UNIVERSITY OF CALIFORNIA

Santa Barbara

Engineering Regulation in Anaerobic Gut Fungi during Lignocellulose Breakdown

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemical Engineering

by

John Kyle Henske

Committee in charge:

Professor Michelle A. O'Malley, Chair

Professor M. Scott Shell

Professor Todd M. Squires

Professor Irene A. Chen

September 2017

The dissertation of John Kyle Henske is approved.

Irene A. Chen

M. Scott Shell

Todd M. Squires

Michelle A. O'Malley, Committee Chair

September 2017

Engineering Regulation in Anaerobic Gut Fungi during Lignocellulose Breakdown

by

John Kyle Henske

ACKNOWLEDGEMENTS

It would not have been possible to accomplish the work described in this thesis without the help and support of many different people in many different forms. I would like to thank my advisor, Michelle O'Malley, for all of her support throughout the past five years. She has always been readily available to provide rapid and useful feedback and has enabled valuable opportunities to present my work at research conferences enabling me to develop an important ability to discuss and explain my work to a wide audience. She even let me teach an entire lecture on the science of brewing beer. I would also like to thank all the members of the O'Malley lab, for the part they played in various experiments either through collaboration on experiments or scientific discussions at group meetings. Special thanks, of course, go to all the members of Team Fungus. Also, thank you to Dr. Kevin Solomon for the role he played as a mentor, particularly during the early stages, while he was a post-doc in the O'Malley lab. I am also grateful for continued support from our collaborator Mike Theodorou who taught me how to culture these unique, and often finnicky, microbes.

I would also like to thank my friends and family for their support. My family has always supported me in everything I have done, giving me confidence and the motivation to always improve. My parents have always supported my decisions, even when that meant me moving clear across the country for graduate school. Despite never having the opportunity to know him, my late grandfather also influenced my decision to study chemical engineering. Stories of his accomplishments as a chemical engineer got me interested in the field. My friends have always provided a vital opportunity to take my mind off some of the more difficult aspects of graduate school. Most of all I would like to thank Allie. She has endured all the highs and

lows of graduate school alongside me. She has celebrated my achievements and encouraged me during the hardships. Her continued support helped me throughout all my accomplishments. It was with the help of all these people that some of the most challenging years of my life were also the most rewarding.

VITA OF JOHN KYLE HENSKE
June 2017

EDUCATION

2012 – 2017 Doctor of Philosophy in Chemical Engineering
University of California, Santa Barbara

2007 – 2012 Bachelor of Science in Chemical Engineering
Northeastern University
Graduated Summa Cum Laude

PROFESSIONAL EMPLOYMENT

2012-2017 Graduate Researcher and Teaching Assistant
Department of Chemical Engineering, UC Santa Barbara
Academic Advisor: Michelle O’Malley

2010-2012 Undergraduate Research Assistant
Department of Chemical Engineering, Northeastern University
Academic Advisor: Shashi Murthy

Jan-Jul 2011 Research and Development Engineering Co-op
GVD Corporation

Jan-Jun 2010 Wet Extrusion Research and Development Co-op
Novartis-MIT Center for Continuous Manufacturing
Charles Cooney Laboratory

Jan-Jun 2009 Process Engineer Co-op: Purification Manufacturing Tech. Support
Genzyme Corporation

PUBLICATIONS

In progress: KV Solomon, SP Gilmore, **JK Henske**, D Thompson, MA O’Malley. Natural antisense transcripts are primary regulators of the gut fungal catabolic response.

In review: JK Henske, S Wilken, KV Solomon, MK Theodorou, MA O’Malley. Metabolic characterization of anaerobic fungi provides a path forward for two-microbe conversion of lignocellulose to bio-based chemicals.

