, Germantown, MD). The concentration and quality of the RNA was measured on a Qubit (Qubit, New York, NY) and Tapestation 2200 (Agilent, Santa Clara, CA). The RNA used for the transcriptome was pooled in equal parts before sequencing. RNA libraries were made using NEBNext Ultra II Directional RNA with mRNA purification beads (NEB, Ipswich, MA), these were subsequently sequenced on a NextSeq 500 (Illumina Inc., San Diego, CA) using High Output 300 Cycle settings and 75 base-pair paired-end reads (the resultant coverage is 470 and 364 for the transcriptome and expression analysis, respectively). The reads were assembled using Trinity (Grabherr *et al.*, 2011). TransDecoder (https://github.com/TransDecoder/TransDecoder/wiki) was used to find the highest likelihood

coding regions in the transcriptome. The transcript abundance was estimated using Kallisto (Bray *et al.*, 2016).

4.4.4 High performance liquid chromatography (HPLC), gas chromatograph (GC) and liquid chromatography/mass spectrometry (LC/MS) measurements

Liquid samples for HPLC analysis were stored in microcentrifuge tubes at -20°C for batch analysis. Sulfuric acid (0.5 M) was added to the samples (1 in 100 volumes), vortexed and allowed to mix at room temperature for 5 minutes. Thereafter the samples were centrifuged for 5 min at 21000×g and filtered using a 0.22μm syringe filter into HPLC vials. The samples were run on an Agilent 1260 Infinity (Agilent, Santa Clara, CA) using a Bio-Rad HPX-87H column (Bio-Rad, Hercules, CA). Samples were run at two column conditions to effectively separate all the fermentation products (Qiu and Jin, 2002; Chinnici *et al.*, 2005). Succinate, lactate, cellobiose, glucose and ethanol were measured at 50°C with a flow rate of 0.5 mL/min. Fumarate, formate and acetate were measured at 25°C with a flow rate of 0.4 mL/min. The mobile phase in both cases was 5mM sulfuric acid and the injection volume 20 μL. Cellobiose, glucose and ethanol were measured with a refractive index detector, and the other compounds on a variable wavelength detector (λ=210 nm). Standard curves for each compound were made at 3 concentrations bracketing the range expected in the samples. Gas samples were analyzed on a Thermo Fisher Scientific TRACE Gas Chromatograph according to a previously established protocol (Gilmore *et al.*, 2019). Standard curves of H_2 were made daily from supplier (Douglas Fluid & Integration Technology, Prosperity, SC) mixed gas at 1, 2, and 5% (mole basis). Fatty acid composition and extracellular amino acid concentrations were measured on an LC/MS using an Agilent Polaris 3 C18-Ether 150x3.0mm (part number

A2021150X030) column. Total run time was 13 minutes, injection volume was 5 ul with column temperature of 30°C and flow rate of 0.3 mL/min. The MS was run in negative ionization mode, with the gas temperature at 230°C and flow rate 12 L/min. Standards were made to bracket the expected concentration ranges of the fatty acids.

4.4.5 Biomass composition measurements

The biomass composition of *N. lanati* was experimentally determined in triplicate for each major component (DNA, RNA, lipid, carbohydrate, protein). Cultures grown in serum bottles on medium C, with cellobiose as the carbon source, were harvested during exponential phase for each measurement. DNA and RNA were immediately isolated from the wet cell pellet using a previously established CTAB protocol (Lankiewicz, Cottrell and Kirchman, 2016). A reference set of triplicate fungal mats were harvested at the same time and lyophilized to estimate the dry mass fraction of DNA and RNA. Lipids and total carbohydrates, which we assumed to be exclusively chitin, were isolated following established protocols (Beck, Hunt and Carlson, 2018). Lipid composition was further refined by mass spectrometry using a fatty acid C18-ether column as described in the methods section. Protein extraction from fungal cell pellets followed a previously developed method (Soh *et al.*, 2014) and the concentration was measured on a Qubit (Q33327, Thermo Fisher Scientific). The amino acid composition of the protein fraction was assumed to follow the amino acid distribution of the predicted proteome.

Additionally, both the growth and non-growth associated maintenance functions (GAM and NGAM) were estimated from experimental data. Briefly, five triplicate sets of Hungate tubes with varying concentrations of cellobiose (1, 2, 3, 4, 5 g/L initially) in medium 2 were

inoculated with 1 mL each (total liquid volume 10 mL) from a single serum bottle of *N. lanati* growing at exponential phase (3 days post inoculation) in medium 2 with cellobiose as the carbon source. Pressure accumulation was measured twice daily to calculate the fungal growth rate (Theodorou *et al.*, 1995). Liquid and gas samples from each triplicate set was harvested during two time points in exponential phase 24 hours apart. The gas samples were analyzed on a GC to determine the H_2 fraction of the gas, and the liquid samples were analyzed on an HPLC for organic acid concentration. At the end of the last measurement the samples were harvested, the fungal cell pellet spun down, lyophilized and weighed. The estimated growth rate for each sample was then used to extrapolate the dry cell mass at the respective time points (Theodorou *et al.*, 1995). The fluxes of the fermentation products could then be estimated by the molar accumulation of each compound divided by the time between measurements and the difference in cell dry masses between these points. The difference in cell masses was taken because each mature cell lyses and dies, thus its remaining biomass no longer contributes to metabolism. Finally, these estimated fluxes were used to constrain the model, and maximize the ATP yield. The GAM and NGAM was then estimated by finding the line of best fit through the plot of maximum ATP yield predicted by the model and the growth rate associated with the fluxes previously measured (Thiele and Palsson, 2010).

4.4.6 ^{13}C metabolic flux analysis for *N. lanati*

Three serially passaged Hungate tubes using [1,2]-^{13}C glucose as the sole carbon source in medium 2 at an initial concentration of 5 g/L were used for the labeling experiment. Each Hungate tube was passaged during exponential growth phase, after which the cell pellet and remaining media were frozen at -20°C for later processing. The media was analyzed for

glucose and fermentation products using the HPLC protocol described above. The pellets were lyophilized, after which GC-MS measurements were used to quantify the isotopic labeling of protein-bound amino acids, glycogen-bound glucose and RNA-bound ribose as described previously (Long and Antoniewicz, 2019). A carbon transition model for flux analysis was constructed using the genome-scale model of *N. lanati* as a basis for the flux reactions and biomass equation. Other carbon transition models were used to check the accuracy of the MFA model (Crown, Long and Antoniewicz, 2016; Liu, Qiao and Stephanopoulos, 2016). INCA was used to perform the flux analysis and sensitivity calculations (Young, 2014). The carbon transition model, constraints and the GC-MS data can be found in the supplement.

4.4.7 Model validation experiments

Carbon utilization and vitamin essentiality experiments were conducted to test the predictive accuracy of the model. Carbon utilization was tested by growing *N. lanati* in medium 2 with each carbon source listed in Table 4.5 at 5 g/L initial concentration instead of cellobiose. A carbon substrate was deemed able to support growth if the fungus could be passaged on it for 4 generations and still produce more than 8 PSIg of accumulated pressure (no carbon blanks produce < 1 PSIg of accumulated pressure). Similarly, the vitamin requirements of *N. lanati* were tested by individually removing each vitamin in medium 2 (listed in Table 4.5) and growing the fungus without it for 4 consecutive generations using cellobiose as the carbon source. Fluxes for comparing model predictions to experimental observations were measured similar to how the fluxes for finding the GAM and NGAM functions were estimated, however only 5 g/L cellobiose loading was used. The total equivalent flux of glucose into the cell was calculated by measuring glucose accumulation and

cellobiose depletion in the media. It was assumed that *N. lanati* imports glucose, and not cellobiose, due to release of beta-glucosidases that decomposed the cellobiose in the media.

V. *In Silico* identification of microbial partners to form consortia with anaerobic fungi

This chapter is based upon an article that was published in Processes, Volume 6, 2018, by St. Elmo Wilken, Mohan Saxena, Linda R. Petzold, and Michelle A. O'Malley, entitled "*In Silico Identification of Microbial Partners to Form Consortia with Anaerobic Fungi"* Copyright Processes. See the published paper for the supplementary information.

5.1 Introduction

Modern biotechnology is well poised to take advantage of the current shift towards a more sustainable chemical industry (Otero and Nielsen, 2010). Harnessing the estimated 1.6 billion tons of energy rich, lignocellulosic agricultural waste generated worldwide each year is a promising avenue towards this goal (Saini, Saini and Tewari, 2015). However, extracting cellulose (40–50%) and hemicellulose (20–40%) from raw plant biomass has proven to be challenging due to the high lignin content of the substrate (Liao *et al.*, 2016). Current industrial techniques used to overcome this barrier include physical, chemical and biological treatment (e.g., milling, acid hydrolysis and enzyme treatment, respectively) (Sindhu, Binod and Pandey, 2016).

Biological conversion attempts to exploit natural mechanisms to produce chemicals from lignocellulose. Currently, two competing alternatives are being investigated: consolidated bioprocessing and microbial consortia approaches (Alper and Stephanopoulos, 2009). The former seeks to engineer a single organism to both degrade biomass and produce a high value commodity chemical (Lynd *et al.*, 2005). The latter seeks to leverage specialist organisms to split the associated metabolic burden between them (Brenner, You and Arnold, 2008).

Exploiting the natural degradation powers of non-model fungi could prove beneficial in this endeavor.

Currently, fungal enzymes from a handful of organisms, e.g. *Trichoderma reesei* or *Aspergillus sp.*, are utilized on an industrial scale to break down plant biomass (Paloheimo *et al.*, 2016). A recent report illustrates the utility of developing consortia between a cellulose degrader like *T. reesei* and the model bacterium *Escherichia coli* (Minty *et al.*, 2013). A potential drawback of this pairing is that *T. reesei* encodes for the smallest diversity of cellulolytic enzymes of any fungus capable of plant cell wall degradation (Martinez *et al.*, 2009). This could necessitate the addition of (expensive) beta-glucosidases, to convert cellobiose to glucose, in some applications. It is hypothesized that under-explored fungal clades, like Neocallimastigomycota, could offer substantial benefits in this regard (S Seppälä *et al.*, 2017).

Anaerobic gut fungi, in the phylum Neocallimastigomycota, found in the gastrointestinal tract of ruminants, have been shown to be prodigious degraders of plant biomass (Resch *et al.*, 2013). Moreover, they possess the highest diversity of lignocellulolytic enzymes, largely untapped, within the fungal kingdom (Solomon *et al.*, 2016). These organisms play a pivotal role in the digestion of plant biomass in herbivores, due to the physical and chemical way in which they degrade plant biomass (Gruninger *et al.*, 2014). Recent work highlights the bounty of biotechnological applications of these fungi (Henske, Wilken, *et al.*, 2018). Given that these organisms typically thrive in consortia, it is desirable to emulate nature to unlock their potential for bioconversion of unpretreated lignocellulose.

However, these organisms are under-studied, and the mechanisms that promote the formation of stable microbial consortia with anaerobic fungi are unknown. Given the wealth

of omics-related data available, we speculate that model driven design could elucidate some of these questions (S Seppälä *et al.*, 2017). Indeed, model driven analysis has successfully been used to study anaerobic organisms (Senger, Yen and Fong, 2014). Necessary components for such analyses are accurate genome-scale models of anaerobic gut fungi and their consortia partners. While a full genome-scale model of the gut fungi is still under active development, it is possible to narrow the field in search of potential consortia partners by making use of extant high-quality genome-scale models to highlight mechanisms of interaction that would promote microbial partnership and consortium stability.

In this work, we present a marriage of experimental and computational tools used to identify suitable consortia partners for anaerobic gut fungi. Given the vast number of potential candidates, it is infeasible to experimentally test all combinations. Instead, we filter microbes by simulation to test their compatibility in silico. As a first approximation, we assume no interaction between the organisms in consortia: the excess fermentable sugars released by fungal hydrolysis of plant biomass, measured experimentally, is available for consumption regardless of the identity of the partner microbe. By predicting the growth rate and waste production of the partner, we can rank order microbes by the likelihood that they would stably co-exist with the gut fungi over the course of active fungal growth in a batch bioreactor. This is a valuable tool to reduce the number of costly and time-consuming wet-lab experiments necessary to identify suitable partners for anaerobic gut fungal-based consortia. Finally, we introduce a novel dynamic flux balance analysis algorithm specifically developed for this task.

5.2 Materials and Methods

5.2.1 Strains and Culture Conditions

Three isolated anaerobic gut fungi were investigated in this work: *Neocallimastix californiae*, *Anaeromyces robustus* and a previously uncharacterized fungus *Neocallimastix sp. S1* (confirmed by ITS sequencing, see the Supplementary Materials). Anaerobic conditions, as described in (Trinci *et al.*, 1994), were maintained for all experiments. Starter cultures for each experiment were grown on complex media (Theodorou *et al.*, 1996), with Reed Canary grass used as a substrate, in 75 mL serum bottles. After four days of growth, these cultures were used to start experiments by inoculating 4 mL from them into the experiment serum bottles. Gas accumulation in the head space of the starter cultures was vented daily. All experiments were conducted in triplicate using 40 mL of M2 media (Teunissen,' *et al.*, 1991) loaded with 2 g of corn stover grass, (4 mm particle size) supplied by the USDA-ARS research center (Madison, WI, USA), in 75 mL serum bottles.

5.2.2 Growth and Metabolite Measurements

Fungal growth was monitored by measuring pressure in the head space of the serum bottles twice daily, approximately 12 h apart (Theodorou *et al.*, 1996). Cultures that accumulated significantly more pressure than a control set, without the carbon source Corn Stover, were deemed to be growing. The gaseous product is primarily composed of hydrogen and carbon dioxide. After the pressure was measured, and prior to venting, 0.2 mL of media was sampled for sugar concentration analysis on a high-performance liquid chromatography (HPLC) device. Samples were stored at −20 °C for batch-wise analysis. After thawing the samples at room temperature, they were centrifuged for 5 min at 21,000×g. By avoiding the pellet, 100

μL was transferred to HPLC vials containing 100 μL de-ionized, 0.45 μm filtered water (1:1 dilution). Subsequently, 20 μL of each sample was run on an Agilent 1260 Infinity HPLC (Agilent, Santa Clara, CA, USA) using a Bio-Rad Aminex HPX-87P column (Part No. 1250098, Bio-Rad, Hercules, CA, USA) with inline filter (Part No. 5067-1551, Agilent, Santa Clara, CA, USA), Bio-rad Micro-Guard De-Ashing column (Part No. 1250118, Bio-Rad, Hercules, CA, USA), and Bio-Rad Micro-Guard CarboP column (Part No. 1250119, Bio-Rad, Hercules, CA, USA) in the following orientation: inline filter → De-Ashing → CarboP → HPX-87P columns. Samples were run with water acting as the mobile phase at a flow rate of 0.6 mL/min and column temperature of 60 °C. Signals were detected using a refractive index detector (RID) with a temperature set point of 40 °C. HPLC standards were created in triplicate for cellobiose, glucose, fructose, xylose and arabinose at 5 g/L, 1 g/L, and 0.1 g/L concentrations in M2. The concentration of each sugar was measured by subtracting the RID signal from a blank M2 sample.

5.2.3 Evaluation and Selection of Model Organisms

The BIGG database is an online repository of curated genome-scale metabolic models (King *et al.*, 2016). Currently (Accessed December 2017) the database consists of 84 models from a wide diversity of organisms. We hypothesized that the higher level of understanding implied by these models may be leveraged into the formation of stable consortia with the relatively understudied anaerobic fungi. The first step in identifying possible consortia partners is to screen the modeled organisms by three criteria: (1) is the organism an obligate aerobe, (2) is the organism pathogenic and (3) is the organism obviously incompatible with the anaerobic fungi? If any of these criteria were positive, the model was discarded. For

example, *Helicobacter pylori* is a modeled pathogen and is therefore excluded. In addition, *Thermotoga maritima* is a modeled hyperthermophilic bacterium; it cannot be co-cultured with the anaerobic fungi and is immediately discarded as a potential consortia partner. By filtering all 84 potential models, we are left with six possible partners, shown in Table 5.1.

Table 5.1: Genome-scale models of potential consortia partners for the un-modeled anaerobic gut fungi used in this work.

Organism	Notes	Reference
Clostridium ljungdahlii DSM 13528	Bacterium, obligate anaerobe, acetogen	(Nagarajan *et al.*, 2013)
Escherichia coli str. K-12 substr. MG1655	Bacterium, facultative anaerobe	(Monk *et al.*, 2017)
Escherichia coli str. ZSC113	Bacterium, facultative anaerobe, glucose deficient	(Curtis and Epstein, 1975)
Lactococcus lactis subsp. *cremoris* MG1363	Bacterium, facultative anaerobe	(Flahaut *et al.*, 2013)
Methanosarcina barkeri str. *Fusaro*	Methanogen, obligate anaerobe	(Feist *et al.*, 2006)
Saccharomyces cerevisiae S288C	Fungus, facultative anaerobe	(Mo, Palsson and Herrgård, 2009)

5.2.4 Dynamic Flux Balance Analysis Formulation

Flux balance analysis (FBA) is a widely used computational tool that simplifies and recasts the metabolic reaction network of a cell into a linear program by making use of a

genome-scale model (Orth, Thiele and Palsson, 2010). Central to FBA is the assumption of metabolic steady state, $\frac{dx}{dt} = \boldsymbol{Sv} = \mathbf{0}$. The space of fluxes, **v**, that satisfy the mass balance implied by the stoichiometric matrix, **S**, is reduced by assuming that the cell strives to maximize an empirically defined biomass objective function, **μ(v)**, subject to additional flux constraints, $\boldsymbol{v_{min}} \leq \boldsymbol{v} \leq \boldsymbol{v_{max}}$. Typically, FBA is applied to systems in a steady state; this poses a problem for modeling anaerobic gut fungi because no continuous reactor has been developed for them yet.

Dynamic flux balance analysis (dFBA) is a well-established tool used to extend FBA to dynamic settings (Varma and Palsson, 1994). It relies on the assumption that intra-cellular dynamics are much faster than extra-cellular dynamics. This allows one to discretize time and apply classical FBA at each time step. The resultant fluxes are then used to update the biomass (X), external substrate (**s**), and product (**p**), concentrations by integrating

$$\begin{aligned} \frac{dX}{dt} &= \mu X \\ \frac{d\boldsymbol{s}}{dt} &= \boldsymbol{v_s} X \\ \frac{d\boldsymbol{p}}{dt} &= \boldsymbol{v_p} X \end{aligned} \tag{5.1}$$

where μ, $\mathbf{v_s}$ and $\mathbf{v_p}$ are the growth rate, substrate and product fluxes, respectively. These are then used to update the flux constraints,

$$\boldsymbol{v_{min}}(\boldsymbol{s}, \boldsymbol{p}) \leq \boldsymbol{v} \leq \boldsymbol{v_{max}}(\boldsymbol{s}, \boldsymbol{p}), \tag{5.2}$$

used in the FBA algorithm for the next time step (Henson and Hanly, 2014). dFBA has been successfully applied to mono-culture (Mahadevan, Edwards and Doyle, 2002; Hjersted and

Henson, 2009) and community (Hanly and Henson, 2011; Hanly, Urello and Henson, 2012) modeling.

An inherent weakness of FBA, and by extension dFBA, is the non-uniqueness of the fluxes that maximize the cellular growth rate (Mahadevan and Schilling, 2003). Sampling from the space of optimal fluxes is feasible for FBA applications because the computational cost is paid only once (typically a mixed integer linear program needs to be solved (Saa and Nielsen, 2016)). For dFBA applications, this is prohibitively expensive due to the iterative nature of the algorithm. However, it is well recognized that non-uniqueness of the fluxes can pose problems when integrating Equation (5.1).

Techniques developed to deal with this problem typically involve hierarchal optimization, subsequent to the biomass maximization, to constrain the fluxes further. One possibility is to maximize the growth rate and then sequentially optimize each external flux using the previous optimization problem as a constraint in the current one (Höffner, Harwood and Barton, 2013; Gomez, Höffner and Barton, 2014). This method effectively deals with the non-uniqueness problem but requires additional assumptions per external flux. These assumptions can dramatically affect the results of the simulation but seem to be a problem only when modeling multiple species (Gomez, Höffner and Barton, 2014).

An alternative method is to perform only a single secondary optimization subsequent to the biomass maximization, in the hope that this constrains the fluxes sufficiently to ameliorate the non-uniqueness issue when performing the integration of Equation (5.1). An example of this approach is to minimize the absolute fluxes, based on the principle of maximum enzyme efficiency (Sánchez, Pérez-Correa and Agosin, 2014). The drawback with this approach is

that it requires the solution of a quadratic program (QP) at each time step. For larger models, this can be computationally expensive.

We chose to keep the imposition of additional assumptions on the modeled systems to a minimum because the work is exploratory in nature. Therefore, we combine aspects of (Gomez, Höffner and Barton, 2014) with the single secondary optimization approach. In our case, the secondary optimization seeks to ensure that the derivative change of each modeled flux is minimized between each time step. The rationale for this is that over small time steps the flux is unlikely to jump suddenly. Therefore, at each time step, the following procedure is followed:

1. The flux bounds, Equation (5.2), are updated. Typically, Michaelis–Menten kinetics are assumed (Hanly and Henson, 2013). Since detailed expression for glucose and xylose uptake rates are not known for all the organisms, we assumed, for comparative fairness,

$$\begin{aligned} v_{min,glucose} &= \max\left(v_{Glc}^{max}, -\frac{G + \Delta t f_G^{produced}}{\Delta t X m_{glucose}}\right) \\ v_{max,glucose} &= 0 \\ v_{min,xylose} &= \max\left(v_{Xyl}^{max}, -\frac{Z + \Delta t f_Z^{produced}}{\Delta t X m_{xylose}} \frac{1}{1 + \frac{G}{0.005}}\right) \\ v_{max,xylose} &= 0 \end{aligned} \tag{5.3}$$

where $f_G^{produced}$, $f_Z^{produced}$ are the fluxes of glucose and xylose produced by the extracellular enzymes, *G*, *Z* are the current concentrations of glucose and xylose, and $m_{glucose}$, m_{xylose} are the respective molar masses. The glucose inhibition term ensures that glucose is preferentially metabolized before xylose (Hanly and Henson, 2011). The maximum flux constants, v_{Glc}^{max}

and v_{Xyl}^{max}, were taken from literature and are supplied in Section 5.2.5. See the Supplement for motivation of the derivation of Equation (5.3).

2. A linear program feasibility problem,

$$\min_{s_1,s_2} \sum_{i=1}^{N} s_{1,i} + s_{2,i} \ (where\ N\ is\ the\ number\ of\ fluxes)$$

$$\text{s.t.}\ \boldsymbol{Sv} + \boldsymbol{s_1} - \boldsymbol{s_2} = \boldsymbol{b}\ (where\ b\ is\ typically\ the\ zero\ vector) \tag{5.4}$$

$$\boldsymbol{v_{min}} \leq \boldsymbol{v} \leq \boldsymbol{v_{max}}$$

$$0 \leq s_{1,i}, s_{2,i} \text{ for all } i \in [1, \ldots, N]$$

is solved to ensure that the genome-scale model is feasible for steps 3 and 4. This problem is solved for the "relaxation variables" $\mathbf{s_1}$ and $\mathbf{s_2}$ (see (Höffner, Harwood and Barton, 2013) for justification).

3. A standard FBA linear program (LP) is solved to determine the optimal growth rate of the organism given the constraints of step 1. This problem,

$$\max_{v} \mu(\boldsymbol{v})$$

$$\text{s.t.}\ \boldsymbol{Sv} + \boldsymbol{s_1} - \boldsymbol{s_2} = \boldsymbol{b} \tag{5.5}$$

$$\boldsymbol{v_{min}} \leq \boldsymbol{v} \leq \boldsymbol{v_{max}}$$

is solved for the unique optimal growth rate μ^*. Given μ^* from Equation (5.5), it is possible to solve for the organism biomass concentration by using $\frac{dX}{dt} = \mu^* X$ for at least one time step into the future.

4. A secondary LP,

$$\min_{v} \sum_{i} \gamma_i \text{ for } i \in M \tag{5.6}$$

$$s.t.\ \boldsymbol{Sv} + \boldsymbol{s_1} - \boldsymbol{s_2} = \boldsymbol{b}$$

$$\boldsymbol{v_{min}} \leq \boldsymbol{v} \leq \boldsymbol{v_{max}}$$

$$\mu(\boldsymbol{v}) = \mu^*$$

$$-\gamma_i \leq 1 - \frac{v_{t-1,i}}{v_{t-1,i} - v_{t-2,i}} - \frac{v_{t,i}}{v_{t-1,i} - v_{t-2,i}} \leq \gamma_i \text{ for } i \in M$$

is solved to ensure that the resultant fluxes used to integrate Equation (5.1) are sufficiently smooth. Here, *M* is the index set of all modeled substrates and products. A full derivation of Equation (5.6) is given in the supplementary material of the paper upon which this chapter is based. Briefly, the objective function asserts that $\sum_i \left|1 - \frac{dv_i}{dt}_t \Big/ \frac{dv_i}{dt}_{t-1}\right| \ \forall i \in M$ is minimized, where the flux derivative at time *t*, $\frac{dv_i}{dt}_t$, is approximated to first order.

5. Using an integration scheme of choice, e.g., backward Euler, the full dynamic profile of the system may be iteratively simulated. If products are being generated at each time step, Equation (5.1) needs to include those fluxes as well.

The primary benefit of Equation (5.6) is that there is only a single secondary LP imposed on the system. From a computational point of view, this is very desirable compared to the other existing algorithms that solve either a QP or multiple sequential LPs.

5.2.5 Simulation Parameters

All simulations restricted the oxygen flux into the system to zero. It was assumed that the gas produced by the fungi is 90% carbon dioxide and 10% hydrogen on a mole basis. This is in line with previous experimental observations. The maximum glucose and xylose uptake

flux constraints, shown in Equation (5.3), were taken from the papers introducing the models (see Table 5.1 for the references). These are summarized in Table 5.2.

Table 5.2: Glucose and xylose maximum uptake rates.

Organism	v_{Glc} [mmol/g_{DW}/h]	v_{Xyl} [mmol/g_{DW}/h]
Clostridium ljungdahlii DSM 13528	5	5
Escherichia coli str. K-12 substr. MG1655	10.5	6
Escherichia coli str. ZSC113	0	6
Lactococcus lactis subsp. *cremoris* MG1363	14.5	0
Methanosarcina barkeri str. *Fusaro*	0	0
Saccharomyces cerevisiae S288C	6.44	0

Note that *M. barkeri* does not consume glucose or xylose. Instead, it autotrophically metabolizes hydrogen and carbon dioxide into methane. The maximum hydrogen uptake rate was set at v_{H2}=41.5[mmol/g_{DW}/h], and the maximum carbon dioxide uptake rate was unbounded (Feist *et al.*, 2006). All products P produced by the fungi, e.g., sugar and gas (in the form of pressure accumulation) were assumed to follow the logistic function,

$$P(t) = \frac{k_1}{1 + e^{-k_2(t-k_3)}} \tag{5.7}$$

where the constants were fit to experimental data. Henry's law was used to model the concentration of dissolved gases (hydrogen, carbon dioxide and methane) in the liquid fraction given the gas pressure. A backward Euler scheme was used to integrate Equation (5.1) with a time step of 0.1 h. The initial conditions for all the substrates and products consumed and

produced by the partner microbes were assumed to be zero. The initial biomass concentration was assumed to be 1 mg/L.

5.3 Results and Discussion

Both experimental and computational data were gathered to evaluate the organisms listed in Table 5.1 for their ability to form stable consortia with anaerobic gut fungi. Batch growth experiments were used to model the rate of sugar release from the raw plant biomass during fungal digestion, as well as the gas accumulation profile. This sheds light on the ability of the fungi to accommodate another organism, likely through nutritional linkage of primary metabolites. Computational experiments were then used to predict growth rates and waste generation of a model partner microbe, given the excess fermentable products determined via the batch experiments.