In review: SP Gilmore, **JK Henske**, J Sexton, KV Solomon, JI Yoo, LM Huyett, A Pressman, Z Cogan, V Kivenson, Y Tan, DL Valentine, MA O’Malley. Genomic analysis of Methanobacterium bryantii, Methanosarcina spelaei, Methanosphaera cuniculi, and

Methanocorpusculum parvum reveals a shift towards energy conservation in the genomes of methanogenic archaea. **BMC Genomics.**

Accepted: CH Haitjema, SP Gilmore, **JK Henske**, KV Solomon, R deGroot, A Kuo, S Mondo, A Salamov, K LaButti, Z Zhao, J Chiniquy, K Barry, H Brewer, S Purvine, A Wright, T van Alen, B Boxma, J Hackstein, S Baker, I Grigoriev, MA O'Malley. A Parts List for Fungal Cellulosomes Revealed by Comparative Genomics. **Nature Microbiology.**

S Seppälä, KV Solomon, SP Gilmore, **JK Henske**, MA O'Malley. Mapping the membrane proteome of anaerobic gut fungi identifies a wealth of carbohydrate binding proteins and transporters. **Microbial Cell Factories**, 15:212, (2016)

G. J. Li et. al (+134 additional authors, including K. V. Solomon, **J. K. Henske**, C. H. Haitjema, S. P. Gilmore, M. K. Theodorou, M. A. O'Malley), "Fungal diversity notes 253-366: taxonomic and phylogenetic contributions to fungal taxa," **Fungal Diversity**, 78(1): 1-237 (2016).

KV Solomon, CH Haitjema, **JK Henske**, SP Gilmore, D Borges-Rivera, A Lipzen, HM Brewer, SO Purvine, AT Wright, MK Theodorou, I Grigoriev, A Regev, DA Thompson, MA O'Malley. Early-branching gut fungi possess a large, comprehensive array of biomass degrading enzymes. **Science**. 351 (6278), 1192-1195. (2016)

JK Henske[‡]**,** KV Solomon[‡], MK Theodorou, MA O'Malley. Robust and effective methodologies for cryopreservation and DNA extraction from anaerobic gut fungi. *Anaerobe*. 38: 39-46. (2016) [‡]Equal author contributions

SP Gilmore, **JK Henske**, MA O'Malley. Driving biomass breakdown through engineered cellulosomes. **Bioengineered**. 6 (4), 204-208. (2015).

Haitjema, C. H., Solomon, K. V., **Henske, J. K.**, Theodorou, M. K. and O'Malley, M. A. Anaerobic gut fungi: Advances in isolation, culture, and cellulolytic enzyme discovery for biofuel production. **Biotechnol. Bioeng.**, 111, 1471–1482. (2014)

PRESENTATIONS

April 2017	American Chemical Society National Conference; San Francisco, CA (talk)
Sept 2016	Clorox-Amgen Graduate Student Symposium; UC Santa Barbara (talk)
Sept 2016	2016 Renewable Carbon Workshop; UC Santa Barbara (talk)
Jul 2016	Synthetic Biology: Engineering Evolution and Design; Chicago, IL (poster)
Mar 2016	Joint Genome Institute User Meeting; Walnut Creek, CA (poster)
Mar 2016	American Chemical Society National Conference; San Diego, CA (talk)
Mar 2016	DOE Contractor-Grantee Meeting; Washington DC (poster)
Jun 2015	Synthetic Biology Engineering Evolution and Design; Boston, MA (poster)
Mar 2015	American Chemical Society National Conference; Denver, CO (talk)
Feb 2015	DOE Contractor-Grantee Meeting; Washington DC (poster)

Jul 2014	Syn. Bio. Engineering Evolution and Design; Manhattan Beach, CA (poster)
Feb 2014	Joint Genome Institute User Meeting; Walnut Creek, CA (poster)
Jan 2014	Southern California Systems Biology Conference; UC Irvine (poster)

AWARDS AND HONORS

2017	UCSB Doctoral Student Travel Grant for ACS Meeting
2017	UCSB Academic Senate Doctoral Travel Grant
2016	UCSB Graduate Dissertation Fellowship
2015	Mellichamp Sustainability Fellow, University of California, Santa Barbara