5.3.1 Anaerobic Fungi Release an Assortment of Products to Enable Consortia Formation

Figure 5.1 shows the experimentally observed sugar release and gas production profiles over time of the three anaerobic fungi we investigated. It can be seen that *A. robustus* produced the highest concentration of soluble sugars and the next to highest accumulated pressure. In accordance with the variance between culture replicates, *N. californiae* displayed more erratic growth. This behavior is uncharacteristic of the fungus when cultured in complex media. We speculate that the M2 defined minimal media was a contributing factor to this phenomenon. *Neocallimastix sp. S1* performed between the other two fungi in terms of stability and sugar/gas production.

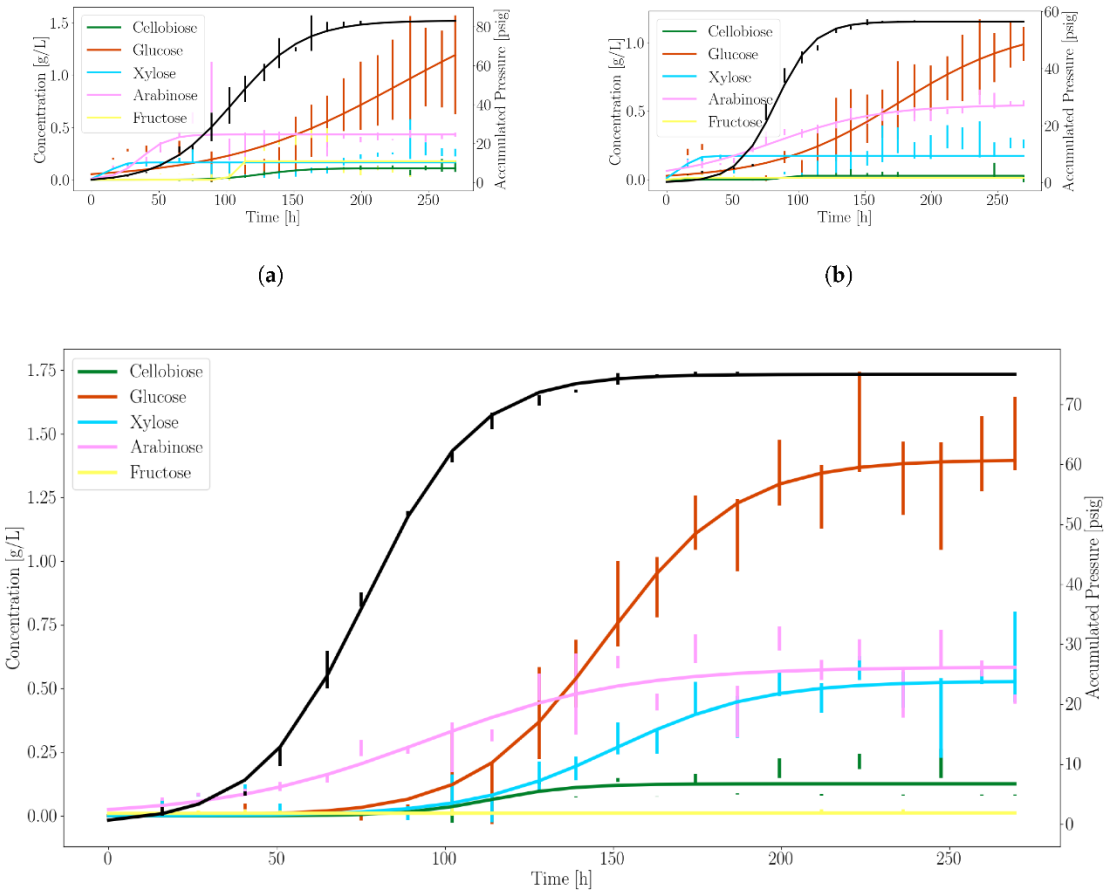

Figure 5.1: Anaerobic gut fungi release excess sugars for microbial partnership during growth on Corn Stover. The solid black line denotes the profile of the accumulated pressure. Other colors represent distinct fermentable sugars generated during growth, as indicated. The vertical bars are standard deviations of errors for each triplicate measurement. (a) *N. californiae*; (b) *Neocallimastix sp. S1*; (c) *A. robustus*. Figure taken from (Wilken *et al.*, 2018).

Based on these data, we selected *A. robustus* as the best candidate for consortia experiments that combine anaerobic fungi with model microbes due to the more stable sugar and gas production rates. Constants used to model substrate production rates for glucose, xylose and pressure accumulation were fit to Equation (5.7) using *A. robustus* data, as shown in Table 5.3.

Table 5.3: Glucose, xylose and gas production rate constants fit to Equation (5.7) for *A. robustus*

Product	k_1 (g/L/h or psi/h)	k_2 (1/h)	k_3 (h)
Glucose	1.39	0.05	148.17
Xylose	0.53	0.05	150.41
Pressure	75.04	0.06	76.51

For completeness, we compare the measured gut fungal net specific growth rates found in M2 defined media, used here, with that of complex media (see Table 5.4). Predictably, the growth rates are lower in minimal defined media. A. robustus consistently outperforms the other fungi when grown on corn stover. The superior growth characteristics of *A. robustus* further motivate its selection as the gut fungus to investigate in greater depth.

Table 5.4: Average anaerobic gut fungal growth rates in defined media compared to rich media.

Organism	Growth rate in M2 (1/h)	Growth rate in MC (1/h)
N. californiae	0.029	0.046
A. robustus	0.033	0.065
N. sp. S1	0.027	No data

5.3.2 Dynamic Simulations Predict Consortia Partner Feasibility

By making use of the dFBA algorithm introduced in Section 5.2.4, and using the experimental data of *A. robustus* to fit Equation (5.7) for both glucose and xylose separately, we can simulate the growth of the co-cultured partner organisms listed in Table 5.1 dynamically. We chose to focus only on glucose and xylose utilization at this stage of

modeling because more is known about the relative preference of each sugar in microbial metabolism (Eiteman, Lee and Altman, 2008). The two classes, fermentable sugar consuming heterotrophs, and hydrogen/carbon dioxide consuming autotrophs, of possible consortia partners were treated separately.

5.3.2.1. Heterotroph Partnership with Anaerobic Fungi

As suggested by Equation (5.3), we assumed, for simplicity, that only glucose and xylose are capable of being fermented by each organism under analysis. Furthermore, we assumed that glucose would be consumed preferentially to xylose whenever possible. Figure 5.2 illustrates the output of the dFBA algorithm when pairing the anaerobic bacterium *C. ljungdahlii* with the gut fungus *A. robustus*. Similar results are available for the other organisms of Table 5.1 in the supplementary material of the paper upon which this chapter is based.

C. ljungdahlii can metabolize both glucose and xylose; this is reflected in the sequential utilization of the substrates in the simulated time course. To determine the effective average growth rate, we fit $\mathrm{X(t)} = c_1 e^{\mu t}$ to the simulated biomass output. The fit indicated that $\mu \approx 0.08$ 1/h. The growth rate is the primary criterion we used to determine suitability for consortia with the gut fungi. We hypothesized that an optimal pairing would occur if the growth rates of the organisms are similar. This would reduce the risk of them out-competing each other. Inter-cellular communication, another pivotal component of consortia, is neglected at this stage of analysis, as it requires detailed experimental data to model.

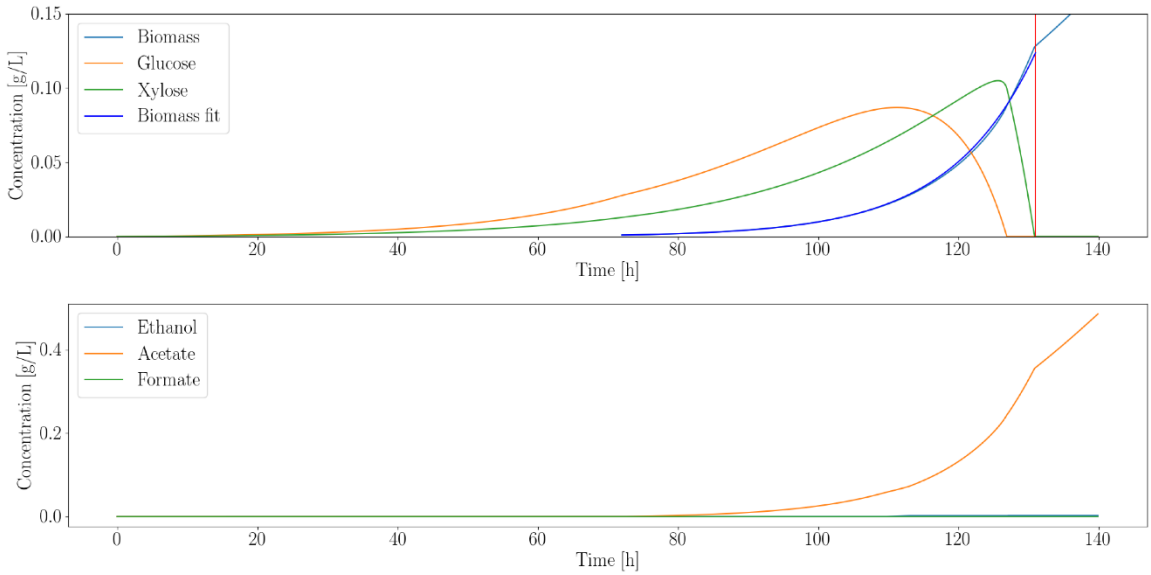

Figure 5.2: Dynamic simulation of *C. ljungdahlii* shows that it consumes all the excess sugars released by *A. robustus*. The vertical red line indicates the point where both sugars were depleted. Even though the fungal enzymes continuously release sugars, the rate at which they release them is exactly equal to the consumption rate beyond the vertical red line. Simulation artifacts cause the growth to continue linearly beyond this point. All the simulations assume an inoculation time at 72 h into the experiment. This allows the slower-growing gut fungi to establish themselves and produce fermentable products prior to the start of the co-culture. Figure taken from (Wilken *et al.*, 2018).

Each modeled organism is also capable of producing metabolic by-products, e.g., ethanol, acetate and formate, that are known to inhibit microbial growth. We also recorded the final concentration of each compound as a secondary criterion to ascertain compatibility with the fungi. The summarized characteristics of each organism, simulated to pair with *A. robustus*, are shown in Table 5.5.

Table 5.5: Growth rate and end point metabolic by-product concentrations produced by each partner microbe assuming inoculation after 72 hours of fungal growth. The end point concentrations are taken when the fermentable substrates were depleted for each organism.

Organism	μ (1/h)	Ethanol (g/L)	Acetate (g/L)	Formate (g/L)
C. ljungdahlii	0.08	0	0.35	0
E. coli MG1655	0.17	0.02	0.02	0.03

E. coli ZSC113	0.04	0.01	0.02	0.03
L. lactis	0.04	0.13	0.32	0.51
S. cerevisiae	0.12	0.02	0	0

The models predicted that both *S. cerevisiae* and *E. coli* MG1655 have a significantly higher growth rate than *A. robustus*. This suggests that maintaining population stability could be difficult for these co-cultures if paired with anaerobic fungi (Goers, Freemont and Polizzi, 2014). While *L. lactis* has a comparable growth rate to *A. robustus*, it is unable to metabolize xylose; therefore, it would directly compete for glucose. Additionally, *L. lactis* produces a wide spectrum of metabolic by-products (ethanol, acetate and formate) at relatively high concentrations; this lessens its attractiveness as a consortia partner. The glucose deficient *E. coli* strain ZSC113 also has a comparable growth rate but produces less metabolic waste products. Additionally, it is genetically amenable to engineering (Bokinsky *et al.*, 2011); this suggests that it could be a favorable organism for consortia formation. Finally, *C. ljungdahlii* is also a competitive choice for consortia. While its growth rate is higher than *A. robustus*, it is not in the range of *S. cerevisiae* and *E. coli* MG1655. C. ljungdahlii can ferment a wide range of sugars as well as autotrophically consume hydrogen (not modeled); this suggests that the organism can take full advantage of the fungal products. Recently, genetic engineering tools have become available for *C. ljungdahlii*, further increasing its viability as a consortia partner.

5.3.2.2 Autotroph Partnership with Anaerobic Fungi

While the organisms shown in Section 5.3.2.1 utilized the fermentable sugars released by the gut fungal enzymes as their carbon source (or preferred carbon source in the case of *C. ljungdahlii*), *M. barkeri*, a methanogen, metabolizes carbon dioxide and hydrogen. It is well known that methanogens are natural consortia partners of gut fungi due to their symbiotic relationship (Haitjema *et al.*, 2014). Methanogens consume the hydrogen gas, a likely growth inhibitor, produced by an intracellular organelle of the fungi called the hydrogenosome (Muller, 1993). Furthermore, it has been shown that methanogens co-cultured with gut fungi significantly increase their cellulolytic efficiency (Marvin-Sikkema *et al.*, 1993).

Figure 5.3 illustrates the simulated growth profile of *M. barkeri*. Negligible quantities of ethanol, acetate and formate are produced, while hydrogen is almost completely consumed. The effective growth rate is 0.03 1/h. Since the gas produced by the fungi drive their growth, it is not surprising that their growth rates are similar.

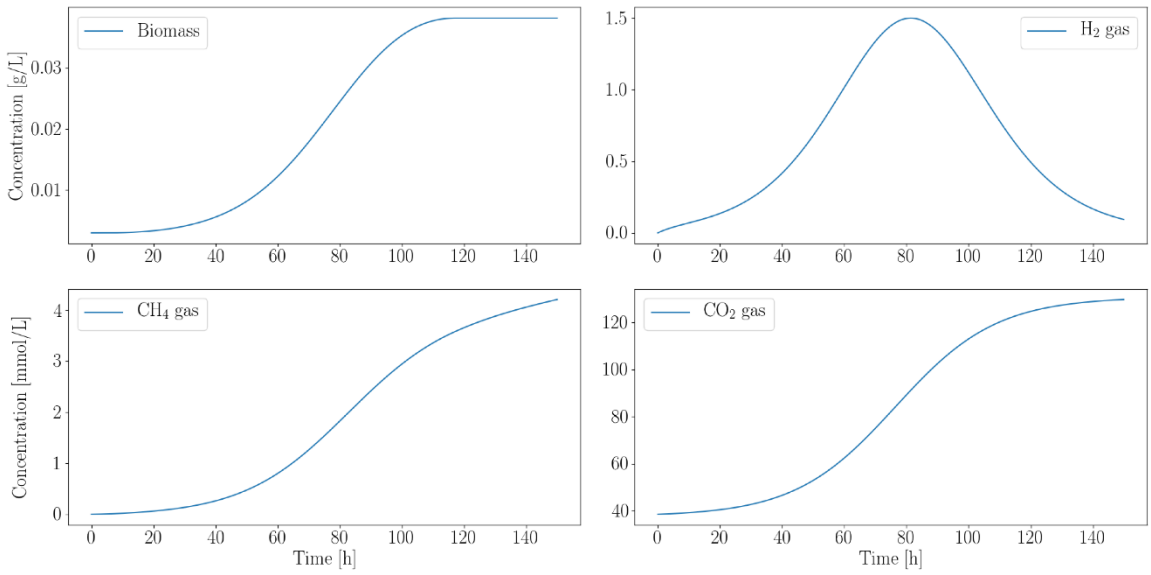

Figure 5.3: Computationally predicted growth profile of *M. barkeri* biomass accumulation over time shows a strong dependence on the fungal metabolic by-products. Hydrogen and carbon dioxide,

produced by the fungi, are consumed by the methanogen. Simultaneous inoculation is assumed because the microbes do not compete for their preferred carbon source. All gas concentrations are in mmol/L. Figure taken from (Wilken *et al.*, 2018).

M. barkeri is also an attractive candidate for synthetic gut fungal consortia due to the mutualism exhibited by the pairing of fungi and methanogens in nature (Peng, Gilmore and O'Malley, 2016). The recent development of genetic technology to manipulate Methanosarcina suggests that the pairing is also feasible for bioproduction (Kohler and Metcalf, 2012). Finally, given the low levels of by-products generated by *M. barkeri*, it is plausible to consider tri-cultures of *A. robustus*, *M. barkeri* and another microbe, like *C. ljungdahlii*. Such a system would be, theoretically, minimally negatively interactive due to the reduced substrate competition. This is a desirable property for community stability.

The benefit of using the dFBA, to screen for consortia partners, is that it is readily generalizable to higher order systems. Known interactions can easily be accounted for, and quantitative predictions of by-product generation can be used to evaluate partner suitability (cf. qualitative literature surveys). The simulation approach is particularly useful for non-model organisms, like anaerobic fungi, because growth rate predictions in their unique culture conditions are not often readily available.

Experimental validation of these predictions will take the form of community composition tracking and by-product generation monitoring. The latter technique is particularly applicable to the anaerobic fungi because it is one of the few non-invasive methods that can be used to measure growth in gut fungal systems (Theodorou *et al.*, 1995). For example, in the case of the *A. robustus* and *M. barkeri* pairing, the methane, carbon dioxide and hydrogen production over time, compared to the mono-cultures, will indicate the success of the co-culture. Similar

indirect measurements could be used to validate the other predictions. However, these detailed experiments are beyond the scope of the current work.

5.4 Conclusions

To assess the suitability of each organism in Table 5.1 to form stable microbial consortia with anaerobic fungi, the identities and contributions of both the gut fungus and partner microbe need to be justified. In this work, experiments were used to select an anaerobic fungus and simulations, making the least number of assumptions, were used to screen possible consortia partners.

The experimental results of Section 5.3.1 indicate that *A. robustus* is a more desirable building block for consortia (or even mono-cultures) compared to other strains tested here—both in terms of higher growth rates on Corn Stover (see Table 5.4) as well as enzyme effectiveness at releasing fermentable sugars (see Figure 5.1). Barring the generation of unknown inhibitory agents, it should be prioritized for further experimentation.

M. barkeri, a methanogen, is a natural consortia partner for gut fungi (Marvin-Sikkema *et al.*, 1993). This is clear from the similar growth rates to *A. robustus* and consumption of hydrogen, a known inhibitor of fungal growth. Additionally, it produces minimal by-products that could retard fungal growth. *C. ljungdahlii* and *E. coli* ZSC113 are also potentially suitable consortia partners. On the other hand, *L. lactis*, *S. cerevisiae* and *E. coli* MG1655 were all ruled out due to their by-product generation or significantly higher growth rates. We introduced a novel dFBA algorithm that is computationally efficient and that does not impose many extra assumptions on the system. Making use of computational tools, such as this, to

reduce the number of costly and time-consuming experiments is a boon to developing and designing scalable synthetic biosystems (Höffner and Barton, 2014).

Moreover, building predictive models of consortia systems can be critical to fully leveraging the inherent capabilities of micro-organisms as it allows engineers additional insight into the mechanics of these complex systems (Mahadevan and Henson, 2012). Fully unlocking the inherent capabilities of non-model organisms, like anaerobic gut fungi, will require novel tools to inexpensively generate and test hypotheses. Current consortia analysis techniques typically assume that the identities of the partner microbes are known and that they are modeled. This work provides a framework that can be used to rationally select the them even if some of the microbes are not modeled.

5.5 Acknowledgments

The authors gratefully acknowledge funding support from the Office of Science (BER), U.S. Department of Energy (DE-SC0010352), the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office, the National Science Foundation (MCB-1553721), and the Dow Discovery Fellowship (to SW). The authors also thank John K. Henske for isolation and ITS characterization of *Neocallimastix sp. S1*, and Paul Weimer from the United States Department of Agriculture (USDA) for providing freshly milled biomass substrates

VI. An Arduino based automatic pressure evaluation system (A-APES) to quantify growth of non-model anaerobes in culture

This chapter is based upon work that is published in the AIChe Journal, Volume 7, 2020, by St. Elmo Wilken, Patrick Leggieri, Corey Kerdman-Andrade, Matthew Reilly, Michael K. Theodorou and Michelle A. O'Malley, entitled "*An Arduino based Automatic Pressure Evaluation System (A-APES) to quantify growth of non-model anaerobes in culture*", Copyright John Wiley and Sons. See the publication for more detailed information regarding the construction and functional tests, as well as the supplementary information mentioned in this chapter.

6.1 Introduction

Cultivation techniques applied to model microbes in biotechnology, like *Escherichia coli* and *Saccharomyces cerevisiae*, are well established, with many commercial tools available to automate data collection and analysis (Junker *et al.*, 1994; Bareither and Pollard, 2011). Moreover, because model microbes are relatively simple to cultivate, and are well-suspended in batch or continuous culture, many lab-scale "do-it-yourself" devices have been constructed to facilitate high throughput, automated experiments that make use of optical density measurements and continuous recording of select metabolites (Boccazzi *et al.*, 2005; Groisman *et al.*, 2005; Klein, Schneider and Heinzle, 2013; Bergenholm *et al.*, 2019) to monitor microbial growth. However, non-model microbes often present unique difficulties that hamper direct application of these technologies and techniques, often necessitating time consuming and/or destructive manual measurements. For example, many such microbes have

complex morphologies, are surface-adherent, and/or feature a complex life cycle (Podolsky *et al.*, 2019).

Anaerobic gut fungi, in the phylum Neocallimastigomycota, are relatively understudied non-model organisms of high biotechnological value due to their vast array of carbohydrate active enzymes (Youssef *et al.*, 2013; Solomon *et al.*, 2016; Haitjema *et al.*, 2017b). However, anaerobic fungi have proven exceptionally difficult to characterize in large part due to challenges in their cultivation. They are strict anaerobes, temperature sensitive, filamentous and typically require specialized media for growth (Haitjema *et al.*, 2014). Further, in contrast to model yeasts or fungi, anaerobic gut fungi are not well suited to cultivation in chemostats because they adhere to their growth substrates, and themselves, through a filamentous rhizoid network (Gruninger *et al.*, 2014). This necessitates either destructive harvesting of samples to benchmark cellular biomass or the use of indirect measurements to permit growth rate calculations.

Indirect measurements for anaerobes typically make use of accumulated pressure of fermentation products as a proxy for growth, and have been widely adopted in the field (Theodorou *et al.*, 1995; Haitjema *et al.*, 2014). For example, for anaerobic gut fungi, gas production rate growth curves are often used to study fungal lignocellulolytic properties and substrate preferences, yet are typically labor and time intensive to generate when fine resolution is required (O'Malley, Theodorou and Kaiser, 2012c; Henske, Wilken, *et al.*, 2018). Typically, the fermentation gas pressure in each sample under consideration must be measured and vented multiple times per day to obtain an accurate estimate of the fungal growth rate. The time intensive nature of measuring accumulated pressure in such cultures has led to the design and construction of devices that automate this process (Davies *et al.*, 2000; Adesogan,

Krueger and Kim, 2005). In essence, these approaches typically combine a pressure transducer with a valve. The transducer measures the accumulated pressure over the course of growth, and the valve vents the closed system to prevent over-pressurization periodically, as shown schematically in Figure 6.1. Alternative designs include liquid displacement flow-meters, but accurate readings can be challenging to attain using such devices (Walker *et al.*, 2009).

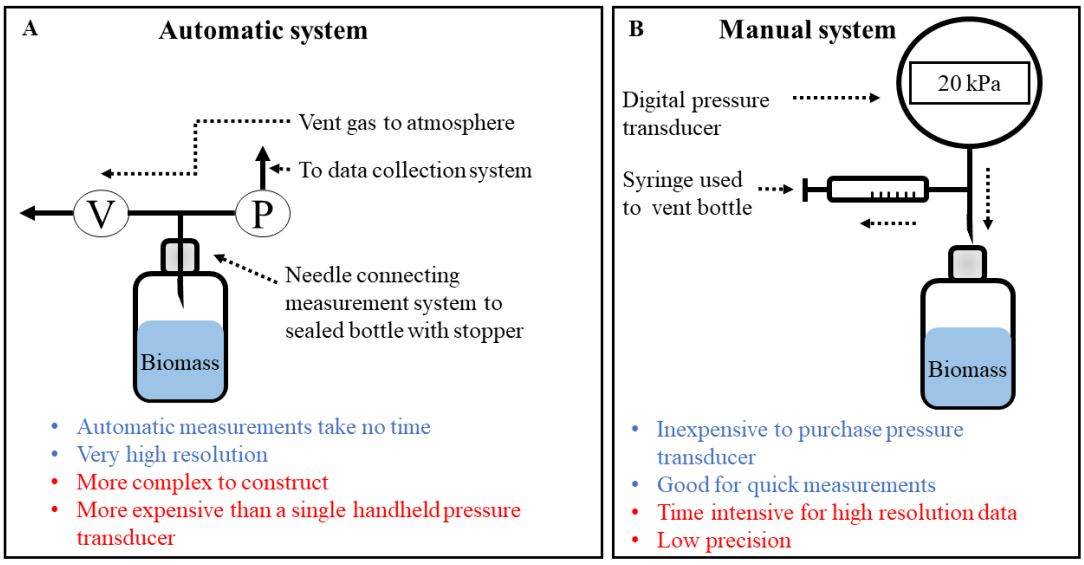

Figure 6.1: Conceptual design of automatic pressure measurement and venting devices(Davies *et al.*, 2000) compared to labor intensive manual measurements. Benefits of each system are shown in blue font, with drawbacks in red. (A) Designs typically make use of a pressure transducer (P) that measures the rate of pressure increase in a sealed bottle, which is correlated to growth in rumen microbiome based systems (Theodorou *et al.*, 1995; Haitjema *et al.*, 2014). To prevent over-pressurization of the sealed bottles a valve (V) can be used to vent the system. (B) Manually measuring and venting the pressure requires the use of a handheld pressure transducer that is used to measure the pressure in the bottle prior to venting. Slight cooling of the bottles is usually observed due to the time it takes to vent the culture outside of an incubator.

Despite the apparent simplicity of the design shown in Figure 6.1.A, these lab-built automated systems have not gained significant traction. This is likely because the electronics required to make these systems work are not simple or readily shareable. Relatively expensive commercial systems, such as the Ankom RF Gas Production System or the OxiTop Respirator system, exist and have been used to study the growth characteristics of anaerobic systems (Tagliapietra *et al.*, 2010; Pabón Pereira, Castañares and Van Lier, 2012). On the other hand, Arduino based systems have recently become popular foundations to build lab automation devices of varying complexity (Urban, 2015, 2018). Importantly, Arduino based systems are low cost and relatively simple to build (Sarik and Kymissis, 2010; Grinias *et al.*, 2016). There is also a growing drive to towards developing "open-hardware", which encompasses the development of low cost, easily shareable, standardized lab automation designs (Sarik and Kymissis, 2010; Gibney, 2016).

Here we use a non-model anaerobic gut fungus as a test bed to design and build a device that can be used to automatically record and release pressure to measure microbial growth. This enables the construction of high-quality growth curves for sensitive, strictly anaerobic microorganisms that are not amenable to direct biomass measurements. Specifically, this device measures and logs the rate of gas production and is particularly applicable to systems where the rate of gas production is correlated with biomass growth. The wireless Arduino based Automatic Pressure Evaluation System device introduced here, named A-APES, is specifically designed to work with strictly anaerobic systems, like rumen microbiome-based cultures. In particular, this system is designed to make use of standard lab equipment (serum bottles, incubators etc.) that are routinely used in the field. Use of this device will enable the collection of cross-lab comparable, high quality data without the need for significant manual

oversight. Additionally, due to the use of the Arduino base and modular apparatus, it is straightforward to extend the system to include additional monitoring channels or simultaneously connect with other measurement devices if desired. The aim is to present a low cost, standardized system that can be built in any lab without the need to understand complex electronics. We describe the design of the system, which includes a "ready to be manufactured" printed circuit board (PCB) that minimizes the amount of assembly and technical know-how required to construct the system.

Furthermore, to demonstrate the utility of the A-APES device, several high-resolution growth curves of an isolated anerobic gut fungus were constructed. Experiments were designed to investigate the influence of pressure venting frequency on the growth rate of anaerobic fungi. Additionally, these high-quality growth curves revealed that gut fungi appear to lack a true exponential phase when grown on lignocellulose. Instead, the growth rate appears to be multiphasic, possibly because the polymeric constituents of lignocellulose are not digested at the same rate by the gut fungus. The effect of venting frequency on the growth rate of the cultures was found not to be significant, suggesting that gas accumulation and venting frequency are not key drivers of the observed fungal growth rate. In future, the ability to accurately and continuously infer the growth rate of anaerobic gut fungi in real-time could be used to perform substrate optimization experiments for which current techniques are lacking in measurement frequency, sensitivity and precision.

6.2 Materials and Methods

6.2.1 Design and construction of A-APES

A schematic diagram of the Arduino based Automatic Pressures Evaluation System (A-APES) device is shown in Figure 6.2. The Supplement contains the Gerber file that was used to manufacture the printed circuit board (PCB), as well as other schematic documents that explain how to construct the entire device. Briefly, A-APES uses two XBEE ZIGBEE Mesh (DIGI, MI) devices for wireless communication between A-APES and a computer that logs the data. The XBEEs are plug-and-play, requiring minimal setup through the free software XCTU from DIGI. The first XBEE is connected to the A-APES device; the second XBEE is connected to the data logging computer using an XBEE USB Dongle (WRL-11812, Sparkfun, CO). A short Python script is used to read and save the data from the USB connection (see the supplied code in the Supplement). Copper tubing, which is connected to an all metal syringe sealed with epoxy, is used to connect the solenoid valve (RSSM-2-12V, Electric Solenoid Valves, NY) and the pressure transducer (PX119-030AI, Omega Engineering, CT) to a bottle that is sealed using a 13 mm thick butyl rubber stopper typical for anaerobic experiments. Insulated 18-gauge wires are used to connect the solenoid valves to an independent power supply via a relay switch (Youngneer 5V relay, Amazon, WA). Additional wires (22-gauge) were used to connect the relay, which controls the solenoid valve, as well as the pressure transducer to an Arduino microcontroller (Arduino Uno R3, Amazon, WA) via the PCB, which used a second power supply. A 16-bit analog-to-digital converter (ADC) (1085, Adafruit, NY) is used to translate the transducer's output to a signal that is interpreted through

the Arduino. More detailed information regarding the construction of the device may be found in Supplement (the construction guide, parts list and code).

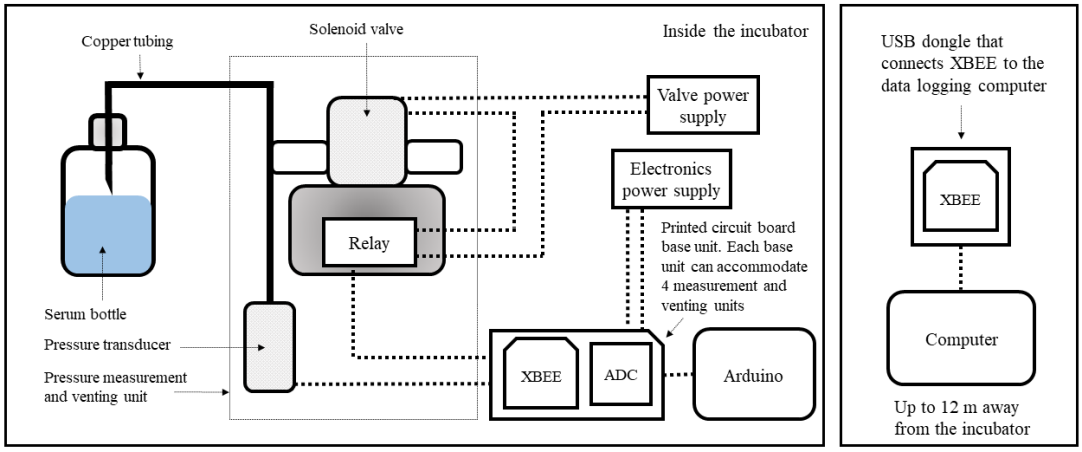

Figure 6.2: A schematic diagram of the primary components of A-APES. In this diagram only a single solenoid valve/pressure transducer unit is shown, but the base system can accommodate up to 4 independent units in total. The construction guide illustrates the assembly process (refer to Supplementary Information).

6.2.2 Tubing and connections leak tests

Prior to the selection of copper tubing for A-APES, various other plastic tubing types were evaluated for their ability to form a gas tight seal between the pressure transducer, the needle and the solenoid valve, as depicted in Figure 6.2. This included Tygon (6516T11, McMaster-Carr, IL), Tygon PVC (8349T12, McMaster-Carr, IL), PFA (EW-06375-01, Cole-Palmer, IL) and CFlex (EW-06424-14, Cole-Palmer, IL) tubing. To test the gas-tightness, each type of tubing was connected to a pressure transducer and left to equilibrate at 39°C in an incubator overnight. Subsequently, a 70 mL serum bottle, half filled with glass beads (2mm diameter,

Chemglass, NJ), was pressurized to approximately 138 kPaa with pure CO_2 gas (representative of the typical operating conditions). This bottle was connected to the transducer and the pressure over time was monitored to ascertain the rate of gas leakage through the tubing. Copper tubing was used in the final design due to its superior gas tight seal, as is discussed later. The entire system was constructed, as shown in the Supplement, and leak tested. This entailed pressurizing three 70 mL serum bottles as before and recording the change in pressure over time.

6.2.3 Experimental evaluation of anaerobic growth

Standard anaerobic gut fungal culturing techniques and conditions were used for all the experiments presented in this work (Haitjema *et al.*, 2014). All experiments used 70 mL (total volume) serum bottles with 0.5 grams of Corn Stover (supplied by the USDA-ARS Research Center, Madison, WI) in 40 mL of MC media (Davies *et al.*, 1993), incubated at 39°C with a 100% CO_2 gas headspace. The filled serum bottles were autoclaved at 121°C for 20 minutes prior to use. An anaerobic gut fungus isolate, *Neocallimastix lanati*, was exclusively used in all the experiments. Each experimental triplicate was inoculated with 2 mL from the same 2-day old serum bottle of growing fungus of the same media composition as the experiment. Additionally, 0.5 mL of 10 mg/mL Chloramphenicol (BP904-100, Fisher Scientific, CA) was added to each bottle to prevent contamination by other microbes. Butyl rubber stoppers were used in all the experiments to ensure a gas tight seal between the serum bottle and the A-APES needle (as described above). Each experiment was run until stationary phase was observed, typically 4-5 days post inoculation. Any deviations from this are noted in the relevant results section. Three independent pressure measurement (transducers) and release valves (solenoids)

were used to enable the measurement of culture growth in a triplicate set of serum bottles. The venting frequency of headspace gas was varied as noted in the results section. Pressure measurements were taken every minute and recorded.

6.2.4 Data analysis

The experimental design resulted in three high resolution pressure measurement datasets per run. The growth rate for each dataset was determined by log transforming the cumulative pressure data and fitting a straight line to time-axis discretized intervals of 12 hours (approximately one doubling time) beginning 20 hours after inoculation. This yielded instantaneous growth rate data over the entire time course as shown in later figures. The 20-hour time offset was used to allow the system to equilibrate post-inoculation. For each replicate, the maximum straight-line slope over all the discretized intervals of the experiment was taken as the maximum growth rate of the dataset. Repeats of runs (each run is a triplicate set) were considered consistent with each other if the p-value of the unequal variance T-test was above 0.05 for over 50% of comparisons between the pressures measured at equivalent time points. The growth rates of different run conditions were also compared using the unequal variance T-test with a cutoff p-value of 0.05. The Julia language was used for all the data analysis and visualization (Bezanson *et al.*, 2017), while Python and C were used to interface the data recording computer with A-APES (code supplied in the Supplement).

6.3. Results and discussion

6.3.1 A-APES is straightforward to construct and is gas tight

Here we introduce an Arduino based Automatic Pressure Evaluation System (A-APES) that can be used to automatically record and vent the pressure in anaerobic cultures. This system allows for the generation of high quality and high-resolution pressure accumulation data that can be used to infer the growth rate of non-model anaerobes in culture. A complete parts list and guide to constructing A-APES is shown in the Supplement. Due to the use of the Arduino base, minimal knowledge of electronics is required to build, modify and operate the system. Moreover, the PCB is designed to reduce the wiring and assembly time required to build the system, which is also relatively inexpensive compared to commercial alternatives. The cost to build the base system, i.e. A-APES with a single pressure measurement and venting unit, is approximately $430 (as of 2020). The cost for a fully equipped base system with 4 independent pressure measurement and venting units is approximately $1000. This equates to a price of $250 per measurement unit, which is 3.2 times cheaper per measurement unit than the equivalent cost of a commercial system. Beyond the cost savings of A-APES, the Arduino base makes the system readily extendible to include other sensors or configurations. Specifically, the high accuracy 16-bit ADC is not restricted to the pressure transducer. Therefore a wide range of commercially available environmental sensors with analogue outputs can also be monitored by the system (Urban, 2018).

Due to limited incubator space and media costs, it is also desirable to minimize the volume of culture vessels used with automated systems. To the best of our knowledge, the smallest operable working volume for a commercially available system is 250 mL. Filling a large bottle

with a relatively small volume of liquid media results in a large headspace volume in the bottle. This larger headspace volume reduces the sensitivity of the measured pressure in the bottle. On the other hand, using more liquid media relative to vessel size results in a smaller head space volume that can exacerbate the effect gas leaks have on the measured pressure. Thus, an important design requirement is that the measurement system is gas tight to accurately measure gas production rates, as well as maintain anaerobicity. A-APES is designed to be gas tight and not constrained to a particular bottle size. For demonstration purposes we used 70 mL total volume glass bottles filled with 40 mL of liquid media. However, it should be noted that the A-APES can potentially be used with a wide range of vessel sizes if they are sealable with butyl-rubber stoppers.

Various tubing types were considered and evaluated during the construction of A-APES, with the goal of identifying the most gas tight configuration. Figure S1 shows that plastic tubing leads to significantly higher gas leak rates, either due to the permeability of CO_2 and/or the barbed connection fittings that were used. Copper tubing was selected because the rate of gas leakage was the lowest (0.01 kPa/h), see Figure S1 in the supplement for details. Since copper is not as flexible as plastic, some strain is placed on the connections when new serum bottles are connected to A-APES. This strain introduces the potential for leaks if the connections are not tight. Sealing the joints with epoxy solves this problem; it was found that the leak rate was halved in the final assembled system when epoxy was used to seal the joints, see Figure S2 in the supplement. However, using epoxy makes the connections permanent – a problem if the system needs to be disassembled and reconfigured. On balance the superior gas tightness ensured by the epoxy was deemed worth the inconvenience of permanent fixtures. The final gas leakage rate for the assembled system is 0.01 kPa/h. Assuming a 5-day

run duration, and 172 kPa of accumulated pressure (typical values recorded), leakage caused an error of less than 1% which we consider to be negligible.

6.3.2 No significant differences were observed between A-APES and manual pressure measurements of anaerobic fungal cultures

Pressure measurement differences between using A-APES and manually measuring and venting culture vessels were investigated by running a side-by-side comparison. It is important that the A-PES system is able to recapitulate pressure accumulation data measured manually because this is the standard in the field and would lend credence to novel observations derived from automatically generated data. To this end, A-APES was programmed to vent a set of triplicate anaerobic fungal cultures every 12 hours, while another set of triplicate cultures were started at the same time, from the same inoculum, and vented manually at the same interval. Figure 6.3.A shows the pressures at each measurement interval, and Figure 6.3.B shows the cumulative pressure profile. In both cases there were no statistically significant differences between the experiments at any point in time, as shown in Figure S3 in the supplement. Furthermore, the automatic experiment had a maximum growth rate of 0.087 ± 0.006 1/h, while the manual experiment had a maximum growth rate of 0.09 ± 0.012 1/h calculated by log transforming data points at the same time and finding the maximum slope for each experiment using these data points. The growth rates were also not statistically significantly different.

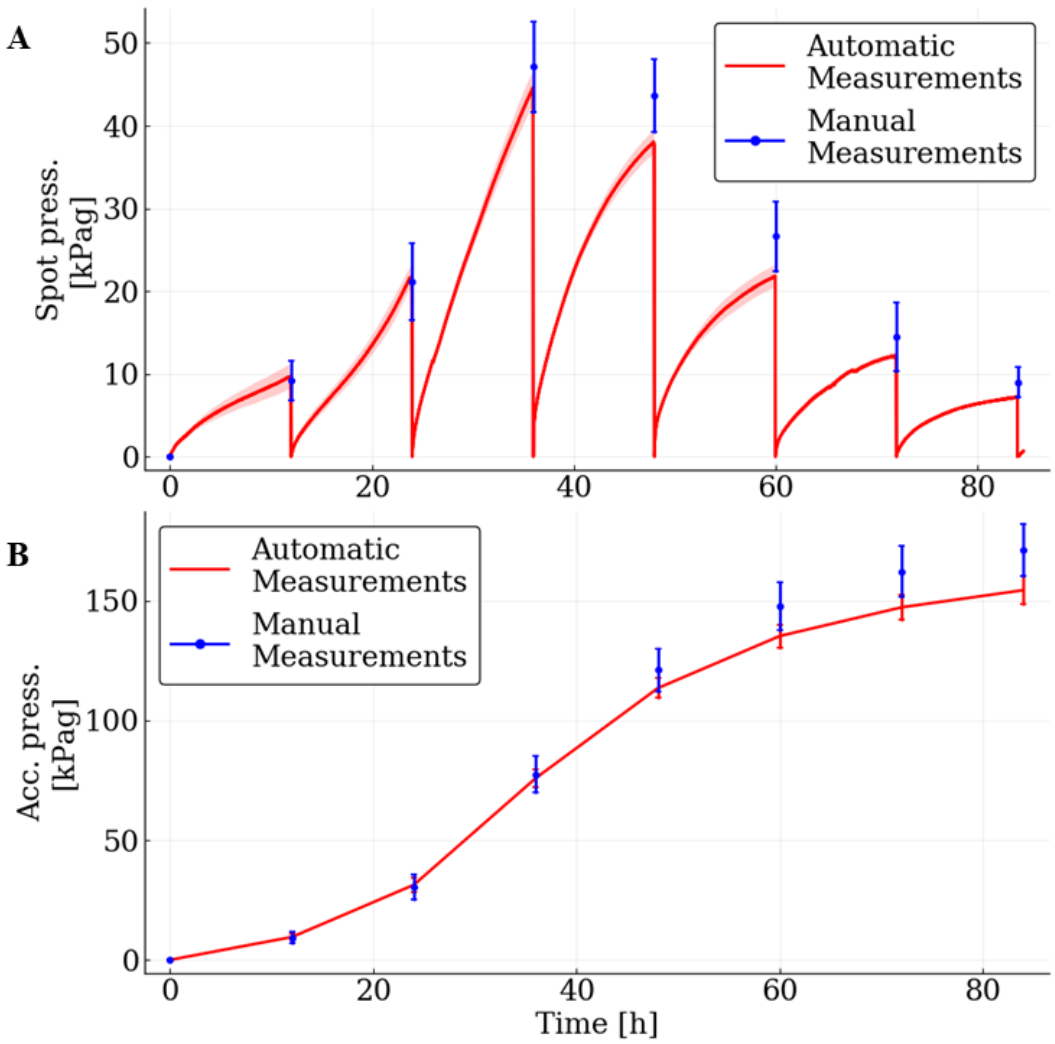

Figure 6.3: No statistically significant differences were found when comparing A-APES pressure measurements to manual pressure measurements of fungal growth. The pressure production measurements of two sets of triplicate *N. lanati* cultures were compared in a side-by-side experiment. Each replicate in both triplicate sets were treated in exactly the same way (2 mL inoculum from the same starter bottle into 40 mL complex media with 0.5 grams of corn stover, see the methods section for more details), except for the measurement method. One set used conventional manual pressure measurements and the other set used A-APES to record the pressure production rate. Both triplicate sets were vented every 12 hours. (A) Spot pressure measurements over time for both sets of triplicates. (B) The accumulated pressure profiles for each case. Neither the spot pressure measurements (Figure

6.3.A), nor the accumulated pressure profile (Figure 6.3.B) was statistically different. The measurement noise was lower using the automatic system (shaded region in Figure 6.3.A represents 1 standard deviation). All error bars represent 1 standard deviation of error from the mean.

It is informative to note some differences between the manually and automatically vented cultures, which were enabled by this comparison. The manually vented cultures cooled down slightly during each measurement bout. While the effect of the temperature fluctuation on growth is likely small when measuring infrequently, it could play a more significant role when smaller test tubes are used instead of individual serum bottles and/or measurements are done more frequently. Additionally, by removing the serum bottles from the incubator some stirring/mixing occurs. This is completely absent from the cultures that were measured using A-APES, as they are never removed, or moved at all, from the incubator. Despite these physical differences, the results suggest that A-APES measures growth rates and pressure profiles with no significant difference to the manual experiment, albeit with reduced manual labor.

6.3.3. A-APES demonstrates high run-to-run consistency

The reproducibility of A-APES was tested by comparing the pressure profiles and growth rates of two runs done at different times using the same venting frequency. Figure 6.4.A shows the measured spot pressures, and Figure 6.4.B shows the cumulative pressure profile over time for both sets of triplicate runs. The cumulative pressure profile is not significantly different over the entire growth curve, while the spot measurements are not significantly different over 89% of the growth curve, see Figure S4 in the supplement. Interestingly, the maximum growth rates were found to be statistically significantly different, irrespective of the time interval used

to calculate, them as shown in Figure S5 in the supplement. The low measurement noise associated with the A-APES system likely makes any experimental or biological noise more noticeable, which gave rise to the significant differences noted in Figure S5.

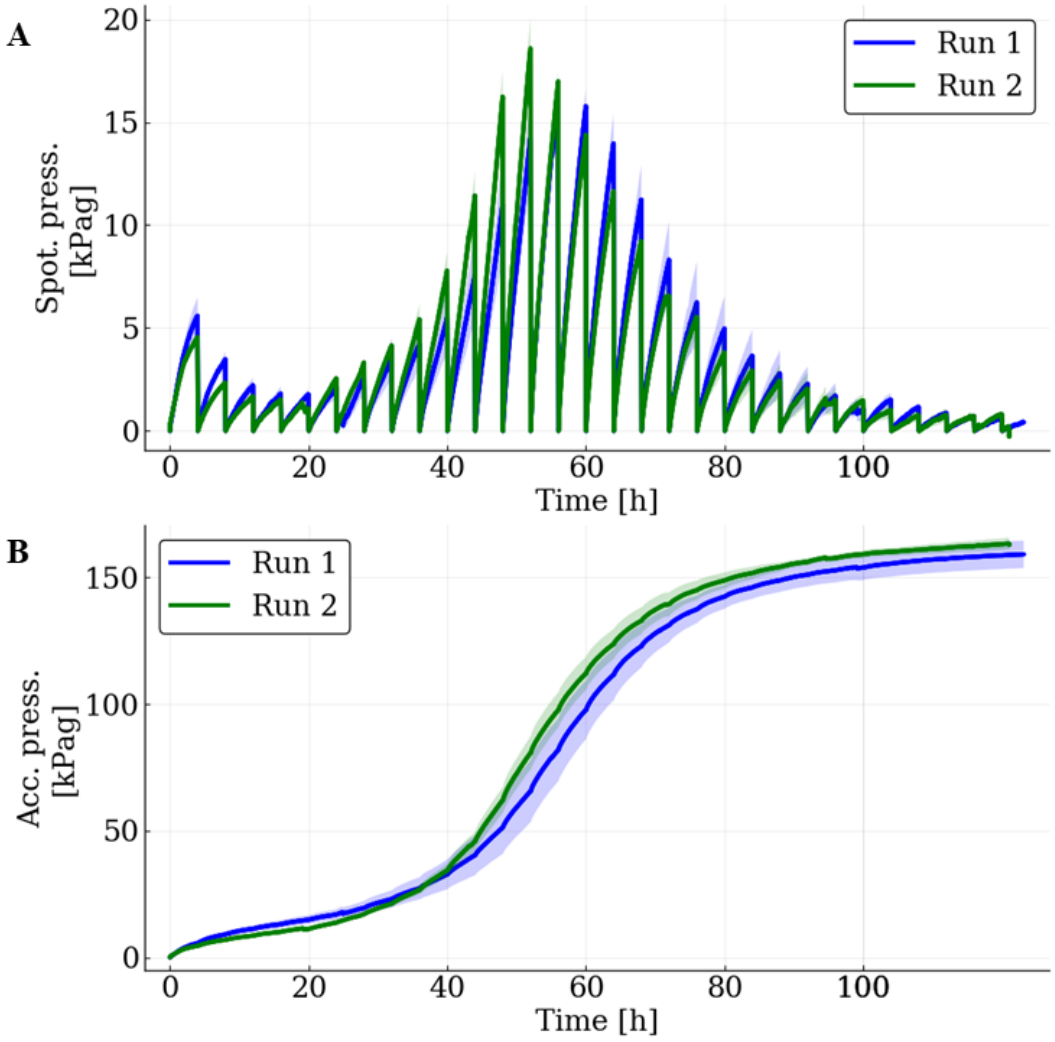

Figure 6.4: A-APES shows high run-to-run measurement consistency with minimal statistically significant differences. Two triplicate experiments (run 1 and run 2, respectively), using exactly the same experimental conditions (2 mL inoculation of *N. lanati*, 40 mL complex media with 0.5 grams of corn stover, venting every 4 hours and recording pressure measurements every minute, see methods section for more details), were run at different times to gauge the reproducibility of pressure measurements using A-APES. (A) The spot pressure measurements for each run. (B) The accumulated

pressure profiles for each run. The shaded area represents 1 standard deviation from the mean curve. The spot pressure measurements (Figure 6.4.A) were not significantly different over 89% of the experimental duration, while the accumulated pressure curves (Figure 6.4.B) were not significantly different over the entire duration of the experiments.

The average difference between the maximum growth rates (as a function of different time discretization) was 0.01 ± 0.002 1/h. A leak test was performed to rule out that a leak in the connections caused the observed differences; this was found not to be the case. Thus, it is likely that these differences have a biological origin, as opposed to indicating problems with A-APES. Indeed, relatively high between run variability has been observed in other gut fungal isolates. For example, the growth rate of *Neocallimastix californiae* has been reported to range from 0.064 ± 0.007 to 0.072 ± 0.002 1/h growing under the same conditions as those used here (Henske, Wilken, *et al.*, 2018; Gilmore *et al.*, 2019). This suggests that there is some inherent biological variation that needs to be accounted for when comparing runs done at different times. Despite these observations, the high similarity in the measured pressure profiles suggest that A-APES is indeed consistent between runs. Furthermore, this result suggests that caution should be exercised when interpreting growth rate differences that are statistically significant yet small (on the order of 0.01 1/h) for this type of organism.

6.3.4 High resolution data yields accurate rate data over the entire growth curve

Manually measured pressure data is typically limited to very few data points, such as measuring and venting an anaerobic culture 3 times per day for 5 days, which results in 15 data points. On the other hand, A-APES can record measurements every minute, yielding much finer resolution that can capture significantly more growth dynamics (~15 vs. ~7200 data points, manual vs. A-APES respectively measured for 5 days). This allows for the

inference of growth rates over the entire time course, with much higher resolution compared to manual methods. Figure 6.5 reveals that the growth rate of *N. lanati*, on a lignocellulosic substrate (corn stover), is variable. In particular, the growth rate seems to plateau for only a short duration (~5 hours), after which it decreases rapidly. By using the high-resolution data afforded by A-APES, it is apparent that classic exponential phase (characterized by a constant maximum growth rate) is absent. Instead a variable growth rate is observed. This information would be obscured by using lower-resolution manual methods. It is possible that the fermentable sugars released during the digestion of the lignocellulose by the fungus are differentially metabolized. This substrate preference could be the cause of the observed variable growth rate. The initially increasing growth rate could be attributed to an excess of easily metabolizable substrates being available, but the enzymes required to unlock them from the lignocellulose first need to be produce, which limits the growth. The harder-to-metabolize substrates are metabolized last, explaining why the growth rate starts to decrease midway through the time course.

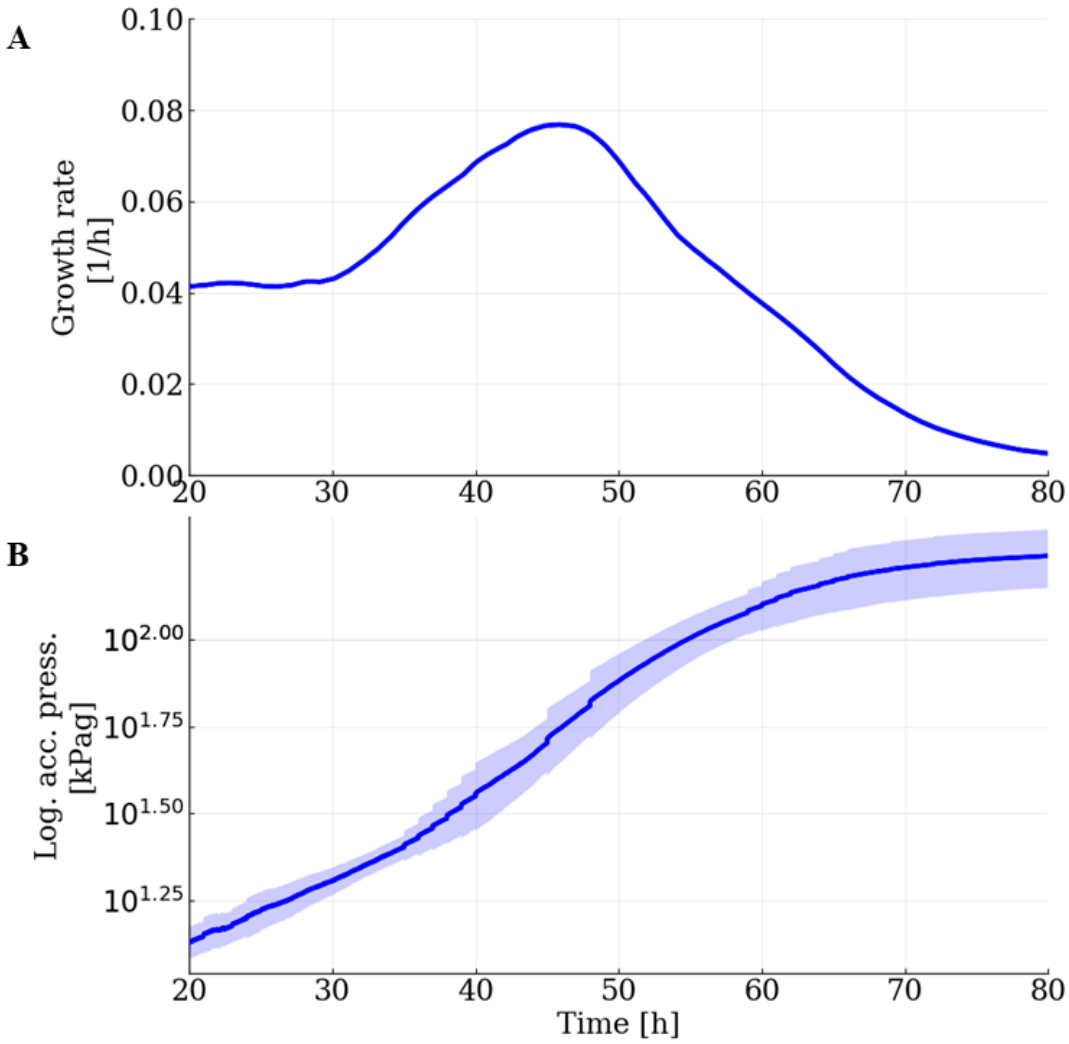

Figure 6.5: High resolution pressure measurements reveal that the growth rate of *N. lanati*, growing on a corn stover, is variable across the growth curve. Pressure was vented every hour, and measurements were taken every minute. Each replicate of the triplicate data shown here was grown in complex media with 0.5 grams of corn stover and inoculated with 2 ml from the same starter bottle, see the methods section for more details. (A) Figure 6.5.A. shows the inferred instantaneous growth rate, calculated over 12-hour intervals, peaks at ~0.08 1/h, but only for a short duration (~5 hours). (B) Figure 6.5.B. shows the corresponding log transformed accumulated pressure curve. In both cases it is apparent that a classic constant rate exponential phase is absent. Differential substrate digestion and

metabolization may explain the variable growth rates. For each figure the shaded region represents 1 standard deviation from the solid blue curve that represents the mean of the measurements

Alternatively, it has been suggested that hydrogen production and accumulation inhibits the gut fungal energy metabolism (Marvin-Sikkema *et al.*, 1994b; Gruninger *et al.*, 2014). To investigate this using A-APES, the venting frequency was varied (every 1, 4, and 12 hours in triplicate), and the growth rates were compared. By venting more frequently, the partial pressure of hydrogen would be reduced, differentially attenuating possible inhibition effects. However, as shown in Figure 6.6, it seems unlikely that this type of inhibition plays an important role in the observed growth rate decrease. Across all three conditions the growth rate profiles were similar and the observed maximum growth rates were approximately similar (~0.08 1/h, within the 0.01 1/h margin noted earlier). This suggests that pressure accumulation, and by extension hydrogen accumulation, does not significantly reduce the growth rate of *N. lanati*. While the reason for this observed growth rate decrease in anaerobic fungi remains unclear, the data suggest there is significant scope to experiment with conditions that optimize growth and to engineer anaerobic gut fungi to grow at their maximum rate for a longer time duration. In sum, the benefit of using A-APES is apparent here: very high-resolution data is available to interrogate the effect of experimental perturbations on sensitive anaerobic systems.