ABSTRACT

Engineering Regulation in Anaerobic Gut Fungi during Lignocellulose Breakdown

by

John Kyle Henske

The development of a renewable, bio-based economy requires efficient methods to extract fermentable sugars from complex plant material. Currently, bioprocessing from crude biomass requires multiple steps including pretreatment to separate lignin from sugar-rich cellulose and hemicellulose, enzymatic hydrolysis to release simple sugars, and microbial fermentation to produce value-added chemicals. Consolidated bioprocessing seeks to improve bioprocessing efficiency by reducing the number of steps required to get from plant biomass to chemical product. To address this challenge, we derived inspiration from natural microbial communities known for degrading biomass. Within the rumen microbiome of large herbivores, anaerobic gut fungi are the primary colonizers of plant material and present an untapped opportunity for consolidated bioprocessing. These unique microorganisms efficiently hydrolyze lignocellulosic biomass into simple sugars, but remain relatively uncharacterized in comparison to industrial production organisms. We implemented Next Generation Sequencing (NGS) technologies alongside biochemical studies to develop a deeper understanding of gut fungi, their metabolism, and the mechanisms by which they break down complex biomass to identify a path forward for their industrial application. We also developed simple, rapid methodologies for cryopreservation and DNA extraction that are critical for the development of industrial microbes.

Sequencing and functional annotation of transcriptomes and genomes of novel isolated species of gut fungi has elucidated their large repertoire of biomass degrading enzymes including cellulases, hemicellulases, and accessory enzymes. These enzymes allow them to efficiently degrade crude biomass, yielding similar growth rates on complex plant material and simple sugars. Remarkably, in isolated batch culture, the biomass degrading power of gut fungi is sufficient to generate surplus fermentable sugars for the growth of additional microorganisms. This ability has been exploited to develop a novel two-stage consolidated bioprocessing scheme that uses anaerobic gut fungi to consolidate the pretreatment and hydrolysis steps in traditional bioprocessing to hydrolyze sugars directly from crude biomass. These sugars can then be fed to the easily metabolically engineered model yeast, *Saccharomyces cerevisiae*, to support growth and bioproduction in a two-stage fermentation scheme.

Further, RNA sequencing studies have provided critical insight into the regulation of biomass degrading activity. Gene expression during growth on varying substrates and in response to a carbon catabolite repressor has revealed conditions required to optimize expression of biomass degrading enzymes. Unannotated sequences that co-regulate with predicted biomass degrading enzymes have also been identified as candidate genes that may host novel biomass degrading function. Together these results reveal important process considerations for the use of gut fungi in industrial bioprocessing to maximize the production of enzymes and the degradation of biomass. While challenges remain for the implementation of gut fungi in industrial bioprocessing, we have demonstrated their potential to consolidate pretreatment and hydrolysis either through engineered culturing schemes or development of improved enzyme cocktails for biomass hydrolysis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

1.1. Motivation

Bio-based production of fuels and chemicals provides an opportunity to reduce the reliance on fossil fuels and move to a more sustainable global economy. Major drivers for this change include the decreasing abundance of materials such as oil, gas, and coal, price volatility of these feedstocks, and a need to reduce greenhouse gas emissions[1]. While the first generation of biofuel production relied primarily on agricultural sources that were also food sources, such as vegetable oils and corn sugar, current efforts have shifted focus to economical use of lignocellulosic feedstocks, that make up the majority of non-food plant materials[2]. These materials that may be agricultural wastes (e.g. corn stover) or invasive species (e.g. reed canary grass). Regardless of substrate, the recalcitrance of lignin in plant cell walls that restricts enzyme activity against the sugar-rich cellulose and hemicellulose within lignocellulosic biomass[3]. Thus, multi-step processed that employ energy-intensive pretreatments prior to enzymatic hydrolysis and subsequent fermentation are frequently employed[4]. To improve efficiency and reduce the cost of bio-based chemical production, consolidated bioprocessing (CBP) seeks to reduce the number of steps required for production of fuels and chemicals from plant material[5].