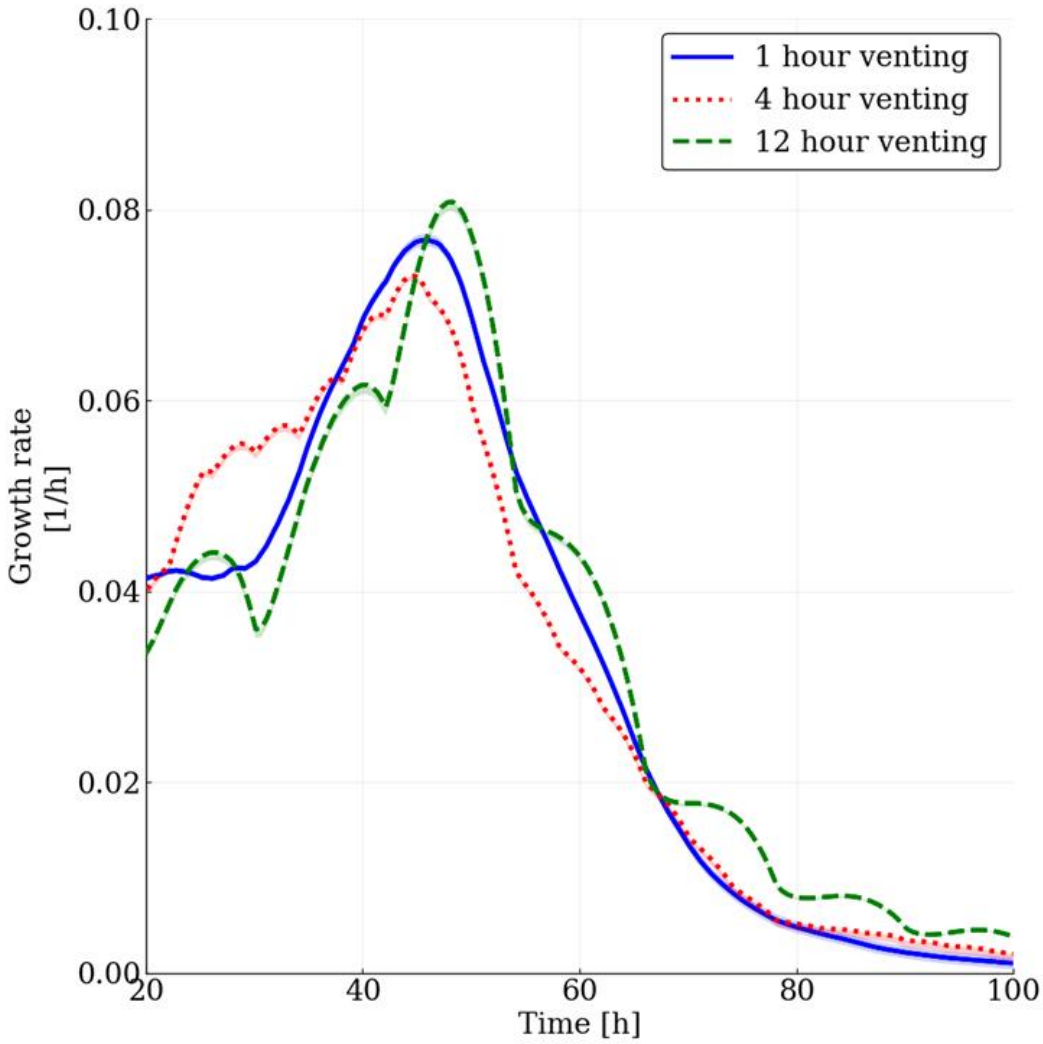

Figure 6.6: The observed instantaneous growth rate is not a function of the venting frequency, suggesting that pressure accumulation does not adversely affect the growth rate of *N. lanati*. Thus, it is unlikely that hydrogen inhibition plays an important role in the observed growth rate decrease. Three triplicate sets of *N. lanati* growing on 40 mL of complex media and 0.5 grams corn stover were vented at 1, 4 and 12-hour intervals to investigate the effect venting time has on the growth rate of the fungus. Higher venting frequencies reduces the buildup of pressure in the closed system, leading to lower concentrations of the gaseous fermentation products. The maximum spot pressure observed during the 1-hour venting experiment was 4.9 kPag, suggesting that there was no significant buildup of hydrogen. In contrast, the maximum spot pressure during the 12-hour venting experiment was 49 kPag. In both

cases the growth rates were comparable. The growth rates were calculated using 12-hour intervals, and the shaded region represents 1 standard deviation from the solid mean curve. Media de-gassing effects can be seen in the periodic behavior observed during the 4-hour and 12-hour curves. The high pressure between venting intervals causes gas to accumulate in the liquid fraction. After venting, the entrained gas escapes into the headspace of the system, which has been reduced to atmospheric pressure, and causes a rapid build-up of pressure that is not related to the current pressure production rate. The higher the venting frequency the more attenuated this de-gassing effect becomes.

6.4 Conclusion

Here we have introduced a fully automated pressure measurement and venting device (A-APES) that can be used to infer the growth rate of microorganisms where gas production is related to biomass accumulation, such as anaerobic gut fungi (Haitjema *et al.*, 2014). The device is also relatively simple to construct and operate. It affords the user high resolution gas production information that can be used to non-invasively study microorganism growth dynamics. Furthermore, due to the Arduino base the device is easy to extend and modify if desired, possibly paving the way for the construction of a lab-scale chemostat tailored for rumen-based microorganism systems. Additionally, we have used this device to reveal the growth dynamics of a non-model anerobic gut fungus. Due to the very high-resolution data afforded by the device, it is apparent that gut fungal growth is punctuated by a short regime of very rapid growth, followed by a much longer regime where the growth rate slows down. This suggests that the slow growth rate associated with anaerobic gut fungi may be heavily influenced by culturing techniques, rather than internal metabolic limitations.

6.5 Acknowledgements

The authors wish to acknowledge funding support from the National Science Foundation (NSF) (MCB-1553721). This work was part of the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the Office of Biological and Environmental Research of the DOE Office of Science through contract DE-AC02–05CH11231 between Lawrence Berkeley National Laboratory and the DOE. S. E. Wilken is grateful for funding support from the Dow Discovery Fellowship, M Reilly would like to acknowledge EPSRC Innovation Fellowship funding (grant code: EP/S001581/1), and M. Reilly and M. K. Theodorou further acknowledge travel support to visit M. A. O'Malley's laboratory at UCSB courtesy of the Harper Adams University entrepreneurial fund.

VII. Conclusions and future directions

Anaerobic gut fungi are a promising clade of non-model fungi to explore for biomass breakdown and biological production, yet challenges remain to fully exploit them in a biotechnological context. These fungi are classified as non-model organisms due to the restrictive culturing conditions they require, their slow growth rate and their genetic intractability. Here we have documented several approaches aimed at elucidating the inner workings of their metabolism and how they can be paired with other organisms to channel lignocellulosic carbon into biotechnologically relevant materials.

7.1 Anaerobic gut fungi are genetically intractable, but omics-based analyses can be used to guide efforts to engineer them

While anaerobic gut fungi are genetically intractable, the vast array of omics datasets that have been collected for them can be used to guide future engineering efforts. Their uniquely GC-depleted genomes show several interesting features that have consequences for metabolic engineering. First, heterologous expression of gut fungal genes in other hosts needs to be carefully considered due to the extreme nucleotide biases present in gut fungal genes. Second, the gut fungi seem to prefer codons, and amino acids, that are also encoded for by GC-depleted sequences. Thus, expressing non-native genes in the gut fungi will likely require codon-optimization to ensure translational compatibility. Third, the high abundance of repetitive genomic sequences may play an important role in the functional efficiency of, for example, their CAZymes due to glycosylation. In sum, this suggests that future genetic engineering strategies will have to be specifically tailored to the unique properties of the gut fungi to optimally leverage their potential.

Their metabolism may also be used to exploit their lignocellulolytic abilities. The genome-scale metabolic model introduced in this thesis will likely prove to be a pivotal building block in understanding their phenotypic responses to environmental perturbations and shaping them to our benefit. Specifically, the metabolic degeneracy highlighted by the model is indicative of an organism that is well adapted to tuning its metabolism to suit the needs of the environment. This suggests that there is room to channel carbon and metabolic fluxes without compromising the organism unduly. These insights have applications in consortia engineering. For example, metabolic models can be used to rationally design and engineer inter-species interactions that prevent organisms from out-competing each other. This could prove valuable in promoting synthetic consortia stability. Moreover, spreading the metabolic burden associated with the production of value-added chemicals between different microbes can also be guided through genome-scale models. The precursor flux availability analysis highlighted in this thesis suggests that the gut fungi can be used in this context in bioprocesses. Additionally, the uncertainty associated with the hydrogenosomal metabolism presents an exciting opportunity to understand how these anaerobes survive, and thrive, in a highly competitive and challenging environment. Specifically, it is tempting to speculate that their growth rate can be significantly enhanced by engineering the hydrogenosome. Our current understanding suggests that the organelle is cofactor limited, suggesting a viable route to engineer it for better performance.

7.2 Anaerobic gut fungi show promise to be incorporated in synthetic consortia

The native environment of anaerobic gut fungi is typically dominated by fast growing prokaryotic members. This raises the question of how the relatively slow growing gut fungi

manage to persist in such a competitive habitat. In this thesis it was demonstrated that despite their low abundance, the gut fungi have a dramatic impact on the fermentation products produced by consortia that feature them. This suggests that, beyond their necessity to decompose lignocellulose into its constitutive sugar monomers for the host animal, their metabolic impact is likely to be important for the stability and function of the lignocellulolytic microbiome. Indeed, their ability to produce H_2 is critical for the growth of methanogens, for example. Model based simulations also indicate that the anaerobic gut fungi form a mutualistic relationship with methanogens due to their ability to cross-feed and not compete with each other for resources to the same extent as other heterotrophic bacteria also present in their microbiome.

Given the current genetic intractability of the gut fungi, their ability to form stable consortia with methanogens and bacteria, which are typically more amenable to genetic engineering, represents a promising alternative approach to utilizing their lignocellulolytic potential. It could be fruitful to pair the gut fungi with organisms that are known to form stable consortia with them and engineer the other members to produce valuable products. In sum, this suggests that there exist possible routes forward for designing consortia that channel carbon to products of interest without necessarily directly engineering the gut fungi.

7.3 Future directions

Several open questions remain that need to be addressed to better understand anaerobic gut fungi and unlock their potential. First is the metabolism and proteins found within the anaerobic fungal hydrogenosome. While the core enzymes of this organelle have been identified, the presence or absence of the bifurcating hydrogenase and/or a functioning proton

pumping mechanism would dramatically affect the way we understand how the gut fungi maintain their energy balance. Organellular isolation, purification and biochemical characterization will be a necessity in this regard. Second is a robust way to genetically engineer the gut fungi. The lack of genetic tools hampers our ability to channel metabolic fluxes (informed by the genome-scale model) to pathways of interest. The ability to insert a constitutively expressed fluorescent tag would also enable a robust way to estimate biomass accumulation, beyond the current gas production inference method. This will prove invaluable for consortia, and even monoculture, modeling. Third, metabolic modeling of other rumen-based microorganisms will facilitate the establishment of stable pairings with the gut fungi that are amenable to systematic analysis. Currently, co-cultures with the gut fungi and microorganisms not isolated from their native habitat tend to be unstable. On the other hand, consortia down-selected from the same rumen-based microbiome have been shown to be very stable. This is likely due to complex metabolic interactions that we do not fully understand yet, e.g. the role methanogens play in diverting metabolic flux to the hydrogenosome. Isolating and modeling an organism that is known to form a stable pairing with a gut fungus will aid in better studying and understanding these interactions. Taken together, these questions can be used to unlock the potential of anaerobic gut fungi for sustainable bioprocessing of lignocellulosic waste.

VIII. Appendices

8.1 Isolating, purifying and characterizing the fungal hydrogenosome

8.1.1 Introduction

Anaerobic gut fungi possess hydrogenosomes (Yarlett *et al.*, 1986; Marvin-Sikkema *et al.*, 1994b; Boxma *et al.*, 2004; Hackstein *et al.*, 2019). The hydrogenosome is a double membrane-enclosed, oxygen sensitive organelle that produces H_2 and ATP typically through substrate level phosphorylation (Muller *et al.*, 2012). These organelles are also found in other anaerobes, and are believed to be related to mitochondria, albeit highly reduced (Boxma *et al.*, 2005). It is believed that the gut fungal hydrogenosome does not possess an electron transport chain, despite one earlier study to the contrary (Marvin-Sikkema *et al.*, 1994a). Notably, other organisms, such as *Nyctotherus ovalis*, possess anaerobic mitochondria-like organelles that have an electron transport chain and produce H_2. These organelles are termed H_2 producing mitochondria instead of hydrogenosomes (Muller *et al.*, 2012). As shown in Chapter IV, much is unknown about the metabolic functioning of the gut fungal hydrogenosome. It is important that the metabolism of this organelle be understood because it likely plays a central role in the energy generation pathways of the gut fungi. Specific questions that need to be answered (see Chapter IV for the motivation):

1. What is the role of pyruvate formate oxidoreductase (PFO) and pyruvate formate lyase (PFL) in the hydrogenosome? Is PFL associated with H_2 production, and if so, how?
2. Is the hydrogenase in the hydrogenosome bifurcating?
3. Do complexes I and II form part of a reduced electron transport chain in the gut fungal hydrogenosome, as possibly suggested in Figure 8.1?

Here we discuss progress made in addressing these questions and some remaining obstacles.

8.1.2 Previous work isolating and purifying the gut fungal hydrogenosome

There are a number of studies that have isolated, purified and attempted to enzymatically characterize the gut fungal hydrogenosome (Yarlett *et al.*, 1986; Marvin-Sikkema *et al.*, 1993, 1994a). The papers by Marvin-Sikkema et al. primarily used gut fungal mycelium, while the paper by Yarlett et al. used zoospores to isolate the hydrogenosomes from the fungal cells. During our attempts to purify hydrogenosomes we primarily relied on the protocol proposed by Marvin-Sikkema and only used fungal mycelia for the extractions. We used both a hydrogenase as well as a malic enzyme assay (described in (Lindmark and Müller, 1973, 1974)) to validate that we isolated an organelle with the "expected" hydrogenosome activity.

While all the isolations we performed had both hydrogenase and malic enzyme activity, the purification part of the protocol (via sucrose density gradient fractionation) did not yield layers with distinct activity, i.e., essentially all the fractions exhibited significant hydrogenase and malic enzyme activities. This suggests that the lysing procedure used by Marvin-Sikkema (grinding the mycelia with sand) is an inappropriate technique. For both enzymatic characterization as well as proteomic analysis, it is crucial that the hydrogenosome fraction is as pure as possible. Mixed samples will make the characterization difficult since some of the uncertainty in the hydrogenosome metabolism centers around enzymes that could be localized to either the cytosol, the hydrogenosome, or both.

Going forward, we suggest that the protocol in Yarlett et al. be followed. Specifically, zoospores should be used instead to mycelia, and the lysis procedure should use something

similar to a Potter/Dounce homogenizer (as suggested in Yarlett et al. 1986). Electron microscopy should be used to compare the density gradient fractions to the micrographs in Yarlett et al. Concurrently, enzymatic assays should be performed to ensure that the fractions display hydrogenosomal activity. We recommend the hydrogenase, malic enzyme, and pyruvate ferredoxin oxidoreductase assays as described in Yarlett et al. Next we describe the current recommended protocol for isolating and purifying hydrogenosomes from anaerobic gut fungi.

8.1.3 Hydrogenosome isolation and purification protocol

This protocol is adapted from Yarlett et. al., "Hydrogenosomes in the rumen fungus *Neocallimastix pariciarum*", Biochemical Journal, 1986.

All work should be performed in the anaerobic chamber as O_2 deactivates the hydrogenase.

Isolation and purification

1. Inoculate 1 liter of gut fungi in complex media using a soluble carbon substrate (e.g., cellobiose). Let it grow at 39°C for 3-4 days or until the culture reaches "mid-exponential" phase i.e., actively growing.
2. Open the bottle(s) inside the anaerobic chamber. Collect the fungal zoospores by filtering the cultures through cheese cloth into 50 mL conical tubes (~20 tubes). Discard the cell matt in the cheese cloth. Centrifuge the tubes (with the filtrate) at 2500x g for 3 min at 39°C (the conical tubes are air-tight enough for these short spins). The zoospores, and some cell debris, will collect at the bottom of the conical tubes. Discard the supernatant in each tube and gently re-dissolve the zoospores in M2 media (i.e., the salt solution mentioned in Yarlett et. al.).
3. Wash the concentrated zoospores at least once (Yarlett et. al. does it twice).
4. Disrupt the pellets with a Teflon/glass Potter (a.k.a. Dounce) homogenizer. Use the disruption buffer (g/l): sucrose, 85.6; EDTA, 0.27; KCI, 1.49; KH2PO4, 1.36; MgCl2, 1.0; Tris/HCl (pH 7.4), 1.21. Alternatively, one could also use a sonicator.
5. Centrifuge (4°C) the homogenate at 2600xg to remove debris – it will pellet out. Use the supernatant, which should contain all the organelles for the next step. Discard the pellet (P1).

6. Perform the following centrifugations at 4°C, with the supernatant from (5), using an ultra-centrifuge at 105x g-min. Collect the sediment (call this P2), keep the supernatant for controls as required, but keep/use the pellet (P2) – it contains the hydrogenosome – for downstream work.
7. Prepare a sucrose density gradient as described in Yarlett and use the ultra-centrifuge to fractionate the hydrogenosome.
8. Yarlett found that the hydrogenosomes collect at ~1.20 g/mL sucrose. Perform assays (hydrogenase, PFO, PFL, malic enzyme etc.) on each fraction and use the fraction with the highest hydrogenase (this is the one Yarlett et. al. favored) activity.
9. Use the mostly pure fraction of hydrogenosome in assays/proteomics for characterization.

Assays

1) The protocol for the hydrogenase and PFO assays, see (Lindmark and Müller, 1973).
2) For the protocol of the malic enzyme, see (Lindmark and Müller, 1974).

Note that both of the Lindmark et. al. papers contain many other useful assays for characterizing the hydrogenosome.

8.1.4 Characterizing the hydrogenosome

Once the hydrogenosomes have been purified we recommend that the other assays in Yarlett et al. be performed. These assays can be used to confirm the presence (or absence) of enzymes in the hydrogenosome model. Additionally, the bifurcating hydrogenase can be assayed using a method similar to that found in (Gerrit J. Schut and Adams, 2009). The electron transport chain may be probed using techniques similar to those used for mitochondria (e.g., fluorescent imaging and/or other assays). Finally, proteomic analysis will add value to the analysis and discovery process.

8.1.5 Conclusion

The hydrogenosome is an understudied yet important organelle in the gut fungal metabolism. It is likely responsible for a large portion of the energy generated by the fungal

cells. Isolating, purifying and enzymatically characterizing its metabolism will be a valuable addition to understanding the cellular metabolism of anaerobic gut fungi.

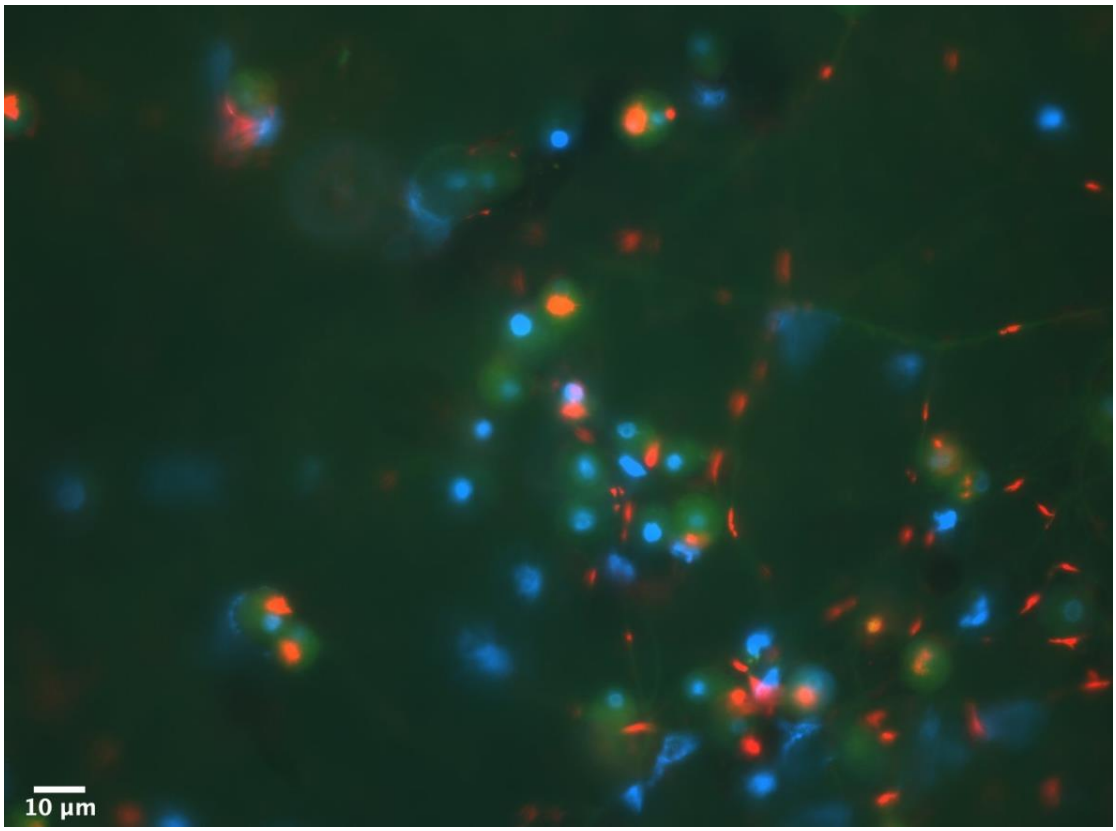

Figure 8.1: Zoospores were stained for the presence of intracellular organelles with electrical potential using the dye JC-1 as well as DAPI to illuminate the nuclear structure. Blue represents the DAPI stain and orange/red represents the JC-1 stain, which indicates the presence of a pH gradient in the hydrogenosome (JC-1 is a mitrochondrial selective strain, but here the mitochondria is replaced with the hydrogenosome). The JC-1 Dye was purchased from Invitrogen (Part No. T3168, Carlsbad, CA, USA) and a standard protocol was used to visualize the presence of electrochemical gradients. Briefly, JC-1 was dissolved in DMSO (1 mg/mL) and frozen until use. Dye aliquots were thawed and added to cultures of anaerobic gut fungal zoospores using final dye concentrations of 1 µg/mL. Zoospores were incubated with JC-1 for 30 minutes anaerobically in standard M2 medium at +39°C. After incubation,

cultures were filtered onto 3 µm polycarbonate membranes (Part No. TSTP02500 MilliporeSigma, Burlington, MA, USA) with a nitrocellulose backing filter (Part No. HAWP04700, MilliporeSigma). Cells were counterstained with DAPI (2 µg/mL) and mounted on glass slides using an antifade mounting solution composed of 4:1 Citiflour:Vectasheild (Part No. AF1, Electron Microscopy Sciences, Hatfield PA, USA: Part No. H-1000, Vector Laboratories, Burlingame, CA, USA). Prepared slides were placed on ice and imaged immediately using a Zeiss Axiovert M200 fluorescence microscope (Carl Zeiss AG, Oberkochen, DE). Image courtesy of Thomas S. Lankiewicz.

8.2 Using neural networks to learn from biological datasets

8.2.1 Introduction

Machine learning is a broad term used to describe computer-based algorithms designed to extract and recognize patterns in data for classification, regression or prediction. With the rise of high throughput experiments and the resultant "big" biological datasets, it has become possible to apply machine learning algorithms to biological problems (Camacho *et al.*, 2018; Costello and Martin, 2018). Of particular importance to biological modeling is the ability to predict the function of a gene from sequencing data. Recent examples include gene annotation (Clauwaert, Menschaert and Waegeman, 2019) and protein localization (Almagro Armenteros *et al.*, 2017) software. In the latter case, a neural network was trained to predict the localization of proteins using only the primary amino acid sequence data with ~70% accuracy. This promising result suggests that complex data can be used to guide experimental and modeling effort with high reliability. In this section neural networks are introduced and used to predict the enrichment of a library of yeast cells with mutations in the binding pocket of an A_2a receptor (Yoo, Daugherty and O'Malley, 2020).

8.2.2 Neural networks

A neural network is a supervised machine learning algorithm that maps inputs to outputs using a series of nonlinear functions. Arbitrary network topologies may be used to effect this transformation. Figure 8.2 depicts a basic feed forward topology using three layers.

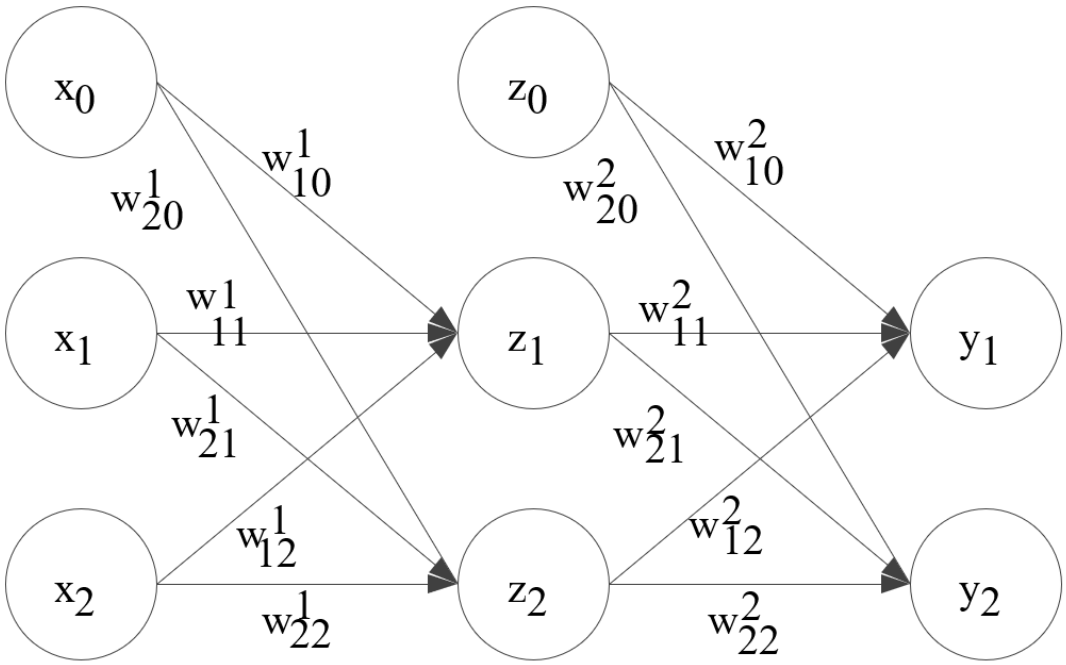

Figure 8.2: A simple feed forward neural network with one input layer (**x**), one hidden layer (**z**), and one output layer (**y**).

Figure 8.2 shows a specific example where there are only two nodes (x_1, x_2) in the input layer. Similarly, there are only two hidden and output layer nodes respectively. The bias terms (x_0, $z_0 = 1.0$) are artificial inputs that are included so that an input of zero to a layer does not necessarily have to result in a constant output. Generalization to an arbitrary number of nodes is straightforward. When more than one hidden layer is present the model is called a deep network. We will only consider conventional feed forward neural networks here. In the case of *D* input nodes, *M* hidden nodes in a single layer, and *K* output nodes, the relationship

between the input layer to a specific output node is shown in equation (8.1). Each layer can have a nonlinear function (termed an activation function) applied to it to allow greater model expressivity. In equation (8.1) these nonlinear functions are denoted by f(…).

$$y_k(\mathbf{x}, \mathbf{w}) = f_2(\sum_{j=0}^{M} w_{kj}^{(2)} f_1(\sum_{i=0}^{D} w_{ji}^{(1)} x_i)) \; with \; k \in [1, 2, ..., K] \tag{8.1}$$

Despite the apparent simplicity of equation (8.1), theoretical analysis has shown that a three-layer neural network is a universal function approximator, i.e. it can, to arbitrary precision, approximate any function on a compact domain given enough hidden layer nodes (Sonoda and Murata, 2017).

A challenge associated with using neural networks for regression or classification is that, although they may be able to model any function (like a discriminant function for classification), finding the set of parameters (e.g. the number of hidden nodes or weights **w**) that achieves this is difficult. This is due to the extremely high dimensionality of these networks for all but trivial problems. Configuring the network for the best results entails finding the optimum number of hidden nodes and the optimum weights. In practice the number of hidden nodes is generally adjusted through trial and error while conventional or unconventional optimization algorithms are used to minimize the error produced over training examples.

The error function and nonlinear activations functions are usually determined based on the type of problem under consideration. In the case of multi-class classification, the cross-entropy error function and the soft-max output function are used (Dreiseitl and Ohno-Machado, 2002). This allows a probabilistic interpretation of the results that aids analysis and

extension. For regression, mean squared error is often used (Dreiseitl and Ohno-Machado, 2002). The nonlinear function mapping the input layer to the hidden layer can theoretically be any function, although some guidelines exist that make training the system much more manageable[4].

Training a neural network entails modifying the weight matrices ($\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ in equation (8.1)) to minimize the prediction or regression error. Due to the aforementioned dimensionality issues, this can be challenging to achieve efficiently. Various optimization strategies are used to find the minimum of the error function, with the most popular strategy being gradient descent, as shown in equation (8.2), where η is known as the learning rate.

$$\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} - \eta \nabla E(\mathbf{w}^{\tau}) \tag{8.2}$$

It may be necessary to run the optimization routine multiple times because the nonlinear (and quite possibly non-convex) nature of the problem almost guarantees that there will be local minima (which the algorithm might get trapped in). Repeatedly running the optimization algorithm with different initial weights of $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ is a way of avoiding this problem and finding the global (or close to it) minimum of the error function.

Error back-propagation (backprop) is an efficient way of calculating $\nabla E(\mathbf{w}^{\tau})$ that is necessary for most derivative based optimization methods(Karnin, 1990). Backprop works by forward propagating an input vector signal $\mathbf{x}$ (i.e. calculating the values of the hidden and output nodes) and then using that information to calculate the partial derivatives over each weighting vector going backwards from the outputs to the inputs. Due to the increasing

[4] See https://stats.stackexchange.com/questions/352036/what-should-i-do-when-my-neural-network-doesnt-learn

popularity of neural network based machine learning architectures, well designed, optimized programming libraries exist (Innes, 2018; Innes *et al.*, 2018). These libraries can be used to rapidly develop neural networks without having to program the fundamental algorithms (e.g. backprop) from scratch.

8.2.3 Predicting GPCR enrichments using neural networks

Making use of machine learning to assist directed protein evolution is becoming an increasingly common way to deal with the exceptionally large design space inherent to protein engineering (Wu *et al.*, 2019). Assuming only the 20 standard amino acids can be used in protein synthesis, and the average length of a protein is 300 amino acids, suggests that 20^{300} distinct proteins can be made. This is an impractically large space to exhaustively explore with experiments to optimize protein performance. Instead, machine learning methods can be used to screen proteins *in silico*. To achieve this, a fraction of the space needs to be explored experimentally to train a machine learning algorithm. By finding patterns in the data, the algorithm can be used to direct experimental effort into areas more likely to yield promising results, as has been demonstrated previously (Wu *et al.*, 2019).

To explore this idea, we have made use of data from a deep mutational scanning library that correlates mutations in the binding pocket of an A_2a GPCR receptor in *S. cerevisiae* to the rate of enrichment of each mutational variant based on experimental data (Yoo, Daugherty and O'Malley, 2020). The goal is to be able to predict the enrichment rate of a variant given only the primary amino acid sequence data of the area that was mutated in the binding pocket. Table 8.1 shows a few examples of the raw data.

A single hidden layer feedforward neural network was constructed to model the relationship between the variants and the reported enrichment rate. A mean squared error function (predicted enrichment vs. actual enrichment) was used as the loss function to be optimized. Each amino acid was encoded as previously described (Sønderby *et al.*, 2015) to form a unique input vector for each variant. Approximately 10% of the raw data was held out for testing, with the balance used for training. Of the ~180,000 variants recorded in the study, only ~2200 had a non-zero reported enrichment. Only this reduced set of data points were used for training/testing due to confidence issues with the remainder of the data[5].

Table 8.1: Example of the format of the raw data used to train the neural network.

Variant (amino acid sequence of binding pocket)	Enrichment rate
SLNIG	790.5
TSWIH	787.2
TTYLH	635.2
…	…

Figure 8.3 shows the training and testing errors when a neural network with 256 hidden nodes was used. For both the training and testing cases the mean squared error is very high. The high error renders the predictions unreliable. Interestingly, it is clear that over-fitting occurs because the training error is significantly lower than the testing error at later epochs. The likely explanation for the poor predictive power of the network is the low number of

[5] The author, Dr. Justin Yoo, recommended this.

training examples and relative complexity of the task. In other regression work that uses neural networks on biological data, more than two orders of magnitude more training data is typically used (Wu *et al.*, 2019).

8.2.4 Conclusion

Here we have developed a small feed forward neural network to attempt to predict the enrichment rate of mutant variants of GPCRs. While neural networks are becoming a powerful tool in biology, the size of the dataset was too small to make reliable predictions.

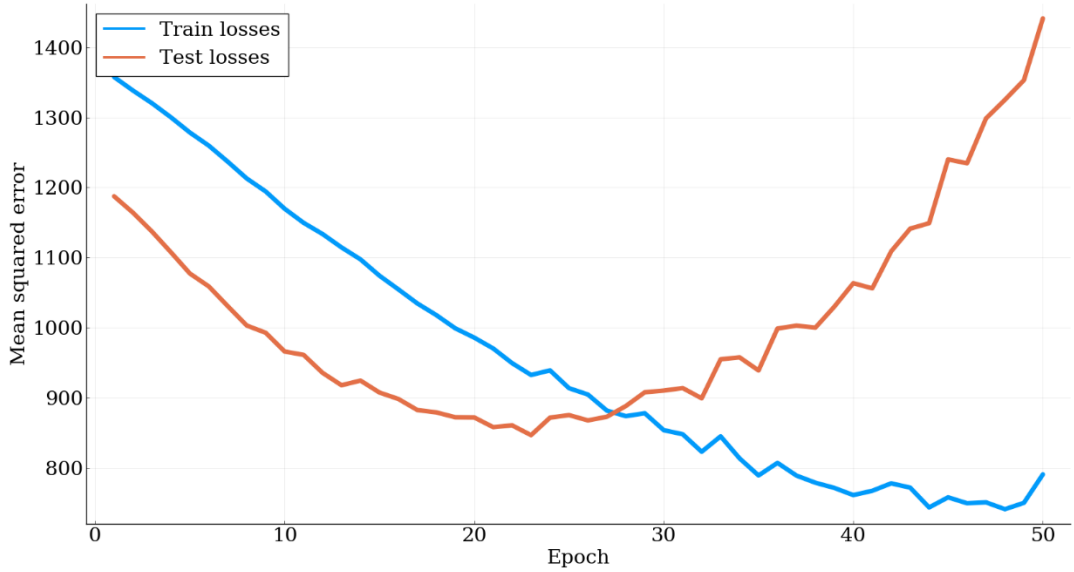

Figure 8.3: Training and testing errors for a neural network composed of 256 hidden nodes, using the ReLu activation function. The high errors are likely due to the small number of training data available. Overfitting is seen to occur at later epochs.

IX. References

Adesogan, A. T., Krueger, N. K. and Kim, S. C. (2005) 'A novel, wireless, automated system for measuring fermentation gas production kinetics of feeds and its application to feed characterization', *Animal Feed Science and Technology*, 123-124 Part 1, pp. 211–223. doi: 10.1016/j.anifeedsci.2005.04.058.

Adney, W. S. *et al.* (1991) 'Anaerobic digestion of lignocellulosic biomass and wastes - Cellulases and related enzymes', *Applied Biochemistry and Biotechnology*. Humana Press, 30(2), pp. 165–183. doi: 10.1007/BF02921684.

Adrio, J. L. and Demain, A. L. (2014) 'Microbial enzymes: tools for biotechnological processes.', *Biomolecules*. Multidisciplinary Digital Publishing Institute (MDPI), 4(1), pp. 117–139. doi: 10.3390/biom4010117.

Akhmanova, A. *et al.* (1999) 'A hydrogenosome with pyruvate formate-lyase: anaerobic chytrid fungi use an alternative route for pyruvate catabolism', *Molecular Microbiology*. John Wiley & Sons, Ltd, 32(5), pp. 1103–1114. doi: 10.1046/j.1365-2958.1999.01434.x.

Akiva, E. *et al.* (2014) 'The Structure–Function Linkage Database', *Nucleic Acids Research*, 42(D1), pp. D521–D530. doi: 10.1093/nar/gkt1130.

Albà, M. M., Tompa, P. and Veitia, R. A. (2007) 'Amino Acid Repeats and the Structure and Evolution of Proteins', in *Gene and Protein Evolution*. Basel: KARGER, pp. 119–130. doi: 10.1159/000107607.

Allen, F. *et al.* (2009) 'Mary Elizabeth Hickox Mandels, 90, bioenergy leader', *Biotechnology for Biofuels*, 2, p. 22. doi: 10.1186/1754-6834-2-22.

Almagro Armenteros, J. J. *et al.* (2017) 'DeepLoc: prediction of protein subcellular localization using deep learning', *Bioinformatics (Oxford, England)*, 33(21), pp. 3387–3395.

doi: 10.1093/bioinformatics/btx431.

Alper, H. and Stephanopoulos, G. (2009) ‘Engineering for biofuels: Exploiting innate microbial capacity or importing biosynthetic potential?’, *Nature Reviews Microbiology*, pp. 715–723. doi: 10.1038/nrmicro2186.

Amore, A., Giacobbe, S. and Faraco, V. (2013) ‘Regulation of cellulase and hemicellulase gene expression in fungi.’, *Current genomics*. Bentham Science Publishers, 14(4), pp. 230–49. doi: 10.2174/1389202911314040002.

Arazoe, T. *et al.* (2015) ‘Tailor-made TALEN system for highly efficient targeted gene replacement in the rice blast fungus’, *Biotechnology and Bioengineering*, 112(7), pp. 1335–1342. doi: 10.1002/bit.25559.

Arima, K., Iwasaki, S. and Tamura, G. (1967) ‘Milk Clotting Enzyme from Microorganisms’, *Agricultural and Biological Chemistry*, 31(5), pp. 540–551. doi: 10.1080/00021369.1967.10858849.

Arkin, A. P. *et al.* (2018) ‘KBase: The United States Department of Energy Systems Biology Knowledgebase’, *Nature Biotechnology*, 36(7), pp. 566–569. doi: 10.1038/nbt.4163.

Arnaud, M. B. *et al.* (2010) ‘The Aspergillus Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the Aspergillus research community’, *Nucleic Acids Research*, 38(suppl 1), pp. D420–D427. doi: 10.1093/nar/gkp751.

Artzi, L., Bayer, E. A. and Moraïs, S. (2017) ‘Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides’, *Nature Reviews Microbiology*, 15(2), pp. 83–95. doi: 10.1038/nrmicro.2016.164.

Atasoglu, C. and Wallace, R. J. (2002) ‘De novo synthesis of amino acids by the ruminal anaerobic fungi, *Piromyces communis* and *Neocallimastix frontalis*’, *FEMS Microbiology*

Letters. Narnia, 212(2), pp. 243–247. doi: 10.1111/j.1574-6968.2002.tb11273.x.

Aung, H. W., Henry, S. A. and Walker, L. P. (2013) 'Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism', *Industrial Biotechnology*, 9(4), pp. 215–228. doi: 10.1089/ind.2013.0013.

Bach, A., Calsamiglia, S. and Stern, M. D. (2005) 'Nitrogen Metabolism in the Rumen', *Journal of Dairy Science*. Elsevier, 88, pp. E9–E21. doi: 10.3168/JDS.S0022-0302(05)73133-7.

Bajpai, P., Mehna, A. and Bajpai, P. K. (1993) 'Decolorization of kraft bleach plant effluent with the white rot fungus Trametes versicolor', *Process Biochemistry*, 28(6), pp. 377–384. doi: 10.1016/0032-9592(93)80024-B.

Banerjee, G., Scott-Craig, J. S. and Walton, J. D. (2010) 'Improving enzymes for biomass conversion: A basic research perspective', *Bioenergy Research*, 3(1), pp. 82–92. doi: 10.1007/s12155-009-9067-5.

Bareither, R. and Pollard, D. (2011) 'A review of advanced small-scale parallel bioreactor technology for accelerated process development: Current state and future need', *Biotechnology Progress*. American Chemical Society (ACS), 27(1), pp. 2–14. doi: 10.1002/btpr.522.

Beck, A. E., Hunt, K. A. and Carlson, R. P. (2018) 'Measuring cellular biomass composition for computational biology applications', *Processes*. MDPI AG, 6(5). doi: 10.3390/pr6050038.

Beckham, G. T. *et al.* (2010) 'The O-Glycosylated Linker from the Trichoderma reesei Family 7 Cellulase Is a Flexible, Disordered Protein', *Biophysical Journal*. Cell Press, 99(11), pp. 3773–3781. doi: 10.1016/J.BPJ.2010.10.032.

Beckham, G. T. *et al.* (2012) 'Harnessing glycosylation to improve cellulase activity', *Current Opinion in Biotechnology*, 23(3), pp. 338–345. doi: 10.1016/j.copbio.2011.11.030.

Bergenholm, D. *et al.* (2019) 'Construction of mini-chemostats for high-throughput strain characterization', *Biotechnology and Bioengineering*, 116(5), pp. 1029–1038. doi: 10.1002/bit.26931.

Bezanson, J. *et al.* (2017) 'Julia: A Fresh Approach to Numerical Computing *', *Society for Industrial and Applied Mathematics*, 59(1). doi: 10.1137/141000671.

Bhattacharya, A. S., Bhattacharya, A. and Pletschke, B. I. (2015) 'Synergism of fungal and bacterial cellulases and hemicellulases: A novel perspective for enhanced bio-ethanol production', *Biotechnology Letters*, 37(6), pp. 1117–1129. doi: 10.1007/s10529-015-1779-3.

Billings, A. F. *et al.* (2015) 'Genome sequence and description of the anaerobic lignin-degrading bacterium Tolumonas lignolytica sp. nov.', *Standards in Genomic Sciences*, 10(1), p. 106. doi: 10.1186/s40793-015-0100-3.

Birdsell, J. (2002) 'Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution', *Mol. Biol. Evol.*, 19, pp. 1181–1197.

Bischof, R. H., Ramoni, J. and Seiboth, B. (2016) 'Cellulases and beyond: the first 70 years of the enzyme producer Trichoderma reesei.', *Microbial cell factories*, 15(1), p. 106. doi: 10.1186/s12934-016-0507-6.

Blackwell, M. (2011) 'The Fungi: 1, 2, 3 ... 5.1 million species?', *American Journal of Botany*, 98(3), pp. 426–438. doi: 10.3732/ajb.1000298.

Blazeck, J. and Alper, H. (2010) 'Systems metabolic engineering: Genome-scale models and beyond', *Biotechnology Journal*, 5(7), pp. 647–659. doi: 10.1002/biot.200900247.

Boccazzi, P. *et al.* (2005) 'Gene expression analysis of Escherichia coli grown in

miniaturized bioreactor platforms for high-throughput analysis of growth and genomic data', *Applied Microbiology and Biotechnology*, 68(4), pp. 518–532. doi: 10.1007/s00253-005-1966-6.

Boch, J. *et al.* (2009) 'Breaking the code of DNA binding specificity of TAL-type III effectors.', *Science (New York, N.Y.)*, 326(5959), pp. 1509–1512. doi: 10.1126/science.1178811.

Bokinsky, G. *et al.* (2011) 'Synthesis of three advanced biofuels from ionic liquid-pretreated switchgrass using engineered Escherichia coli', *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), pp. 19949–19954. doi: 10.1073/pnas.1106958108.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Bonugli-Santos, R. C. *et al.* (2015) 'Marine-derived fungi: diversity of enzymes and biotechnological applications', *Frontiers in Microbiology*. Frontiers, 6, p. 269. doi: 10.3389/fmicb.2015.00269.

Borneman, W. S. *et al.* (1991) 'Isolation and characterization of p-coumaroyl esterase from the anaerobic fungus Neocallimastix strain MC-2.', *Applied and environmental microbiology*, 57(8), pp. 2337–2344.

Bothast, R. J. and Schlicher, M. A. (2005) 'Biotechnological processes for conversion of corn into ethanol', *Applied Microbiology and Biotechnology*, 67(1), pp. 19–25. doi: 10.1007/s00253-004-1819-8.

Boxma, B. *et al.* (2004) 'The anaerobic chytridiomycete fungus Piromyces sp. E2

produces ethanol via pyruvate:formate lyase and an alcohol dehydrogenase E', *Molecular Microbiology*. John Wiley & Sons, Ltd, 51(5), pp. 1389–1399. doi: 10.1046/j.1365-2958.2003.03912.x.

Boxma, B. *et al.* (2005) 'An anaerobic mitochondrion that produces hydrogen', *Nature*. Nature Publishing Group, 434(7029), pp. 74–79. doi: 10.1038/nature03343.

Boxma, B. *et al.* (2007) 'The [FeFe] hydrogenase of Nyctotherus ovalis has a chimeric origin', *BMC Evolutionary Biology*. BioMed Central, 7(1), p. 230. doi: 10.1186/1471-2148-7-230.

Boyarskiy, S. and Tullman-Ercek, D. (2015) 'Getting pumped: Membrane efflux transporters for enhanced biomolecule production', *Current Opinion in Chemical Biology*. Elsevier Ltd, pp. 15–19. doi: 10.1016/j.cbpa.2015.05.019.

Bray, N. L. *et al.* (2016) 'Near-optimal probabilistic RNA-seq quantification', *Nature Biotechnology*. Nature Publishing Group, 34(5), pp. 525–527. doi: 10.1038/nbt.3519.

Brenner, K., You, L. and Arnold, F. H. (2008) 'Engineering microbial consortia: a new frontier in synthetic biology', *Trends in Biotechnology*, pp. 483–489. doi: 10.1016/j.tibtech.2008.05.004.

Brown, M. E. and Chang, M. C. (2014) 'Exploring bacterial lignin degradation', *Current Opinion in Chemical Biology*, 19, pp. 1–7. doi: 10.1016/j.cbpa.2013.11.015.

Brownlee, A. G. (1989) 'Remarkably AT-rich genomic DNA from the anaerobic fungus Neocallimastix.', *Nucleic acids research*, 17(4), pp. 1327–35.

Calkins, S. S. *et al.* (2018) 'Development of an RNA interference (RNAi) gene knockdown protocol in the anaerobic gut fungus *Pecoramyces ruminantium* strain C1A', *PeerJ*, 6, p. e4276. doi: 10.7717/peerj.4276.

Camacho, C. *et al.* (2009) 'BLAST+: architecture and applications', *BMC Bioinformatics*, 10(1), p. 421. doi: 10.1186/1471-2105-10-421.

Camacho, D. M. *et al.* (2018) 'Next-Generation Machine Learning for Biological Networks', *Cell*. Cell Press, pp. 1581–1592. doi: 10.1016/j.cell.2018.05.015.

Camiolo, S. *et al.* (2019) 'An analysis of codon bias in six red yeast species', *Yeast*. John Wiley & Sons, Ltd, 36(1), pp. 53–64. doi: 10.1002/yea.3359.

Campanaro, S. *et al.* (2016) 'Metagenomic analysis and functional characterization of the biogas microbiome using high throughput shotgun sequencing and a novel binning strategy', *Biotechnology for Biofuels*, 9(1), p. 26. doi: 10.1186/s13068-016-0441-1.

Carlson, M. *et al.* (2019) 'PFAM.db: A set of protein ID mappings for PFAM', *R package version 3.8.2*.

Caspi, R. *et al.* (2018) 'The MetaCyc database of metabolic pathways and enzymes', *Nucleic Acids Research*. Oxford University Press, 46(D1), pp. D633–D639. doi: 10.1093/nar/gkx935.

Chakraborty, R. and Coates, J. D. (2004) 'Anaerobic degradation of monoaromatic hydrocarbons', *Applied Microbiology and Biotechnology*, 64(4), pp. 437–446. doi: 10.1007/s00253-003-1526-x.

Chan, P. P. and Lowe, T. M. (2019) 'tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences', in *Methods in molecular biology (Clifton, N.J.)*, pp. 1–14. doi: 10.1007/978-1-4939-9173-0_1.

Chandel, A. K. *et al.* (2015) 'Biodelignification of lignocellulose substrates: An intrinsic and sustainable pretreatment strategy for clean energy production', *Critical Reviews in Biotechnology*, 35(3), pp. 281–293. doi: 10.3109/07388551.2013.841638.

Chandra, R. P. *et al.* (2007) 'Substrate pretreatment: The key to effective enzymatic hydrolysis of lignocellulosics?', *Advances in Biochemical Engineering/Biotechnology*, 108(May), pp. 67–93. doi: 10.1007/10_2007_064.

Chang, Y. *et al.* (2015) 'Phylogenomic Analyses Indicate that Early Fungi Evolved Digesting Cell Walls of Algal Ancestors of Land Plants', *Genome Biology and Evolution*, 7(6), pp. 1590–1601. doi: 10.1093/gbe/evv090.

Chen, I.-M. A. *et al.* (2017) 'IMG/M: integrated genome and metagenome comparative data analysis system', *Nucleic Acids Research*, 45(D1), pp. D507–D516. doi: 10.1093/nar/gkw929.

Chen, W. *et al.* (2012) 'Genomic characteristics comparisons of 12 food-related filamentous fungi in tRNA gene set, codon usage and amino acid composition', *Gene*. Elsevier, 497(1), pp. 116–124. doi: 10.1016/J.GENE.2012.01.016.

Chen, Y. *et al.* (2012) 'Kraft lignin biodegradation by Novosphingobium sp. B-7 and analysis of the degradation process', *Bioresource Technology*, 123, pp. 682–685. doi: 10.1016/j.biortech.2012.07.028.

Cheng, Y. S. *et al.* (2014) 'Structural analysis of a glycoside hydrolase family 11 xylanase from Neocallimastix patriciarum: Insights into the molecular basis of a thermophilic enzyme', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., 289(16), pp. 11020–11028. doi: 10.1074/jbc.M114.550905.

Cheng, Y. S. *et al.* (2015) 'Improving the catalytic performance of a GH11 xylanase by rational protein engineering', *Applied Microbiology and Biotechnology*. Springer Verlag, 99(22), pp. 9503–9510. doi: 10.1007/s00253-015-6712-0.

Cherry, J. M. *et al.* (2012) 'Saccharomyces Genome Database: the genomics resource of

budding yeast', *Nucleic Acids Research*, 40(Database issue), pp. D700–705. doi: 10.1093/nar/gkr1029.

Chinnici, F. *et al.* (2005) 'Optimization of the determination of organic acids and sugars in fruit juices by ion-exclusion liquid chromatography', *Journal of Food Composition and Analysis*, 18(2–3), pp. 121–130. doi: 10.1016/j.jfca.2004.01.005.

Chokhawala, H. A. *et al.* (2015) 'Mutagenesis of Trichoderma reesei endoglucanase I: impact of expression host on activity and stability at elevated temperatures', *BMC Biotechnology*. BioMed Central, 15(1), p. 11. doi: 10.1186/s12896-015-0118-z.

Chubukov, V. *et al.* (2018) 'Synthetic and systems biology for microbial production of commodity chemicals', *npj Systems Biology and Applications*. Nature Publishing Group, pp. 1–11. doi: 10.1038/npjsba.2016.9.

Clauwaert, J., Menschaert, G. and Waegeman, W. (2019) 'DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns', *Nucleic Acids Research*, 47(6), p. 36. doi: 10.1093/nar/gkz061.

Coker, J. A. (2016) 'Extremophiles and biotechnology: current uses and prospects.', *F1000Research*. Faculty of 1000 Ltd, 5. doi: 10.12688/f1000research.7432.1.

Corradi, N. *et al.* (2010) 'The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis.', *Nature communications*, 1, p. 77. doi: 10.1038/ncomms1082.

Costello, Z. and Martin, H. G. (2018) 'A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data', *npj Systems Biology and Applications*. Nature Publishing Group, 4(1), p. 19. doi: 10.1038/s41540-018-0054-3.

Cowan, D. *et al.* (2005) 'Metagenomic gene discovery: Past, present and future', *Trends*

in Biotechnology, 23(6), pp. 321–329. doi: 10.1016/j.tibtech.2005.04.001.

Crous, P. W. *et al.* (2012) 'Fungal Planet description sheets: 107–127', *Persoonia : Molecular Phylogeny and Evolution of Fungi*, 28, pp. 138–182. doi: 10.3767/003158512X652633.

Crown, S. B., Long, C. P. and Antoniewicz, M. R. (2016) 'Optimal tracers for parallel labeling experiments and 13 C metabolic flux analysis: A new precision and synergy scoring system', *Metabolic Engineering*. Academic Press Inc., 38, pp. 10–18. doi: 10.1016/j.ymben.2016.06.001.