Anaerobic gut fungi found in the guts of large herbivores present an opportunity for CBP as they possess a wide range of enzymes required to efficiently break down crude plant material[6-9]. Due to the immense number of cellulases, hemicellulases, and other hydrolytic enzymes secreted by gut fungi[10,11] they represent an untapped resource of enzymatic machinery for consolidated bioprocessing. However, due to the difficulty of their isolation, culture, and characterization, little has been done to apply their capabilities to bioprocessing. While gut fungi have potential for application in bio-based production from lignocellulosic materials, it is necessary to first fill the gaps in knowledge. We have obtained transcriptomic and genetic sequence information to identify the critical biomass degrading and metabolic functions present within novel isolated strains of gut fungi. We also determined mechanisms of regulation under conditions relevant for bioprocessing applications to highlight optimal conditions for lignocellulolytic enzyme production and biomass degradation. All this information is critical for the development of anaerobic gut fungi as novel, engineered bio-based production platform organisms.

1.2. Organization of the Dissertation

This dissertation describes the isolation, characterization, and application of anaerobic gut fungi for the degradation of plant material in bio-based production. These understudied organisms have immense potential for application in the breakdown of biomass, but have not been employed due to the relatively poor understanding of their genetics and metabolism as well as a lack of genetic tools to modify them. Through the study of these organisms we were able to develop simple methods to isolate high quality genomic DNA for assembly of full genomes and long term cryogenic storage of isolated cultures, gain an understanding of how they regulate biomass degrading enzymes, and develop a framework for the implementation of anaerobic gut fungi for the hydrolysis of sugars from crude biomass.

The dissertation is comprised of six chapters. The first chapter introduces the field of bio-based fuel and chemical production, including first generation biofuels and more recent improvements, and describes the potential for anaerobic gut fungi to fit into these processes. The second chapter describes the isolation and characterization of anaerobic gut fungi using growth experiments as well as DNA and RNA sequencing for genome and transcriptome acquisition. The third chapter describes the development of simple methods for cryopreservation and extraction of high quality DNA, two tools that are critical to any organism in industrial use. The fourth chapter details the application of anaerobic gut fungi into a consolidated bioprocessing scheme that leverages a new understanding of metabolic capabilities to combine pretreatment and hydrolysis steps, providing sugars to model microbes for production through nutrient linkage. The fifth chapter examines the response of gut fungi to a simple catabolite repressor (e.g. sugar), detailing global regulation of genes and highlighting the impact of glucose on the expression of carbohydrate active enzymes. The sixth and final chapter summarizes the implications of this research as whole and discusses the next steps and challenges for implementation of anaerobic gut fungi into bio-based fuel and chemical production.

1.3. Renewable, bio-based production of fuels and chemicals

While sustainable production of fuels and chemicals from plant biomass is desirable, the recalcitrance of plant biomass must be overcome to create efficient, cost effective processes[5,12]. Starches from maize and simple sugars from sugarcane and sugar beets can be easily obtained and used for fermentations[13]. In this regard, first generation biofuels primarily focused on the production of ethanol and biodiesel from edible plant sources such as these, as they are a simple resource for 6-carbon sugars that can easily be fermented by yeast, *Saccharomyces cerevisiae*. In countries like Brazil, sugarcane is commonly used for bioethanol production, whereas in the United States corn is the primary feedstock for first generation biofuels[13]. However, the production of these first-generation biofuels puts increased strain on the agricultural industry by increasing demands on important agricultural commodities traditional used for food and animal feed. In fact, the agricultural commodities that are common feedstocks for first-generation fuels, such as sugarcane, maize, and cassava also comprise a large share of the diets of food-insecure people worldwide leading to dangerous implications for global food security[14].

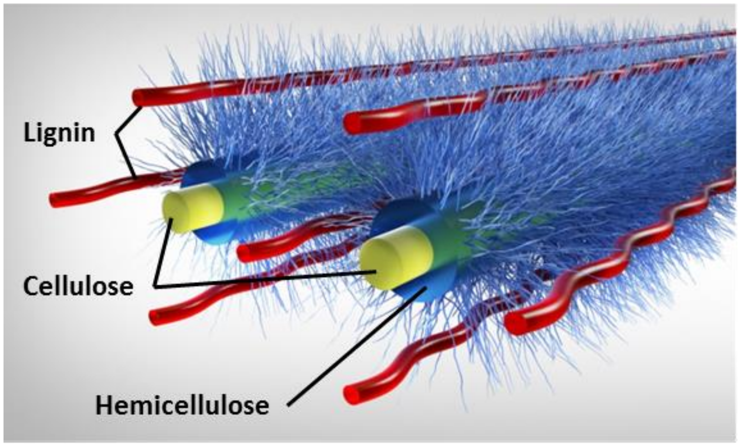


Figure 1.1. Lignocellulose is a complex structure of cellulose, hemicellulose, and lignin