Curran, K. A. and Alper, H. S. (2012) 'Expanding the chemical palate of cells by combining systems biology and metabolic engineering', *Metabolic Engineering*, 14, pp. 289–297. doi: 10.1016/j.ymben.2012.04.006.

Curtis, S. J. and Epstein, W. (1975) 'Phosphorylation of D glucose in Escherichia coli mutants defective in glucosephosphotransferase, mannosephosphotransferase, and glucokinase', *Journal of Bacteriology,* 122(3), pp. 1189–1199.

Dashtban, M., Schraft, H. and Qin, W. (2009) 'Fungal bioconversion of lignocellulosic residues: Opportunities & perspectives', *International Journal of Biological Sciences*, 5(6), pp. 578–595. doi: 10.7150/ijbs.5.578.

Davies, D. R. *et al.* (1993) 'Distribution of anaerobic fungi in the digestive tract of cattle and their survival in faeces', *Journal of General Microbiology*, 139(6), pp. 1395–1400. doi: 10.1099/00221287-139-6-1395.

Davies, Z. S. *et al.* (2000) 'An automated system for measuring gas production from forages inoculated with rumen fluid and its use in determining the effect of enzymes on grass silage', *Animal Feed Science and Technology*, 83(3–4), pp. 205–221. doi: 10.1016/S0377-

8401(99)00138-8.

Davis, R. *et al.* (2013) ‘Process design and economics for the conversion of lignocellulosic biomass to hydrocarbons: Dilute-acid and enzymatic deconstruction of biomass to sugars and biological conversion of sugars to hydrocarbons’, *National Renewable Energy Laboratory*, p. NREL/TP-5100-60223.

Dean, R. A. *et al.* (2005) ‘The genome sequence of the rice blast fungus Magnaporthe grisea’, *Nature*, 434(7036), pp. 980–986. doi: 10.1038/nature03449.

Demirbas, A. and Demirbas, M. F. (2010) ‘Biorefineries’, in, pp. 159–181. doi: 10.1007/978-1-84996-050-2_7.

Deshpande, N. *et al.* (2008) ‘Protein glycosylation pathways in filamentous fungi’, *Glycobiology*. Narnia, 18, pp. 626–637. doi: 10.1093/glycob/cwn044.

Diener, A. C. and Fink, G. R. (1996) ‘DLH1 is a functional Candida albicans homologue of the meiosis-specific gene DMC1’, *Genetics*, 143(2), pp. 769–776.

Dighton, J. (2007) ‘Nutrient cycling by saprotrophic fungi in terrestrial habitats’, *Environmental and Microbial Relationships*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 287–300. doi: 10.1007/978-3-540-71840-6_16.

Dollhofer, V. *et al.* (2019) ‘The biotechnological potential of the anaerobic gut fungi’, in *The Mycota*.

Doyle, J. J. and Doyle, J. L. (1987) ‘A rapid DNA isolation procedure for small quantities of fresh leaf tissue’, *PHYTOCHEMICAL BULLETIN*. Available at: https://worldveg.tind.io/record/33886 (Accessed: 4 August 2020).

Doyle, S. (2011) ‘Fungal proteomics: from identification to function’, *FEMS microbiology letters*, 321(1), pp. 1–9. doi: 10.1111/j.1574-6968.2011.02292.x.

Dreiseitl, S. and Ohno-Machado, L. (2002) ‘Logistic regression and artificial neural network classification models: A methodology review’, *Journal of Biomedical Informatics*. Academic Press Inc., 35(5–6), pp. 352–359. doi: 10.1016/S1532-0464(03)00034-0.

Duarte, I. and Huynen, M. A. (2019) ‘Contribution of Lateral Gene Transfer to the evolution of the eukaryotic fungus Piromyces sp. E2: Massive bacterial transfer of genes involved in carbohydrate metabolism’, *bioRxiv*, (2001), p. 514042. doi: 10.1101/514042.

Duplessis, S. *et al.* (2011) ‘Obligate biotrophy features unraveled by the genomic analysis of rust fungi’, *Proceedings of the National Academy of Sciences*, 108(22), pp. 9166–9171. doi: 10.1073/pnas.1019315108.

Durand, R. *et al.* (1997) ‘Transient expression of the beta-glucuronidase gene after biolistic transformation of the anaerobic fungus Neocallimastix frontalis.’, *Current genetics*, 31(2), pp. 158–61.

Durand, R., Rascle, C. and Fèvre, M. (1999) ‘Expression of a catalytic domain of a Neocallimastix frontalis endoxylanase gene (xyn3) in Kluyveromyces lactis and Penicillium roqueforti.’, *Applied microbiology and biotechnology*, 52(2), pp. 208–214.

Duret, L. and Galtier, N. (2009) ‘Biased gene conversion and the evolution of mammalian genomic landscapes’, *Annual Review of Genomics and Human Genetics*, 10, pp. 285–311.

Ebrahim, A. *et al.* (2013) ‘COBRApy: COnstraints-Based Reconstruction and Analysis for Python’, *BMC Systems Biology*. BioMed Central, 7(1), p. 74. doi: 10.1186/1752-0509-7-74.

Eiteman, M. A., Lee, S. A. and Altman, E. (2008) ‘A co-fermentation strategy to consume sugar mixtures effectively’, *Journal of Biological Engineering*, 2. doi: 10.1186/1754-1611-2-3.

Elliott, A. R. *et al.* (1999) ‘Transformation of Bacillus subtilis using the particle inflow gun and submicrometer particles obtained by the polyol process’, *Analytical Biochemistry*, 269(2), pp. 418–420. doi: 10.1006/abio.1999.4036.

Falade, A. O. *et al.* (2017) ‘Lignin peroxidase functionalities and prospective applications.’, *MicrobiologyOpen*. Wiley-Blackwell, 6(1). doi: 10.1002/mbo3.394.

Feist, A. M. *et al.* (2006) ‘Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri’, *Molecular Systems Biology*, 2. doi: 10.1038/msb4100046.

Feist, A. M. and Palsson, B. O. (2010) ‘The biomass objective function’, *Current Opinion in Microbiology*, pp. 344–349. doi: 10.1016/j.mib.2010.03.003.

Field, J. A. *et al.* (1993) ‘Screening for ligninolytic fungi applicable to the biodegradation of xenobiotics’, *Trends in biotechnology*, 11(2), pp. 44–49. doi: 10.1016/0167-7799(93)90121-O.

Fischer, M., Durand, R. and Fèvre, M. (1995) ‘Characterization of the promoter region of the enolase-encoding gene enol from the anaerobic fungus Neocallimastix frontalis: Sequence and promoter analysis.’, *Current genetics*, 28(1), pp. 80–6.

Flahaut, N. A. L. *et al.* (2013) ‘Genome-scale metabolic model for Lactococcus lactis MG1363 and its application to the analysis of flavor formation’, *Applied Microbiology and Biotechnology*, 97(19), pp. 8729–8739. doi: 10.1007/s00253-013-5140-2.

Flamholz, A. *et al.* (2012) ‘EQuilibrator - The biochemical thermodynamics calculator’, *Nucleic Acids Research*, 40(D1). doi: 10.1093/nar/gkr874.

Flint, H. J. *et al.* (2008) ‘Polysaccharide utilization by gut bacteria: Potential for new insights from genomic analysis’, *Nature Reviews Microbiology*. Nature Publishing Group, pp.

121–131. doi: 10.1038/nrmicro1817.

Floudas, D. *et al.* (2012) ‘The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes.’, *Science (New York, N.Y.)*, 336(6089), pp. 1715–9. doi: 10.1126/science.1221748.

Fondon III, J. W. and Garner, H. R. (2004) *Molecular origins of rapid and continuous morphological evolution*, *Harvard Medical School*. Available at: www.pnas.orgcgidoi10.1073pnas.0408118101 (Accessed: 12 June 2019).

Gadd, G. M. (2007) ‘Geomycology: biogeochemical transformations of rocks, minerals, metals and radionuclides by fungi, bioweathering and bioremediation’, *Mycological Research*, 111(1), pp. 3–49. doi: 10.1016/j.mycres.2006.12.001.

Galagan, J. E. *et al.* (2003) ‘The genome sequence of the filamentous fungus Neurospora crassa’, *Nature*, 422(6934), pp. 859–868. doi: 10.1038/nature01554.

Galbe, M. and Zacchi, G. (2012) ‘Pretreatment: The key to efficient utilization of lignocellulosic materials’, *Biomass and Bioenergy*. Elsevier Ltd, 46, pp. 70–78. doi: 10.1016/j.biombioe.2012.03.026.

Galtier, N. (2001) ‘GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis’, *Genetics*, 159(1), pp. 907–911. doi: 10.3138/physio.61.1.51.

Garcia-Campayo, V. and Wood, T. M. (1993) ‘Purification and characterisation of a beta-D-xylosidase from the anaerobic rumen fungus Neocallimastix frontalis.’, *Carbohydrate research*, 242, pp. 229–245.

Gasiunas, G. *et al.* (2012) ‘Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria.’, *Proceedings of the National Academy of Sciences of the United States of America*, 109(39), pp. E2579–86. doi:

10.1073/pnas.1208507109.

Gazis, R. *et al.* (2016) ‘The genome of Xylona heveae provides a window into fungal endophytism’, *Fungal Biology*, 120(1), pp. 26–42. doi: 10.1016/j.funbio.2015.10.002.

Gentzsch, M. and Tanner, W. (1996) ‘The PMT gene family: protein O-glycosylation in Saccharomyces cerevisiae is vital.’, *The EMBO Journal*. John Wiley & Sons, Ltd, 15(21), pp. 5752–5759. doi: 10.1002/j.1460-2075.1996.tb00961.x.

Gerngross, T. U. (2004) ‘Advances in the production of human therapeutic proteins in yeasts and filamentous fungi’, *Nature Biotechnology*. Nature Publishing Group, 22(11), pp. 1409–1414. doi: 10.1038/nbt1028.

Gessner, M. O. *et al.* (2007) ‘Fungal decomposers of plant litter in aquatic ecosystems’, *Environmental and Microbial Relationships*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 301–324. doi: 10.1007/978-3-540-71840-6_17.

Gharechahi, J. and Salekdeh, G. H. (2018) ‘A metagenomic analysis of the camel rumen’s microbiome identifies the major microbes responsible for lignocellulose degradation and fermentation’, *Biotechnology for Biofuels*. BioMed Central Ltd., 11(1). doi: 10.1186/s13068-018-1214-9.

Gianoulis, T. A. *et al.* (2012) ‘Genomic analysis of the hydrocarbon-producing, cellulolytic, endophytic fungus Ascocoryne sarcoides’, *PLoS Genet*, 8(3), p. e1002558. doi: 10.1371/journal.pgen.1002558.

Gibney, E. (2016) ‘“Open-hardware” pioneers push for low-cost lab kit’, *Nature*. Nature Publishing Group, pp. 147–148. doi: 10.1038/531147a.

van der Giezen, M. *et al.* (1998) ‘The hydrogenosomal malic enzyme from the anaerobic fungus neocallimastix frontalis is targeted to mitochondria of the methylotrophic yeast

Hansenula polymorpha.', *Current genetics*, 33(2), pp. 131–135.

Gilbert, H. J. *et al.* (1992) 'Homologous catalytic domains in a rumen fungal xylanase: evidence for gene duplication and prokaryotic origin.', *Molecular microbiology*, 6(15), pp. 2065–2072.

Gilmore, S. P. *et al.* (2019) 'Top-Down Enrichment Guides in Formation of Synthetic Microbial Consortia for Biomass Degradation', *ACS Synthetic Biology*. American Chemical Society (ACS). doi: 10.1021/acssynbio.9b00271.

Gilmore, S. P., Henske, J. K. and O'Malley, M. A. (2015) 'Driving biomass breakdown through engineered cellulosomes.', *Bioengineered*, 6(4), pp. 204–208. doi: 10.1080/21655979.2015.1060379.

Glass, N. L., Grotelueschen, J. and Metzenberg, R. L. (1990) 'Neurospora crassa A mating-type region.', *Proceedings of the National Academy of Sciences of the United States of America*, 87(13), pp. 4912–4916.

Glémin, S. (2015) 'Quantification of GC-biased gene conversion in the human genome', *Genome Research*, (25), pp. 1215–1228.

Glenn, J. K. and Gold, M. H. (1983) 'Decolorization of several polymeric dyes by the lignin-degrading basidiomycete Phanerochaete chrysosporium', *Applied and Environmental Microbiology*, 45(6), pp. 1741–1747.

Goers, L., Freemont, P. and Polizzi, K. M. (2014) 'Co-culture systems and technologies: Taking synthetic biology to the next level', *Journal of the Royal Society Interface*. Royal Society. doi: 10.1098/rsif.2014.0065.

Goffeau, A. *et al.* (1996) 'Life with 6000 genes', *Science*, 274(5287), pp. 546, 563–567.

Gomez, J. A., Höffner, K. and Barton, P. I. (2014) 'DFBAlab: A fast and reliable

MATLAB code for dynamic flux balance analysis', *BMC Bioinformatics*. BioMed Central Ltd., 15(1). doi: 10.1186/s12859-014-0409-8.

Grabherr, M. G. *et al.* (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nature Biotechnology*, 29(7), pp. 644–652. doi: 10.1038/nbt.1883.

Greene, E. R. *et al.* (2015) 'Glycosylation of Cellulases: Engineering Better Enzymes for Biofuels', *Advances in Carbohydrate Chemistry and Biochemistry*. Academic Press, 72, pp. 63–112. doi: 10.1016/BS.ACCB.2015.08.001.

Grigoriev, I. V. *et al.* (2014) 'MycoCosm portal: gearing up for 1000 fungal genomes', *Nucleic Acids Research*. Narnia, 42(D1), pp. D699–D704. doi: 10.1093/nar/gkt1183.

Grinias, J. P. *et al.* (2016) 'An Inexpensive, Open-Source USB Arduino Data Acquisition Device for Chemical Instrumentation', *Journal of Chemical Education*, 93(7), pp. 1316–1319. doi: 10.1021/acs.jchemed.6b00262.

Groisman, A. *et al.* (2005) 'A microfluidic chemostat for experiments with bacterial and yeast cells', *Nature Methods*, 2(9), pp. 685–689. doi: 10.1038/nmeth784.

Groussin, M. *et al.* (2017) 'Unraveling the processes shaping mammalian gut microbiomes over evolutionary time', *Nature Communications*. Nature Publishing Group, 8. doi: 10.1038/ncomms14319.

Gruninger, R. J. *et al.* (2014) 'Anaerobic fungi (phylum Neocallimastigomycota): Advances in understanding their taxonomy, life cycle, ecology, role and biotechnological potential', *FEMS Microbiology Ecology*. Blackwell Publishing Ltd, 90(1), pp. 1–17. doi: 10.1111/1574-6941.12383.

Guerriero, G. *et al.* (2015) 'Destructuring plant biomass: Focus on fungal and

extremophilic cell wall hydrolases', *Plant Science*. Elsevier Ireland Ltd, 234, pp. 180–193. doi: 10.1016/j.plantsci.2015.02.010.

Güllert, S. *et al.* (2016) 'Deep metagenome and metatranscriptome analyses of microbial communities affiliated with an industrial biogas fermenter, a cow rumen, and elephant feces reveal major differences in carbohydrate hydrolysis strategies', *Biotechnology for Biofuels*, 9(1), p. 121. doi: 10.1186/s13068-016-0534-x.

Hackstein, J. H. P. *et al.* (2019) 'Hydrogenosomes of Anaerobic Fungi: An Alternative Way to Adapt to Anaerobic Environments', in. Springer, Cham, pp. 159–175. doi: 10.1007/978-3-030-17941-0_7.

Haghighi Mood, S. *et al.* (2013) 'Lignocellulosic biomass to bioethanol, a comprehensive review with a focus on pretreatment', *Renewable and Sustainable Energy Reviews*, 27, pp. 77–93. doi: 10.1016/j.rser.2013.06.033.

Haitjema, C. H. *et al.* (2014) 'Anaerobic gut fungi: Advances in isolation, culture, and cellulolytic enzyme discovery for biofuel production', *Biotechnology and Bioengineering*, 111(8), pp. 1471–1482. doi: 10.1002/bit.25264.

Haitjema, C. H. *et al.* (2017a) 'A parts list for fungal cellulosomes revealed by comparative genomics', *Nature Microbiology*. Nature Publishing Group, 2(May), pp. 1–8. doi: 10.1038/nmicrobiol.2017.87.

Haitjema, C. H. *et al.* (2017b) 'A parts list for fungal cellulosomes revealed by comparative genomics', *Nature Microbiology*. Nature Publishing Group, 2(May), pp. 1–8. doi: 10.1038/nmicrobiol.2017.87.

Hamilton, W. L. *et al.* (2017) 'Extreme mutation bias and high AT content in Plasmodium falciparum.', *Nucleic acids research*. Oxford University Press, 45(4), pp. 1889–1901. doi:

10.1093/nar/gkw1259.

Hanafy, R. A. *et al.* (2017) '*Pecoramyces ruminantium* , gen. nov., sp. nov., an anaerobic gut fungus from the feces of cattle and sheep', *Mycologia*. Taylor & Francis, 109(2), pp. 231–243. doi: 10.1080/00275514.2017.1317190.

Hanly, T. J. and Henson, M. A. (2011) 'Dynamic flux balance modeling of microbial co-cultures for efficient batch fermentation of glucose and xylose mixtures', *Biotechnology and Bioengineering*, 108(2), pp. 376–385. doi: 10.1002/bit.22954.

Hanly, T. J. and Henson, M. A. (2013) 'Unstructured modeling of a synthetic microbial consortium for consolidated production of ethanol', in *IFAC Proceedings Volumes (IFAC-PapersOnline)*. IFAC Secretariat, pp. 157–162. doi: 10.3182/20131216-3-IN-2044.00003.

Hanly, T. J., Urello, M. and Henson, M. A. (2012) 'Dynamic flux balance modeling of S. cerevisiae and E. coli co-cultures for efficient consumption of glucose/xylose mixtures', *Applied Microbiology and Biotechnology*, 93(6), pp. 2529–2541. doi: 10.1007/s00253-011-3628-1.

Harhangi, H. R. *et al.* (2002) 'A highly expressed family 1 beta-glucosidase with transglycosylation capacity from the anaerobic fungus Piromyces sp. E2.', *Biochimica et biophysica acta*, 1574(3), pp. 293–303.

Hartfield, M. (2016) 'Evolutionary genetic consequences of facultative sex and outcrossing', *Journal of Evolutionary Biology*, 29(1), pp. 5–22. doi: 10.1111/jeb.12770.

Hebraud, M. and Fevre, M. (1990a) 'Purification and characterization of an aspecific glycoside hydrolase from the anaerobic ruminal fungus Neocallimastix frontalis.', *Applied and environmental microbiology*, 56(10), pp. 3164–3169.

Hebraud, M. and Fevre, M. (1990b) 'Purification and characterization of an extracellular

beta-xylosidase from the rumen anaerobic fungus Neocallimastix frontalis.', *FEMS microbiology letters*, 60(1–2), pp. 11–16.

Heirendt, L. *et al.* (2019) 'Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0', *Nature Protocols*. Nature Publishing Group, 14(3), pp. 639–702. doi: 10.1038/s41596-018-0098-2.

Henry, C. S. *et al.* (2010) 'High-throughput generation, optimization and analysis of genome-scale metabolic models', *Nature Biotechnology*, 28(9), pp. 977–982. doi: 10.1038/nbt.1672.

Henske, J. K., Gilmore, S. P., *et al.* (2018) 'Biomass-degrading enzymes are catabolite repressed in anaerobic gut fungi', *AIChE Journal*, 64(12), pp. 4263–4270. doi: 10.1002/aic.16395.

Henske, J. K., Wilken, S. E., *et al.* (2018) 'Metabolic characterization of anaerobic fungi provides a path forward for bioprocessing of crude lignocellulose', *Biotechnology and Bioengineering*, 115(4), pp. 874–884. doi: 10.1002/bit.26515.

Henson, M. A. and Hanly, T. J. (2014) 'Dynamic flux balance analysis for synthetic microbial communities', *IET Systems Biology*. Institution of Engineering and Technology, 8(5), pp. 214–229. doi: 10.1049/iet-syb.2013.0021.

Hershberg, R. and Petrov, D. (2010) 'Evidence that mutation is universally biased towards AT in bacteria', *PLoS Genetics*, 6(9), p. e1001115.

Hess, M. *et al.* (2011) 'Metagenomic discovery of biomass-degrading genes and genomes from cow rumen', *Science*, 331(6016), pp. 463–467. doi: 10.1126/science.1200387.

Hibbett, D. S. *et al.* (2007) 'A higher-level phylogenetic classification of the Fungi', *Mycological Research*, 111(5), pp. 509–547. doi: 10.1016/j.mycres.2007.03.004.

Hildebrand, F., Meyer, A. and Eyre-Walker, A. (2010) 'Evidence of selection upon genomic GC-content in bacteria', *PLoS Genetics*, 6(9), p. e1001107.

Himmel, M. E. *et al.* (2007) 'Biomass recalcitrance: Engineering plants and enzymes for biofuels production', *Science*, 454, pp. 804–807. doi: 10.1126/science.1137016.

Hjersted, J. L. and Henson, M. A. (2009) 'Steady-state and dynamic flux balance analysis of ethanol production by Saccharomyces cerevisiae', *IET Systems Biology*, 3(3), pp. 167–179. doi: 10.1049/iet-syb.2008.0103.

Höffner, K. and Barton, P. I. (2014) 'Design of microbial consortia for industrial biotechnology', in *Computer Aided Chemical Engineering*. Elsevier B.V., pp. 65–74. doi: 10.1016/B978-0-444-63433-7.50008-0.

Höffner, K., Harwood, S. M. and Barton, P. I. (2013) 'A reliable simulator for dynamic flux balance analysis', *Biotechnology and Bioengineering*, 110(3), pp. 792–802. doi: 10.1002/bit.24748.

Hofrichter, M. *et al.* (2010) 'New and classic families of secreted fungal heme peroxidases', *Applied Microbiology and Biotechnology*, 87(3), pp. 871–897. doi: 10.1007/s00253-010-2633-0.

Hong, J. *et al.* (2001) 'Cloning of a gene encoding a highly stable endo-beta-1,4-glucanase from Aspergillus niger and its expression in yeast', *Journal of Bioscience and Bioengineering*, 92(5), pp. 434–441.

Horn, S. J. *et al.* (2012) 'Novel enzymes for the degradation of cellulose', *Biotechnology for Biofuels*. doi: 10.1186/1754-6834-5-45.

Houston, K. *et al.* (2016) 'The Plant Cell Wall: A Complex and Dynamic Structure As Revealed by the Responses of Genes under Stress Conditions.', *Frontiers in plant science*, 7,

p. 984. doi: 10.3389/fpls.2016.00984.

Hrdy, I. *et al.* (2004) ‘Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I’, *Nature*. Nature Publishing Group, 432(7017), pp. 618–622. doi: 10.1038/nature03149.

Hull, C. M., Raisner, R. M. and Johnson, A. D. (2000) ‘Evidence for mating of the “asexual” yeast Candida albicans in a mammalian host’, *Science*, 289(5477), pp. 307–310. doi: 10.1126/science.289.5477.307.

Ilmberger, N. (2013) ‘Cellulases in ionic liquids-the long term stability of Aspergillus sp. cellulase’, *Catalysts*, 3, pp. 584–587.

Innes, M. *et al.* (2018) ‘Fashionable Modelling with Flux’, *arXiv*. Available at: http://arxiv.org/abs/1811.01457 (Accessed: 8 August 2020).

Innes, M. (2018) ‘Flux: Elegant machine learning with Julia’, *Journal of open source software*. doi: 10.21105/joss.00602.

James, T. Y. *et al.* (2013) ‘Shared Signatures of Parasitism and Phylogenomics Unite Cryptomycota and Microsporidia’, *Current Biology*, 23(16), pp. 1548–1553. doi: 10.1016/j.cub.2013.06.057.

Jiang, W. *et al.* (2013) ‘RNA-guided editing of bacterial genomes using CRISPR-Cas systems’, *Nature Biotechnology*. Nature Publishing Group, 31(3), pp. 233–239. doi: 10.1038/nbt.2508.

Jones, C. M., Hernández Lozada, N. J. and Pfleger, B. F. (2015) ‘Efflux systems in bacteria and their metabolic engineering applications’, *Applied Microbiology and Biotechnology*. Springer Verlag, pp. 9381–9393. doi: 10.1007/s00253-015-6963-9.

Jorgensen, H., Kristensen, J. B. and Felby, C. (2007) ‘Enzymatic conversion of

lignocellulose into fermentable sugars: challenges and opportunities', *Biofuels, Bioproducts and Biorefining*, 1(3), pp. 119–134. doi: 10.1002/bbb.

Junker, B. H. *et al.* (1994) 'On-line and in-situ monitoring technology for cell density measurement in microbial and animal cell cultures', *Bioprocess Engineering*. Springer-Verlag, 10(5–6), pp. 195–207. doi: 10.1007/BF00369530.

Kämper, J. *et al.* (2006) 'Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis', *Nature*, 444(7115), pp. 97–101. doi: 10.1038/nature05248.

Kanaly, R. A. and Harayama, S. (2000) 'Biodegradation of high-molecular-weight polycyclic aromatic hydrocarbons by bacteria', *Journal of Bacteriology*, 182(8), pp. 2059–2067. doi: 10.1128/JB.182.8.2059-2067.2000.

Kanehisa, M. *et al.* (2016) 'KEGG as a reference resource for gene and protein annotation', *Nucleic Acids Research*. Oxford University Press, 44(D1), pp. D457–D462. doi: 10.1093/nar/gkv1070.

Kang, D. *et al.* (2019) 'MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies'. PeerJ Inc. doi: 10.7287/peerj.preprints.27522v1.

Karlin, S. *et al.* (2002) 'Amino acid runs in eukaryotic proteomes and disease associations.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 99(1), pp. 333–8. doi: 10.1073/pnas.012608599.

Karnin, E. D. (1990) 'A Simple Procedure for Pruning Back-Propagation Trained Neural Networks', *IEEE Transactions on Neural Networks*, pp. 239–242. doi: 10.1109/72.80236.

Karplus, K., Barrett, C. and Hughey, R. (1998) 'Hidden Markov models for detecting remote protein homologies.', *Bioinformatics*, 14(10), pp. 846–856. doi:

10.1093/bioinformatics/14.10.846.

Kell, D. B. *et al.* (2015) ‘Membrane transporter engineering in industrial biotechnology and whole cell biocatalysis’, *Trends in Biotechnology*. Elsevier Ltd, pp. 237–246. doi: 10.1016/j.tibtech.2015.02.001.

Kern, H. W. and Kirk, T. K. (1987) ‘Influence of molecular size and ligninase pretreatment on degradation of lignins by Xanthomonas sp. strain 99’, *Applied and Environmental Microbiology*, 53(9), pp. 2242–2246.

Kerr, T. J., Kerr, R. D. and Benner, R. (1983) ‘Isolation of a bacterium capable of degrading peanut hull lignin’, *Applied and Environmental Microbiology*, 46(5), pp. 1201–1206.

Khodayari, A. *et al.* (2014) ‘A kinetic model of Escherichia coli core metabolism satisfying multiple sets of mutant flux data.’, *Metabolic engineering*. Elsevier, 25, pp. 50–62. doi: 10.1016/j.ymben.2014.05.014.

Kiktev, D. A. *et al.* (2018) ‘GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae’, *Proceedings of the National Academy of Sciences*, 115(30), pp. E7109–E7118. doi: 10.1073/pnas.1807334115.

Kim, H. *et al.* (2014) ‘Analysis of cellodextrin transporters from Neurospora crassa in Saccharomyces cerevisiae for cellobiose fermentation’, *Applied Microbiology and Biotechnology*. Springer Verlag, 98(3), pp. 1087–1094. doi: 10.1007/s00253-013-5339-2.

King, Z. A. *et al.* (2015) ‘Next-generation genome-scale models for metabolic engineering’, *Current Opinion in Biotechnology*. Elsevier Ltd, pp. 23–29. doi: 10.1016/j.copbio.2014.12.016.

King, Z. A. *et al.* (2016) ‘BiGG Models: A platform for integrating, standardizing and

sharing genome-scale models', *Nucleic Acids Research*. Oxford University Press, 44(D1), pp. D515–D522. doi: 10.1093/nar/gkv1049.

Klein-Marcuschamer, D. *et al.* (2012) 'The challenge of enzyme cost in the production of lignocellulosic biofuels', *Biotechnology and Bioengineering*, 109(4), pp. 1083–1087. doi: 10.1002/bit.24370.

Klein, T., Schneider, K. and Heinzle, E. (2013) 'A system of miniaturized stirred bioreactors for parallel continuous cultivation of yeast with online measurement of dissolved oxygen and off-gas', *Biotechnology and Bioengineering*, 110(2), pp. 535–542. doi: 10.1002/bit.24633.

Kleinstiver, Benjamin P *et al.* (2015) 'Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition', *Nature Biotechnology*. Nature Publishing Group, 33(12), pp. 1293–1298. doi: 10.1038/nbt.3404.

Kleinstiver, Benjamin P. *et al.* (2015) 'Engineered CRISPR-Cas9 nucleases with altered PAM specificities', *Nature*. Nature Publishing Group, 523(7561), pp. 481–485. doi: 10.1038/nature14592.

Knauer, R. and Lehle, L. (1999) 'The oligosaccharyltransferase complex from yeast', *Biochimica et Biophysica Acta (BBA) - General Subjects*. Elsevier, 1426(2), pp. 259–273. doi: 10.1016/S0304-4165(98)00128-7.

Knight, R. D., Freeland, S. J. and Landweber, L. F. (2001) 'A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes', *Genome Biology*. BioMed Central, 2(4), p. research0010.1. doi: 10.1186/gb-2001-2-4-research0010.

Kohler, P. R. A. and Metcalf, W. W. (2012) 'Genetic manipulation of Methanosarcina

spp.', *Frontiers in Microbiology*. Frontiers Research Foundation. doi: 10.3389/fmicb.2012.00259.

Komar, A. A. (2016) 'The Yin and Yang of codon usage', *Human Molecular Genetics*, 25(R2), pp. R77–R85. doi: 10.1093/hmg/ddw207.

Krogh, A. *et al.* (2001) 'Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes', *Journal of Molecular Biology*. Academic Press, 305(3), pp. 567–580. doi: 10.1006/jmbi.2000.4315.

Kuhad, R. C., Gupta, R. and Singh, A. (2011) 'Microbial cellulases and their industrial applications.', *Enzyme research*, 2011, p. 280696. doi: 10.4061/2011/280696.

Kuyper, M. *et al.* (2003) 'High-level functional expression of a fungal xylose isomerase: The key to efficient ethanolic fermentation of xylose by Saccharomyces cerevisiae?', *FEMS Yeast Research*, 4(1), pp. 69–78. doi: 10.1016/S1567-1356(03)00141-7.

Lamed, R. *et al.* (1985) 'Major characteristics of the cellulolytic system of Clostridium thermocellum coincide with those of the purified cellulosome', *Enzyme and Microbial Technology*, 7(1), pp. 37–41. doi: 10.1016/0141-0229(85)90008-0.

Langston, J. A. *et al.* (2011) 'Oxidoreductive cellulose depolymerization by the enzymes cellobiose dehydrogenase and glycoside hydrolase 61', *Applied and Environmental Microbiology*, 77(19), pp. 7007–7015. doi: 10.1128/AEM.05815-11.

Lankiewicz, T. S., Cottrell, M. T. and Kirchman, D. L. (2016) 'Growth rates and rRNA content of four marine bacteria in pure cultures and in the Delaware estuary', *ISME Journal*. Nature Publishing Group, 10(4), pp. 823–832. doi: 10.1038/ismej.2015.156.

Leberer, E. *et al.* (1996) 'Signal transduction through homologs of the Ste20p and Ste7p protein kinases can trigger hyphal formation in the pathogenic fungus Candida albicans.',

Proceedings of the National Academy of Sciences of the United States of America, 93(23), pp. 13217–22.

Lee, J. M. *et al.* (1993) 'Cloning of a xylanase gene from the ruminal fungus Neocallimastix patriciarum 27 and its expression in Escherichia coli.', *Canadian journal of microbiology*, 39(1), pp. 134–139.

Levasseur, A. *et al.* (2013) 'Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes', *Biotechnology for Biofuels*, 6(1), p. 41. doi: 10.1186/1754-6834-6-41.

Li, X.-L. *et al.* (2004) 'Properties of a recombinant beta-glucosidase from polycentric anaerobic fungus Orpinomyces PC-2 and its application for cellulose hydrolysis.', *Applied biochemistry and biotechnology*, 113–116, pp. 233–250.

Li, X. and Calza, R. E. (1991) 'Purification and characterization of an extracellular β-glucosidase from the rumen fungus Neocallimastix frontalis EB188', *Enzyme and Microbial Technology*, 13(8), pp. 622–628. doi: 10.1016/0141-0229(91)90075-L.

Li, X. L. *et al.* (2007) 'Expression of an AT-rich xylanase gene from the anaerobic fungus Orpinomyces sp. strain PC-2 in and secretion of the heterologous enzyme by Hypocrea jecorina', *Applied Microbiology and Biotechnology*, 74(6), pp. 1264–1275. doi: 10.1007/s00253-006-0787-6.

Li, X. L., Chen, H. and Ljungdahl, L. G. (1997) 'Two cellulases, CelA and CelC, from the polycentric anaerobic fungus Orpinomyces strain PC-2 contain N-terminal docking domains for a cellulase-hemicellulase complex.', *Applied and Environmental Microbiology*, 63(12), pp. 4721–4728.

Li, Yuanfei *et al.* (2019) 'Combined Genomic, Transcriptomic, Proteomic, and

Physiological Characterization of the Growth of Pecoramyces sp. F1 in Monoculture and Co-culture With a Syntrophic Methanogen', *Frontiers in Microbiology*. Frontiers Media S.A., 10(MAR), p. 435. doi: 10.3389/fmicb.2019.00435.

Liao, J. C. *et al.* (2016) 'Fuelling the future: Microbial engineering for the production of sustainable biofuels', *Nature Reviews Microbiology*. Nature Publishing Group, pp. 288–304. doi: 10.1038/nrmicro.2016.32.

Lieven, C. *et al.* (2020) 'MEMOTE for standardized genome-scale metabolic model testing', *Nature Biotechnology*. Nature Research, pp. 272–276. doi: 10.1038/s41587-020-0446-y.

Liggenstoffer, A. S. *et al.* (2010) 'Phylogenetic diversity and community structure of anaerobic gut fungi (phylum Neocallimastigomycota) in ruminant and non-ruminant herbivores.', *The ISME journal*, 4(10), pp. 1225–35. doi: 10.1038/ismej.2010.49.

Lindmark, D. G. and Müller, M. (1973) 'Hydrogenosome, a Cytoplasmic Organelle of the Anaerobic Flagellate Tritrichomonas foetus, and Its Role in Pvruvate Metabolism', *THE JOURNAL OF BIOLOGICAL CHEMISTRY*, 248(22), p. 1073. Available at: http://www.jbc.org/ (Accessed: 9 August 2020).

Lindmark, D. G. and Müller, M. (1974) 'Biochemical Cytology of Trichomonad Flagellates. II. Subcellular Distribution of Oxidoreductases and Hydrolases in Monocercomonas sp.*', *The Journal of Protozoology*. John Wiley & Sons, Ltd, 21(2), pp. 374–378. doi: 10.1111/j.1550-7408.1974.tb03673.x.

Liu, G., Zhang, J. and Bao, J. (2016) 'Cost evaluation of cellulase enzyme for industrial-scale cellulosic ethanol production based on rigorous Aspen Plus modeling', *Bioprocess and Biosystems Engineering*. Springer Berlin Heidelberg, 39(1), pp. 133–140. doi:

10.1007/s00449-015-1497-1.

Liu, H. *et al.* (2018) ‘Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias’, *Nature Ecology and Evolution*. Springer US, 2(1), pp. 164–173. doi: 10.1038/s41559-017-0372-7.

Liu, J.-H. *et al.* (1997) ‘Plant seed oil-bodies as an immobilization matrix for a recombinant xylanase from the rumen fungus Neocallimastix patriciarum’, *Molecular Breeding*. Kluwer Academic Publishers, 3(6), pp. 463–470. doi: 10.1023/A:1009604119618.

Liu, J.-R. *et al.* (2005) ‘Direct cloning of a xylanase gene from the mixed genomic DNA of rumen fungi and its expression in intestinal Lactobacillus reuteri.’, *FEMS microbiology letters*, 251(2), pp. 233–241. doi: 10.1016/j.femsle.2005.08.008.

Liu, N., Qiao, K. and Stephanopoulos, G. (2016) ‘13C Metabolic Flux Analysis of acetate conversion to lipids by Yarrowia lipolytica’, *Metabolic Engineering*. Academic Press Inc., 38, pp. 86–97. doi: 10.1016/j.ymben.2016.06.006.

Lombard, V. *et al.* (2014) ‘The carbohydrate-active enzymes database (CAZy) in 2013’, *Nucleic Acids Research*, 42(D1). doi: 10.1093/nar/gkt1178.

Long, C. P. and Antoniewicz, M. R. (2019) ‘High-resolution 13C metabolic flux analysis’, *Nature Protocols*. Nature Publishing Group, 14(10), pp. 2856–2877. doi: 10.1038/s41596-019-0204-0.

Lowe, S. E., Theodorou, M. K. and Trinci, A. P. (1987a) ‘Cellulases and xylanase of an anaerobic rumen fungus grown on wheat straw, wheat straw holocellulose, cellulose, and xylan.’, *Applied and environmental microbiology*. American Society for Microbiology (ASM), 53(6), pp. 1216–1223.

Lowe, S. E., Theodorou, M. K. and Trinci, A. P. (1987b) ‘Growth and fermentation of an

anaerobic rumen fungus on various carbon sources and effect of temperature on development.', *Applied and environmental microbiology*, 53(6), pp. 1210–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3606103 (Accessed: 17 January 2020).

Lu, H. *et al.* (2017) 'Comprehensive reconstruction and in silico analysis of *Aspergillus niger* genome-scale metabolic network model that accounts for 1210 ORFs', *Biotechnology and Bioengineering*. John Wiley & Sons, Ltd, 114(3), pp. 685–695. doi: 10.1002/bit.26195.

Lynd, L. R. *et al.* (2005) 'Consolidated bioprocessing of cellulosic biomass: An update', *Current Opinion in Biotechnology*, pp. 577–583. doi: 10.1016/j.copbio.2005.08.009.

Machado, D. *et al.* (2018) 'Fast automated reconstruction of genome-scale metabolic models for microbial species and communities', *Nucleic Acids Research*, 46(15), pp. 7542–7553. doi: 10.1093/nar/gky537.

Magee, B. B. (2002) 'Induction of Mating in Candida albicans by Construction of MTLa and MTLalpha Strains', *Science*, 289(5477), pp. 310–313. doi: 10.1126/science.289.5477.310.

Mahadevan, R., Edwards, J. S. and Doyle, F. J. (2002) 'Dynamic Flux Balance Analysis of diauxic growth in Escherichia coli', *Biophysical Journal*. Biophysical Society, 83(3), pp. 1331–1340. doi: 10.1016/S0006-3495(02)73903-9.

Mahadevan, R. and Henson, M. A. (2012) 'Genome-based modeling and design of metabolic interactions in microbial communities', *Computational and Structural Biotechnology Journal*. Research Network of Computational and Structural Biotechnology, p. e201210008. doi: 10.5936/csbj.201210008.

Mahadevan, R. and Schilling, C. H. (2003) 'The effects of alternate optimal solutions in constraint-based genome-scale metabolic models', *Metabolic Engineering*. Academic Press

Inc., 5(4), pp. 264–276. doi: 10.1016/j.ymben.2003.09.002.

Mantovani, C. F., Geimba, M. P. and Brandelli, A. (2005) 'Enzymatic clarification of fruit juices by fungal pectin lyase', *Food Biotechnology*, 19(3), pp. 173–181. doi: 10.1080/08905430500316284.

Martinez, D. *et al.* (2008) 'Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina)', *Nature Biotechnology*, 26(5), pp. 553–560. doi: 10.1038/nbt1403.

Martinez, D. *et al.* (2009) 'Genome, transcriptome, and secretome analysis of wood decay fungus *Postia placenta* supports unique mechanisms of lignocellulose conversion', *Proceedings of the National Academy of Sciences*, 106(6), pp. 1954–1959. doi: 10.1073/pnas.0809575106.

Martínez, I. *et al.* (2008) 'Replacing Escherichia coli NAD-dependent glyceraldehyde 3-phosphate dehydrogenase (GAPDH) with a NADP-dependent enzyme from Clostridium acetobutylicum facilitates NADPH dependent pathways', *Metabolic Engineering*. Academic Press, 10(6), pp. 352–359. doi: 10.1016/j.ymben.2008.09.001.

Marvin-Sikkema, F. D. *et al.* (1993) 'Characterization of hydrogenosomes and their role in glucose metabolism of Neocallimastix sp. L2', *Archives of Microbiology*. Springer-Verlag, 160(5), pp. 388–396. doi: 10.1007/BF00252226.

Marvin-Sikkema, F. D. *et al.* (1994a) 'Metabolic energy generation in hydrogenosomes of the anaerobic fungus Neocallimastix: evidence for a functional relationship with mitochondria', *Mycological Research*. Elsevier, 98(2), pp. 205–212. doi: 10.1016/S0953-7562(09)80187-1.

Marvin-Sikkema, F. D. *et al.* (1994b) 'Metabolic energy generation in hydrogenosomes

of the anaerobic fungus Neocallimastix: evidence for a functional relationship with mitochondria', *Mycological Research*, 98(2), pp. 205–212. doi: 10.1016/S0953-7562(09)80187-1.

McGinnis, S. and Madden, T. L. (2004) 'BLAST: at the core of a powerful and diverse set of sequence analysis tools', *Nucleic Acids Research*, 32(suppl 2), pp. W20–W25. doi: 10.1093/nar/gkh435.

Mee, M. T. *et al.* (2014) 'Syntrophic exchange in synthetic microbial communities', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 111(20), pp. E2149–E2156. doi: 10.1073/pnas.1405641111.

Meerupati, T. *et al.* (2013) 'Genomic mechanisms accounting for the adaptation to parasitism in nematode-trapping fungi', *PLoS Genet*, 9(11), p. e1003909. doi: 10.1371/journal.pgen.1003909.

Mertens, S. *et al.* (2015) 'A Large Set of Newly Created Interspecific Saccharomyces Hybrids Increases Aromatic Diversity in Lager Beers', *Applied and Environmental Microbiology*, 81(23), pp. 8202–8214. doi: 10.1128/AEM.02464-15.Editor.

Meunier, J. and Duret, L. (2004) 'Recombination drives the evolution of GC-content in the human genome', *Molecular Biology and Evolution*, 21(6), pp. 984–990. doi: 10.1093/molbev/msh070.

Mih, N. and Palsson, B. O. (2019) 'Expanding the uses of genome-scale models with protein structures', *Molecular Systems Biology*, 15(11). doi: 10.15252/msb.20188601.

Minty, J. J. *et al.* (2013) 'Design and characterization of synthetic fungal-bacterial consortia for direct production of isobutanol from cellulosic biomass', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of

Sciences, 110(36), pp. 14592–14597. doi: 10.1073/pnas.1218447110.

Mo, M. L., Palsson, B. and Herrgård, M. J. (2009) 'Connecting extracellular metabolomic measurements to intracellular flux states in yeast', *BMC Systems Biology*, 3. doi: 10.1186/1752-0509-3-37.

Monciardini, P. *et al.* (2014) 'Discovering new bioactive molecules from microbial sources', *Microbial Biotechnology*, 7(3), pp. 209–220. doi: 10.1111/1751-7915.12123.

Mondo, S. J. *et al.* (2017) 'Widespread adenine N6-methylation of active genes in fungi', *Nature Genetics*, 49(6), pp. 964–968. doi: 10.1038/ng.3859.

Monk, J. M. *et al.* (2017) 'iML1515, a knowledgebase that computes Escherichia coli traits', *Nature Biotechnology*. Nature Publishing Group, pp. 904–908. doi: 10.1038/nbt.3956.

Morrison, J. M., Elshahed, M. S. and Youssef, N. H. (2016) 'Defined enzyme cocktail from the anaerobic fungus Orpinomyces sp. strain C1A effectively releases sugars from pretreated corn stover and switchgrass', *Scientific Reports*. Nature Publishing Group, 6(May), p. 29217. doi: 10.1038/srep29217.

Mountfort, D. O. and Asher, R. A. (1985) 'Production and regulation of cellulase by two strains of the rumen anaerobic fungus Neocallimastix frontalis.', *Applied and environmental microbiology*, 49(5), pp. 1314–1322.

Mukherjee, S. *et al.* (2017) '1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life', *Nature Biotechnology*. Nature Publishing Group, 35(7), pp. 676–683. doi: 10.1038/nbt.3886.

Muller, M. (1993) 'Review Article: The hydrogenosome', *Journal of General Microbiology*. Microbiology Society, 139(12), pp. 2879–2889. doi: 10.1099/00221287-139-12-2879.

Muller, M. *et al.* (2012) ‘Biochemistry and Evolution of Anaerobic Energy Metabolism in Eukaryotes’, *Microbiology and Molecular Biology Reviews*. American Society for Microbiology, 76(2), pp. 444–495. doi: 10.1128/mmbr.05024-11.

Müller, V., Chowdhury, N. P. and Basen, M. (2018a) ‘Electron Bifurcation: A Long-Hidden Energy-Coupling Mechanism’, *Annual Review of Microbiology*. Annual Reviews, 72(1), pp. 331–353. doi: 10.1146/annurev-micro-090816-093440.

Müller, V., Chowdhury, N. P. and Basen, M. (2018b) ‘Electron Bifurcation: A Long-Hidden Energy-Coupling Mechanism’, *Annual Review of Microbiology*. Annual Reviews , 72(1), pp. 331–353. doi: 10.1146/annurev-micro-090816-093440.

Murphy, C. L. *et al.* (2019) ‘Horizontal Gene Transfer as an Indispensable Driver for Evolution of Neocallimastigomycota into a Distinct Gut-Dwelling Fungal Lineage’, *Applied and Environmental Microbiology*. American Society for Microbiology, 85(15). doi: 10.1128/aem.00988-19.

Nagarajan, H. *et al.* (2013) ‘Characterizing acetogenic metabolism using a genome-scale metabolic reconstruction of Clostridium ljungdahlii’, *Microbial Cell Factories*. BioMed Central Ltd., 12(1). doi: 10.1186/1475-2859-12-118.

Nicholson, M. J., Theodorou, M. K. and Brookman, J. L. (2005) ‘Molecular analysis of the anaerobic rumen fungus Orpinomyces - Insights into an AT-rich genome’, *Microbiology*, 151(1), pp. 121–133. doi: 10.1099/mic.0.27353-0.

Nierman, W. C. *et al.* (2005) ‘Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus’, *Nature*, 438(7071), pp. 1151–1156. doi: 10.1038/nature04332.

Nieuwenhuis, B. P. S. and James, T. Y. (2016) ‘The frequency of sex in fungi’,

Philosophical Transactions of the Royal Society B: Biological Sciences. doi: 10.1098/rstb.2015.0540.

Nørholm, M. H. H. (2019) 'Meta synthetic biology: controlling the evolution of engineered living systems.', *Microbiol Biotechnol*, 12(1), pp. 35–37.

Nurk, S. *et al.* (2017) 'MetaSPAdes: A new versatile metagenomic assembler', *Genome Research*. Cold Spring Harbor Laboratory Press, 27(5), pp. 824–834. doi: 10.1101/gr.213959.116.

O'Leary, N. A. *et al.* (2016) 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research*, 44(D1), pp. D733–D745. doi: 10.1093/nar/gkv1189.

O'Malley, M. A., Theodorou, M. K. and Kaiser, C. A. (2012a) 'Evaluating expression and catalytic activity of anaerobic fungal fibrolytic enzymes cative to Piromyces sp E2 in Saccharomyces cerevisiae', *Environmental Progress and Sustainable Energy*, 31(1), pp. 37–46. doi: 10.1002/ep.

O'Malley, M. A., Theodorou, M. K. and Kaiser, C. A. (2012b) 'Evaluating expression and catalytic activity of anaerobic fungal fibrolytic enzymes native topiromyces sp E2 inSaccharomyces cerevisiae', *Environmental Progress & Sustainable Energy*. John Wiley & Sons, Ltd, 31(1), pp. 37–46. doi: 10.1002/ep.10614.

O'Malley, M. A., Theodorou, M. K. and Kaiser, C. A. (2012c) 'Evaluating expression and catalytic activity of anaerobic fungal fibrolytic enzymes native topiromyces sp E2 inSaccharomyces cerevisiae', *Environmental Progress & Sustainable Energy*. John Wiley & Sons, Ltd, 31(1), pp. 37–46. doi: 10.1002/ep.10614.

Obembe, O. O. *et al.* (2007) 'Promiscuous, non-catalytic, tandem carbohydrate-binding

modules modulate the cell-wall structure and development of transgenic tobacco (Nicotiana tabacum) plants.', *Journal of plant research*, 120(5), pp. 605–617. doi: 10.1007/s10265-007-0099-7.

Ohm, R. A. *et al.* (2012) 'Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi', *PLoS Pathog*, 8(12), p. e1003037. doi: 10.1371/journal.ppat.1003037.

Ohm, R. A. *et al.* (2014) 'Genomics of wood-degrading fungi', *Fungal Genetics and Biology*, 72, pp. 82–90. doi: 10.1016/j.fgb.2014.05.001.

Olm, M. R. *et al.* (2017) 'DRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication', *ISME Journal*. Nature Publishing Group, 11(12), pp. 2864–2868. doi: 10.1038/ismej.2017.126.

Orpin, C. G. (1975) 'Studies on the rumen flagellate Neocallimastix frontalis.', *Journal of general microbiology*, 91(2), pp. 249–62. doi: 10.1099/00221287-91-2-249.

Orpin, C. G. (1977) 'The occurrence of chitin in the cell walls of the rumen organisms Neocallimastix frontalis, Piromonas communis and Sphaeromonas communis.', *Journal of general microbiology*, 99(1), pp. 215–218. doi: 10.1099/00221287-99-1-215.

Orpin, C. G. and Greenwood, Y. (1986) 'The role of haems and related compounds in the nutrition and zoosporogenesis of the rumen chytridiomycete Neocallimastix frontalis H8', *Journal of General Microbiology*. Microbiology Society, 132(8), pp. 2179–2185. doi: 10.1099/00221287-132-8-2179.

Orth, J. D., Thiele, I. and Palsson, B. O. (2010) 'What is flux balance analysis?', *Nature Biotechnology*, pp. 245–248. doi: 10.1038/nbt.1614.

Otero, J. M. J. M. and Nielsen, J. (2010) *Industrial systems biology*, *Biotechnology and*

Bioengineering. doi: 10.1002/bit.22592.

Oyola, S. O. *et al.* (2012) 'Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes', *BMC Genomics*. BioMed Central, 13(1), p. 1. doi: 10.1186/1471-2164-13-1.

Pabón Pereira, C. P., Castañares, G. and Van Lier, J. B. (2012) 'An OxiTop® protocol for screening plant material for its biochemical methane potential (BMP)', *Water Science and Technology*. IWA Publishing, 66(7), pp. 1416–1423. doi: 10.2166/wst.2012.305.

Paloheimo, M. *et al.* (2016) 'Production of Industrial Enzymes in Trichoderma reesei', in, pp. 23–57. doi: 10.1007/978-3-319-27951-0_2.

Pan, S. *et al.* (2017) 'Model-enabled gene search (MEGS) allows fast and direct discovery of enzymatic and transport gene functions in the marine bacterium Vibrio fischeri', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., 292(24), pp. 10250–10261. doi: 10.1074/jbc.M116.763193.

Parks, D. H. *et al.* (2015) 'CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome Research*. Cold Spring Harbor Laboratory Press, 25(7), pp. 1043–1055. doi: 10.1101/gr.186072.114.

Payne, C. M. *et al.* (2015) 'Fungal cellulases', *Chemical Reviews*. American Chemical Society, 115(3), pp. 1308–1448. doi: 10.1021/cr500351c.

Pearce, P. D. and Bauchop, T. (1985) 'Glycosidases of the rumen anaerobic fungus Neocallimastix frontalis grown on cellulosic substrates.', *Applied and environmental microbiology*, 49(5), pp. 1265–1269.

Peng, X. *et al.* (2018) 'Methods for genomic characterization and maintenance of anaerobic fungi', in *Methods in Molecular Biology*. Humana Press Inc., pp. 53–67. doi:

10.1007/978-1-4939-7804-5_5.

Peng, X. “Nick”, Gilmore, S. P. and O’Malley, M. A. (2016) ‘Microbial communities for bioprocessing: lessons learned from nature’, *Current Opinion in Chemical Engineering*. Elsevier Ltd, pp. 103–109. doi: 10.1016/j.coche.2016.09.003.

Petersen, T. N. *et al.* (2011) ‘SignalP 4.0: Discriminating signal peptides from transmembrane regions’, *Nature Methods*, pp. 785–786. doi: 10.1038/nmeth.1701.

Piao, H. *et al.* (2014) ‘Identification of novel biomass-degrading enzymes from genomic dark matter: Populating genomic sequence space with functional annotation’, *Biotechnology and Bioengineering*, 111(8), pp. 1550–1565. doi: 10.1002/bit.25250.

Podolsky, I. *et al.* (2019) ‘Harnessing Nature’s Anaerobes for Biotechnology and Bioprocessing.’, *Annual Review of Chemical and Biomolecular Engineering*. doi: https://doi.org/10.1146/annurev-chembioeng-060718-030340.

Poidevin, L. *et al.* (2009) ‘Heterologous production of the Piromyces equi cinnamoyl esterase in Trichoderma reesei for biotechnological applications.’, *Letters in applied microbiology*, 49(6), pp. 673–678. doi: 10.1111/j.1472-765X.2009.02734.x.

Qiu, J. and Jin, X. (2002) ‘Development and optimization of organic acid analysis in tobacco with ion chromatography and suppressed conductivity detection’, *Journal of Chromatography A*, 950(1–2), pp. 81–88. doi: 10.1016/S0021-9673(02)00034-1.

Ramanjaneyulu, G. and Rajasekhar Reddy, B. (2016) ‘Optimization of xylanase production through response surface methodology by Fusarium sp. BVKT R2 isolated from forest soil and its application in saccharification.’, *Frontiers in microbiology*, 7, p. 1450. doi: 10.3389/fmicb.2016.01450.

Ranganathan, A. *et al.* (2017) ‘Utilizing Anaerobic Fungi for Two-stage Sugar Extraction

and Biofuel Production from Lignocellulosic Biomass', *Frontiers in Microbiology*. Frontiers Research Foundation, 8(APR), p. 635. doi: 10.3389/fmicb.2017.00635.

Raymond, M. *et al.* (1998) 'A Ste6p/P-glycoprotein homologue from the asexual yeast Candida albicans transports the a-factor mating pheromone in Saccharomyces cerevisiae', *Molecular Microbiology*, 27(3), pp. 587–598. doi: 10.1046/j.1365-2958.1998.00704.x.

Reichenberger, E. R. *et al.* (2015) 'Prokaryotic Nucleotide Composition Is Shaped by Both Phylogeny and the Environment', *Genome biology and evolution*, 7(5), pp. 1380–1389.

Resch, M. G. *et al.* (2013) 'Fungal cellulases and complexed cellulosomal enzymes exhibit synergistic mechanisms in cellulose deconstruction', *Energy and Environmental Science*, 6(6), pp. 1858–1867. doi: 10.1039/c3ee00019b.

Reymond, P. *et al.* (1991) 'Molecular cloning of genes from the rumen anaerobic fungus Neocallimastix frontalis: expression during hydrolase induction', *FEMS Microbiology Letters*, 77(1), pp. 107–112. doi: 10.1111/j.1574-6968.1991.tb04330.x.

Reymond, P. *et al.* (1992) 'Sequence of the phosphoenolpyruvate carboxykinase-encoding cDNA from the rumen anaerobic fungus Neocallimastix frontalis: Comparison of the amino acid sequence with animals and yeast', *Gene*, 110(1), pp. 57–63. doi: 10.1016/0378-1119(92)90444-T.

Rhind, N. *et al.* (2011) 'Comparative functional genomics of the fission yeasts', *Science*, 332(6032), pp. 930–936. doi: 10.1126/science.1203357.

Rigden, D. J. (2005) 'Analysis of glycoside hydrolase family 98: Catalytic machinery, mechanism and a novel putative carbohydrate binding module', *FEBS Letters*, 579(25), pp. 5466–5472. doi: 10.1016/j.febslet.2005.09.011.

Riley, R. *et al.* (2014) 'Extensive sampling of basidiomycete genomes demonstrates

inadequacy of the white-rot/brown-rot paradigm for wood decay fungi', *Proceedings of the National Academy of Sciences*, 111(27), pp. 9923–9928. doi: 10.1073/pnas.1400592111.

Rocha-Martin, J. *et al.* (2014) 'Emerging strategies and integrated systems microbiology technologies for biodiscovery of marine bioactive compounds.', *Marine drugs*. Multidisciplinary Digital Publishing Institute (MDPI), 12(6), pp. 3516–3559. doi: 10.3390/md12063516.

Rogers, J. N. *et al.* (2017) 'An assessment of the potential products and economic and environmental impacts resulting from a billion ton bioeconomy', *Biofuels, Bioproducts and Biorefining*. John Wiley & Sons, Ltd, 11(1), pp. 110–128. doi: 10.1002/bbb.1728.

Ropars, J. *et al.* (2016) 'Evidence for the sexual origin of heterokaryosis in arbuscular mycorrhizal fungi', *Nature Microbiology*, 1(6), pp. 1–9. doi: 10.1038/nmicrobiol.2016.33.

Rowe, E., Palsson, B. O. and King, Z. A. (2018) 'Escher-FBA: a web application for interactive flux balance analysis', *BMC systems biology*. NLM (Medline), 12(1), p. 84. doi: 10.1186/s12918-018-0607-5.

Rubin, E. M. (2008) 'Genomics of cellulosic biofuels', *Nature*, 454(7206), pp. 841–845. doi: 10.1038/nature07190.

Saa, P. A. and Nielsen, L. K. (2016) 'Ll-ACHRB: A scalable algorithm for sampling the feasible solution space of metabolic networks', *Bioinformatics*. Oxford University Press, 32(15), pp. 2330–2337. doi: 10.1093/bioinformatics/btw132.

Sadhu, C. *et al.* (1992) 'A G-protein alpha subunit from asexual Candida albicans functions in the mating signal transduction pathway of Saccharomyces cerevisiae and is regulated by the a1-alpha 2 repressor.', *Molecular and Cellular Biology*, 12(5), pp. 1977–1985. doi: 10.1128/mcb.12.5.1977.

Saini, J. K., Saini, R. and Tewari, L. (2015) *Lignocellulosic agriculture wastes as biomass feedstocks for second-generation bioethanol production: concepts and recent developments*, *3 Biotech*. Springer Verlag. doi: 10.1007/s13205-014-0246-5.

Sammond, D. W. *et al.* (2012) 'Cellulase Linkers Are Optimized Based on Domain Type and Function: Insights from Sequence Analysis, Biophysical Measurements, and Molecular Simulation', *PLoS ONE*. Edited by V. Arcus. Public Library of Science, 7(11), p. e48615. doi: 10.1371/journal.pone.0048615.

Sánchez, B. J., Pérez-Correa, J. R. and Agosin, E. (2014) 'Construction of robust dynamic genome-scale metabolic model structures of Saccharomyces cerevisiae through iterative re-parameterization', *Metabolic Engineering*. Academic Press Inc., 25, pp. 159–173. doi: 10.1016/j.ymben.2014.07.004.

Sanderson, K. (2011) 'Lignocellulose: A chewy problem.', *Nature*, 474(7352), pp. S12–S14. doi: 10.1038/474S012a.

Sarik, J. and Kymissis, I. (2010) 'Lab kits using the arduino prototyping platform', in *Proceedings - Frontiers in Education Conference, FIE*. doi: 10.1109/FIE.2010.5673417.

Sarkar, N. *et al.* (2012) 'Bioethanol production from agricultural wastes: An overview', *Renewable Energy*. Elsevier Ltd, 37(1), pp. 19–27. doi: 10.1016/j.renene.2011.06.045.

Scarlat, N. *et al.* (2015) 'The role of biomass and bioenergy in a future bioeconomy: Policies and facts', *Environmental Development*. Elsevier, 15, pp. 3–34. doi: 10.1016/j.envdev.2015.03.006.

Schneider, R. E. *et al.* (2011) 'The Trichomonas vaginalis hydrogenosome proteome is highly reduced relative to mitochondria, yet complex compared with mitosomes', *International Journal for Parasitology*. Australian Society for Parasitology Inc., 41(13–14),

pp. 1421–1434. doi: 10.1016/j.ijpara.2011.10.001.

Schuetz, R., Kuepfer, L. and Sauer, U. (2007) 'Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*', *Molecular Systems Biology*, 3(1), p. 119. doi: 10.1038/msb4100162.

Schut, Gerrit J and Adams, M. W. W. (2009) 'The iron-hydrogenase of Thermotoga maritima utilizes ferredoxin and NADH synergistically: a new perspective on anaerobic hydrogen production.', *Journal of bacteriology*. American Society for Microbiology Journals, 191(13), pp. 4451–7. doi: 10.1128/JB.01582-08.

Schut, Gerrit J. and Adams, M. W. W. (2009) 'The iron-hydrogenase of Thermotoga maritima utilizes ferredoxin and NADH synergistically: A new perspective on anaerobic hydrogen production', *Journal of Bacteriology*. American Society for Microbiology Journals, 191(13), pp. 4451–4457. doi: 10.1128/JB.01582-08.

Schwarz, W. H. (2001) 'The cellulosome and cellulose degradation by anaerobic bacteria', *Applied Microbiology and Biotechnology*, 56(5–6), pp. 634–649.

Segata, N. *et al.* (2013) 'PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes', *Nature Communications*, 4. doi: 10.1038/ncomms3304.

Sekowska, A. *et al.* (2016) 'Generation of mutation hotspots in ageing bacterial colonies', *Scientific Reports*, 6, p. 2.

Senger, R. S., Yen, J. Y. and Fong, S. S. (2014) 'A review of genome-scale metabolic flux modeling of anaerobiosis in biotechnology', *Current Opinion in Chemical Engineering*. Elsevier Ltd, pp. 33–42. doi: 10.1016/j.coche.2014.08.003.

Seppälä, S. *et al.* (2016) 'Mapping the membrane proteome of anaerobic gut fungi identifies a wealth of carbohydrate binding proteins and transporters', *Microbial Cell*

Factories. Edited by A. P. Mitchell. BioMed Central, 15(1), p. 212. doi: 10.1186/s12934-016-0611-7.

Seppälä, S *et al.* (2017) 'The importance of sourcing enzymes from non-conventional fungi for metabolic engineering and biomass breakdown.', *Metabolic Engineering*, 44, pp. 45–59.

Seppälä, Susanna *et al.* (2017) *The importance of sourcing enzymes from non-conventional fungi for metabolic engineering and biomass breakdown*, *Metabolic Engineering*. doi: 10.1016/j.ymben.2017.09.008.

Seppälä, S. *et al.* (2019) 'Heterologous transporters from anaerobic fungi bolster fluoride tolerance in Saccharomyces cerevisiae', *Metabolic Engineering Communications*. Elsevier, 9, p. e00091. doi: 10.1016/J.MEC.2019.E00091.

Seshadri, R. *et al.* (2018) 'Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection', *Nature Biotechnology*, 36(4), pp. 359–367. doi: 10.1038/nbt.4110.

Simeonidis, E. and Price, N. D. (2015) 'Genome-scale modeling for metabolic engineering', *Journal of Industrial Microbiology and Biotechnology*. Springer Verlag, pp. 327–338. doi: 10.1007/s10295-014-1576-3.

Sindhu, R., Binod, P. and Pandey, A. (2016) 'Biological pretreatment of lignocellulosic biomass - An overview', *Bioresource Technology*. Elsevier Ltd, pp. 76–82. doi: 10.1016/j.biortech.2015.08.030.

Singh Arora, D. and Kumar Sharma, R. (2010) 'Ligninolytic fungal laccases and their biotechnological applications', *Applied Biochemistry and Biotechnology*, 160(6), pp. 1760–1788. doi: 10.1007/s12010-009-8676-y.

Smidt, H. *et al.* (2001) '*Clostridium beijerinckii* cells expressing *Neocallimastix patriciarum* glycoside hydrolases show enhanced lichenan utilization and solvent production', *Appl. Environ. Microbiol.*, 67(11), pp. 5127–5133. doi: 10.1128/AEM.67.11.5127.

Soh, L. *et al.* (2014) 'Evaluating microalgal integrated biorefinery schemes: Empirical controlled growth studies and life cycle assessment', *Bioresource Technology*. Elsevier Ltd, 151, pp. 19–27. doi: 10.1016/j.biortech.2013.10.012.

Solden, L. M. *et al.* (2018) 'Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem', *Nature Microbiology*. Nature Publishing Group, 3(11), pp. 1274–1284. doi: 10.1038/s41564-018-0225-4.

Solieri, L. *et al.* (2015) 'Fast method for identifying inter- and intra-species Saccharomyces hybrids in extensive genetic improvement programs based on yeast breeding', *Journal of Applied Microbiology*, 119(1), pp. 149–161. doi: 10.1111/jam.12827.

Solomon, K. V *et al.* (2016) 'Early-branching gut fungi possess a large, comprehensive array of biomass-degrading enzymes', *Science*, 351(6278), pp. 1192–1196. doi: 10.1126/science.aad1431.

Sonan, G. K. *et al.* (2007) 'The linker region plays a key role in the adaptation to cold of the cellulase from an Antarctic bacterium.', *The Biochemical journal*. Portland Press Limited, 407(2), pp. 293–302. doi: 10.1042/BJ20070640.

Sønderby, S. K. *et al.* (2015) 'Convolutional LSTM networks for subcellular localization of proteins', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, pp. 68–80. doi: 10.1007/978-3-319-21233-3_6.

Sonoda, S. and Murata, N. (2017) 'Neural network with unbounded activation functions

is universal approximator', *Applied and Computational Harmonic Analysis*. Academic Press Inc., 43(2), pp. 233–268. doi: 10.1016/j.acha.2015.12.005.

Sørensen, A. *et al.* (2013) 'Fungal beta-glucosidases: A bottleneck in industrial use of lignocellulosic materials', *Biomolecules*, 3(3), pp. 612–631. doi: 10.3390/biom3030612.

Staben, C. and Yanofsky, C. (1990) 'Neurospora crassa a mating-type region (sexual reproduction/vegetative incompatibility/perithecium formation/filamentous fungus)', *Genetics*, 87(July), pp. 4917–4921.

Stairs, C. W., Roger, A. J. and Hampl, V. (2011) 'Eukaryotic Pyruvate Formate Lyase and Its Activating Enzyme Were Acquired Laterally from a Firmicute', *Molecular Biology and Evolution*. Narnia, 28(7), pp. 2087–2099. doi: 10.1093/molbev/msr032.

Steensels, J., Snoek, T., *et al.* (2014) 'Improving industrial yeast strains: exploiting natural and artificial diversity.', *FEMS microbiology reviews*. Wiley-Blackwell, 38(5), pp. 947–95. doi: 10.1111/1574-6976.12073.

Steensels, J., Meersman, E., *et al.* (2014) 'Large-Scale Selection and Breeding To Generate Industrial Yeasts with Superior Aroma Production', *Applied and Environmental Microbiology*, 80(22), pp. 6965–6975. doi: 10.1128/aem.02235-14.

Stewart, R. D. *et al.* (2018) 'Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen', *Nature Communications*. Springer US, 9(1), pp. 1–11. doi: 10.1038/s41467-018-03317-6.

Stewart, R. D. *et al.* (2019) 'Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery', *Nature Biotechnology*. Nature Publishing Group, 37(8), pp. 953–961. doi: 10.1038/s41587-019-0202-3.

Sukumaran, R. K. *et al.* (2009) 'Cellulase production using biomass feed stock and its

application in lignocellulose saccharification for bio-ethanol production', *Renewable Energy*. Pergamon, 34(2), pp. 421–424. doi: 10.1016/J.RENENE.2008.05.008.

Sukumaran, R. K., Singhania, R. R. and Pandey, A. (2005) 'Microbial cellulases - Production, applications and challenges', *Journal of Scientific and Industrial Research*, 64(11), pp. 832–844.

Svartström, O. *et al.* (2017) 'Ninety-nine de novo assembled genomes from the moose (Alces alces) rumen microbiome provide new insights into microbial plant biomass degradation', *ISME Journal*. Nature Publishing Group, 11(11), pp. 2538–2551. doi: 10.1038/ismej.2017.108.

Tagliapietra, F. *et al.* (2010) 'In vitro rumen fermentation: Effect of headspace pressure on the gas production kinetics of corn meal and meadow hay', *Animal Feed Science and Technology*, 158(3–4), pp. 197–201. doi: 10.1016/j.anifeedsci.2010.04.003.

Tedersoo, L. *et al.* (2014) 'Global diversity and geography of soil fungi', *Science*, 346(6213).

Teunissen,', M. J. *et al.* (1991) *Comparison of growth characteristics of anaerobic fungi isolated from ruminant and non-ruminant herbivores during cultivation in a defined medium*, *Journal of General Microbiology*.

Teunissen, M. J. *et al.* (1992) 'Purification and characterization of an extracellular beta-glucosidase from the anaerobic fungus Piromyces sp. strain E2', *Archives of Microbiology*. Springer-Verlag, 158(4), pp. 276–281. doi: 10.1007/BF00245245.

Theodorou, M. K. *et al.* (1995) 'Determination of growth of anaerobic fungi on soluble and cellulosic substrates using a pressure transducer', *Microbiology*. Microbiology Society, 141(3), pp. 671–678. doi: 10.1099/13500872-141-3-671.

Theodorou, M. K. *et al.* (1996) ‘Anaerobic fungi in the digestive tract of mammalian herbivores and their potential for exploitation.’, *The Proceedings of the Nutrition Society*, 55(3), pp. 913–926. doi: 10.1079/PNS19960088.

Theodorou, M. K., Gascoyne, D. J. and Beever, D. E. (1984) ‘The role of consecutive batch culture in rumen microbiology’, *Canadian Journal of Animal Science*, 64(5), pp. 47–48. doi: 10.4141/cjas84-150.

Thiele, I. and Palsson, B. (2010) ‘A protocol for generating a high-quality genome-scale metabolic reconstruction’, *Nature Protocols*, 5(1), pp. 93–121. doi: 10.1038/nprot.2009.203.

Thies, S. *et al.* (2016) ‘Metagenomic discovery of novel enzymes and biosurfactants in a slaughterhouse biofilm microbial community.’, *Scientific reports*. Nature Publishing Group, 6, p. 27035. doi: 10.1038/srep27035.

Tisserant, E. *et al.* (2013) ‘Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis’, *Proceedings of the National Academy of Sciences*, 110(50), pp. 20117–20122. doi: 10.1073/pnas.1313452110.

Tomàs-Gamisans, M., Ferrer, P. and Albiol, J. (2016) ‘Integration and Validation of the Genome-Scale Metabolic Models of Pichia pastoris: A Comprehensive Update of Protein Glycosylation Pathways, Lipid and Energy Metabolism’, *PLOS ONE*. Edited by W. L. Araujo. Public Library of Science, 11(1), p. e0148031. doi: 10.1371/journal.pone.0148031.

Traeger, S. *et al.* (2013) ‘The genome and development-dependent transcriptomes of Pyronema confluens: A window into fungal evolution’, *PLoS Genet*, 9(9), p. e1003820. doi: 10.1371/journal.pgen.1003820.

Trinci, A. P. J. *et al.* (1994) ‘Anaerobic fungi in herbivorous animals’, *Mycological Research*, pp. 129–152. doi: 10.1016/S0953-7562(09)80178-0.

Tsirigos, K. D. *et al.* (2015) 'The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides', *Nucleic Acids Research*, 43(W1), pp. W401–W407. doi: 10.1093/nar/gkv485.

U.S. Department of Energy (2016) *2016 Billion-ton report: Advancing domestic resources for a thriving bioeconomy, volume 1: Economic availability of feedstocks. M. H. Langholtz, L. M. Eaton (Eds.), ORNL/TM-2016/160. Oak Ridge National Laboratory, Oak Ridge, TN*.

U.S. Department of Energy (2017) *2016 Billion-ton report: Advancing domestic resources for a thriving bioeconomy, volume 2: Environmental sustainability effects of select scenarios from volume 1. R. A. Efroymson, M. H. Langholtz, K.E. Johnson, and B. J. Stokes (Eds.), ORNL/TM-2016/727. Oak Ridge National Laboratory, Oak Ridge, TN.*

Uehling, J. *et al.* (2017) 'Comparative genomics of Mortierella elongata and its bacterial endosymbiont Mycoavidus cysteinexigens', *Environmental Microbiology*. doi: 10.1111/1462-2920.13669.

'UniProt: a worldwide hub of protein knowledge' (2019) *Nucleic Acids Research*. Narnia, 47(D1), pp. D506–D515. doi: 10.1093/nar/gky1049.

Urban, P. L. (2015) 'Universal electronics for miniature and automated chemical assays', *Analyst*. Royal Society of Chemistry, 140(4), pp. 963–975. doi: 10.1039/c4an02013h.

Urban, P. L. (2018) 'Prototyping Instruments for the Chemical Laboratory Using Inexpensive Electronic Modules', *Angewandte Chemie International Edition*, 57(34), pp. 11074–11077. doi: 10.1002/anie.201803878.

Vanwonterghem, I. *et al.* (2016) 'Genome-centric resolution of microbial diversity, metabolism and interactions in anaerobic digestion', *Environmental Microbiology*, 18(9), pp. 3144–3158. doi: 10.1111/1462-2920.13382.

Vardakou, M. *et al.* (2008) 'Understanding the structural basis for substrate and inhibitor recognition in eukaryotic GH11 xylanases.', *Journal of molecular biology*, 375(5), pp. 1293–1305. doi: 10.1016/j.jmb.2007.11.007.

Varghese, N. J. *et al.* (2015) 'Microbial species delineation using whole genome sequences', *Nucleic Acids Research*, 43(14), pp. 6761–6771. doi: 10.1093/nar/gkv657.

Varma, A. and Palsson, B. O. (1994) 'Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110', *Applied and Environmental Microbiology*, 60(10), pp. 3724–3731.

Videvall, E. (2018) 'Plasmodium parasites of birds have the most AT-rich genes of eukaryotes', *Microbial Genomics*, 4(2). doi: 10.1099/mgen.0.000150.

Vongsangnak, W. *et al.* (2016) 'Genome-scale metabolic modeling of Mucor circinelloides and comparative analysis with other oleaginous species', *Gene*, 583(2), pp. 121–129. Available at: https://www.sciencedirect.com/science/article/pii/S0378111916300956 (Accessed: 11 July 2019).

Walker, M. *et al.* (2009) 'Potential errors in the quantitative evaluation of biogas production in anaerobic digestion processes', *Bioresource Technology*. Elsevier, 100(24), pp. 6339–6346. doi: 10.1016/j.biortech.2009.07.018.

Wang, C. *et al.* (2019) 'Efficient production of glycyrrhetinic acid in metabolically engineered Saccharomyces cerevisiae via an integrated strategy', *Microbial Cell Factories*. BioMed Central, 18(1), p. 95. doi: 10.1186/s12934-019-1138-5.

Wang, D. *et al.* (2013) 'Draft genome sequence of Rhizopus chinensis CCTCCM201021, used for brewing traditional Chinese alcoholic beverages', *Genome Announcements*, 1(2), pp.

e00195-12. doi: 10.1128/genomeA.00195-12.

Wang, H.-C., Chen, Y.-C. and Hseu, R.-S. (2014) 'Purification and characterization of a cellulolytic multienzyme complex produced by Neocallimastix patriciarum J11.', *Biochemical and biophysical research communications*, 451(2), pp. 190–195. doi: 10.1016/j.bbrc.2014.07.088.

Wang, H. *et al.* (2018) 'RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor', *PLoS Computational Biology*. Public Library of Science, 14(10). doi: 10.1371/journal.pcbi.1006541.

Wang, T. Y. *et al.* (2011) 'Functional characterization of cellulases identified from the cow rumen fungus Neocallimastix patriciarum W5 by transcriptomic and secretomic analyses', *Biotechnology for Biofuels*, 4. doi: 10.1186/1754-6834-4-24.

Weimer, P. J., Russell, J. B. and Muck, R. E. (2009) 'Lessons from the cow: What the ruminant animal can teach us about consolidated bioprocessing of cellulosic biomass', *Bioresource Technology*, 100(21), pp. 5323–5331. doi: 10.1016/j.biortech.2009.04.075.

Wilken, S. *et al.* (2018) 'In Silico Identification of Microbial Partners to Form Consortia with Anaerobic Fungi', *Processes*, 6(1), p. 7. doi: 10.3390/pr6010007.

Wilken, S. E. *et al.* (2019) 'Linking "omics" to function unlocks the biotech potential of non-model fungi', *Current Opinion in Systems Biology*, pp. 9–17. doi: 10.1016/j.coisb.2019.02.001.

Wilken, S. E. *et al.* (2020) 'Genomic and proteomic biases inform metabolic engineering strategies for anaerobic fungi', *Metabolic Engineering Communications*. Elsevier B.V., 10. doi: 10.1016/j.mec.2019.e00107.

Wilkens, C. *et al.* (2017) 'GH62 arabinofuranosidases: Structure, function and

applications', *Biotechnology Advances*. Elsevier Inc., pp. 792–804. doi: 10.1016/j.biotechadv.2017.06.005.

Wilkinson, T. J. *et al.* (2018) 'CowPI: A Rumen Microbiome Focussed Version of the PICRUSt Functional Inference Software', *Frontiers in Microbiology*, 9. doi: 10.3389/fmicb.2018.01095.

Williams, A. G. and Orpin, C. G. (1987) 'Polysaccharide-degrading enzymes formed by three species of anaerobic rumen fungi grown on a range of carbohydrate substrates.', *Canadian journal of microbiology*, 33(5), pp. 418–426.

Wolin, M. J. (1981) 'Fermentation in the rumen and human large intestine', *Science*, 213(4515), pp. 1463–1468. doi: 10.1126/science.7280665.

Woo, H. L. *et al.* (2014) 'Complete genome sequence of the lignin-degrading bacterium Klebsiella sp. strain BRL6-2', *Standards in Genomic Sciences*, 9(1), p. 19. doi: 10.1186/1944-3277-9-19.

Wood, T. M. *et al.* (1986) 'A highly active extracellular cellulase from the anaerobic rumen fungus Neocallimastix frontalis', *FEMS Microbiology Letters*, 34(1), pp. 37–40.

Wright, P. E. and Dyson, H. J. (2015) 'Intrinsically disordered proteins in cellular signalling and regulation', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 16(1), pp. 18–29. doi: 10.1038/nrm3920.

Wu, H. *et al.* (2012) 'On the molecular mechanism of GC content variation among eubacterial genomes', *Biology Direct*, 7, p. 2.

Wu, I. and Arnold, F. H. (2013) 'Engineered thermostable fungal Cel6A and Cel7A cellobiohydrolases hydrolyze cellulose efficiently at elevated temperatures', *Biotechnology and Bioengineering*. Wiley Subscription Services, Inc., A Wiley Company, 110(7), pp. 1874–

1883. doi: 10.1002/bit.24864.

Wu, Z. *et al.* (2019) 'Machine learning-assisted directed protein evolution with combinatorial libraries', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 116(18), pp. 8852–8858. doi: 10.1073/pnas.1901979116.

Ximenes, E. A. *et al.* (2005) 'A mannanase, ManA, of the polycentric anaerobic fungus Orpinomyces sp. strain PC-2 has carbohydrate binding and docking modules', *Canadian Journal of Microbiology*, 51(7), pp. 559–568. doi: 10.1139/w05-033.

Xue, G. P. *et al.* (1992) 'Cloning and expression of multiple cellulase cDNAs from the anaerobic rumen fungus Neocallimastix patriciarum in Escherichia coli.', *J Gen Microbiol.*, 138(7), pp. 1413–1420.

Xue, G. P. *et al.* (1997) 'Improvement of expression and secretion of a fungal xylanase in the rumen bacterium Butyrivibrio fibrisolvens OB156 by manipulation of promoter and signal sequences.', *Journal of biotechnology*, 54(2), pp. 139–148.

Xue, G. P., Gobius, K. S. and Orpin, C. G. (1992) 'A novel polysaccharide hydrolase cDNA (celD) from Neocallimastix patriciarum encoding three multi-functional catalytic domains with high endoglucanase, cellobiohydrolase and xylanase activities', *Journal of General Microbiology*, 138, pp. 2397–2403.

Yarlett, N. *et al.* (1986) 'Hydrogenosomes in the rumen fungus Neocallimastix patriciarum', *Biochemical Journal*. Portland Press, 236(3), pp. 729–739. doi: 10.1042/bj2360729.

Yoo, J. I., Daugherty, P. S. and O'Malley, M. A. (2020) 'Bridging non-overlapping reads illuminates high-order epistasis between distal protein sites in a GPCR', *Nature*

Communications. Nature Research, 11(1), pp. 1–12. doi: 10.1038/s41467-020-14495-7.

Young, J. D. (2014) 'INCA: a computational platform for isotopically non-stationary metabolic flux analysis', *Bioinformatics*. Oxford University Press, 30(9), pp. 1333–1335. doi: 10.1093/bioinformatics/btu015.

Youssef, N. H. *et al.* (2013) 'The genome of the anaerobic fungus Orpinomyces sp. strain C1A reveals the unique evolutionary history of a remarkable plant biomass degrader.', *Applied and environmental microbiology*, 79(15), pp. 4620–34. doi: 10.1128/AEM.00821-13.

Zetsche, B. *et al.* (2015) 'Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System', *Cell*. Cell Press, 163(3), pp. 759–771. doi: 10.1016/J.CELL.2015.09.038.

Zhang, G. *et al.* (2016) 'Bioprospecting metagenomics of a microbial community on cotton degradation: Mining for new glycoside hydrolases', *Journal of Biotechnology*, 234, pp. 35–42. doi: 10.1016/j.jbiotec.2016.07.017.

Zoltan, I. K. and John, J. W. (1933) 'Process of treating plant juices and extracts (Patent US 1932833 A)'.

Zorova, L. D. *et al.* (2018) 'Mitochondrial membrane potential', *Analytical Biochemistry*. Academic Press Inc., 552, pp. 50–59. doi: 10.1016/j.ab.2017.07.009.

Zou, Y. *et al.* (2019) '1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses', *Nature Biotechnology*. Nature Publishing Group, 37(2), pp. 179–185. doi: 10.1038/s41587-018-0008-8.