Plant cell walls are primarily composed of crystalline cellulose, hemicellulose, and lignin in the form of microfibrils. In these microfibrils, crystalline cellulose is at the core with amorphous hemicellulose, and waxy lignin around it. The lignin and hemicellulose surround the crystalline cellulose, making it more difficult for enzymes to access.

Second-generation biofuel production aims to avoid the use of foods for fuel and chemical production, instead turning to lignocellulosic plant biomass as feedstocks[2]. Fibrous plant material contains additional sugars within structural biopolymers, but these sugars are difficult to extract. The cell walls of plants have evolved to resist degradation by microbes and their enzymes[15] and contain three major biopolymer components: cellulose, hemicellulose, and lignin. The composition of these three components can vary greatly in different types of biomass generally with cellulose ranging from 15-50%, hemicellulose ranging from 10-30%, and lignin ranging from 10-20% of biomass by dry weight[16,17]. Cellulose is a polymer comprised of hundreds or thousands of glucose molecules joined by β(1,4) glucosidic bonds and the action of several different types of cellulose degrading enzymes (cellulases) including endo-glucanases, exo-glucanases, and β-glucosidases are required to hydrolyze cellulose into its glucose monomers[18]. Furthermore, cellulose is found in both crystalline and amorphous forms; the amorphous form is more susceptible to enzymatic digestion, but the crystalline cellulose core of cell wall microfibrils is resistant due to its rigid, compact structure[15].

While cellulose is the most abundant polysaccharide in nature, hemicellulose is also present in large quantities in plant material. Unlike cellulose, the composition of hemicellulose can vary. Hemicelluloses are heterogeneous polymers of pentose sugars (ie – xylose, arabinose), hexose sugars (ie – glucose, mannose, galactose), and sugar acids (ie – glucuronic acid, ferulic acid)[19]. Thus, the enzymatic activity required to degrade hemicellulose varies with its composition. These enzymes include, but are not limited to xylanases, xylosidases, mannanases, mannosidases, galactanases, and arabinanases[20].

The third major component of biomass, lignin, is more heterogeneous and complex then hemicellulose. Lignins are complex aromatic heteropolymers comprised of monomeric units that are primarily derived from three hydroxycinnamyl alcohol monomers (*p*-coumaryl, coniferal, and sinapyl alcohols)[21]. The presence of lignin in biomass has a negative impact on the processing steps required for biofuel production reducing the accessibility of cellulolytic enzymes to cellulose[22]. To combat this issue, the first step of bio-based production is typically a pretreatment step intended to separate lignin, hemicellulose, and cellulose[16,23,24].

1.3.1. Production of fuels and chemicals from biomass

Primary methods for the conversion of biomass to fuels and other chemical products are thermochemical and biochemical processes. Thermochemical conversion processes primarily consist of combustion, pyrolysis, gasification, and liquefaction to produce a combination of solid (charcoal), liquid (bio-oils), and gaseous fuel compounds. These methods often require elevated temperatures and/or pressures, solvents, and catalysts that can lead to expensive and energy intensive operations[25-27]. Biochemical methods use microbial fermentations to produce both fuels and chemicals from carbohydrate sources, but typically rely on pretreatments that may be harsh and energy intensive to extract sugars from biomass (Figure 1.2). Pretreatment technologies primarily aim to make crude biomass more susceptible to enzymatic hydrolysis, such that fermentable sugars can be easily obtained from cellulose. These pretreatments include mechanical size reduction as well as acid, base, solvent, and ionic liquid incubations to increase enzyme access to the cellulose locked within lignocellulosic biomass[23,25]. After these pretreatments, enzymatic hydrolysis is employed to hydrolyze cellulose into its simple sugar constituents that are more amenable to microbial fermentations. These hydrolysis steps require the action of a suite of enzymes to break down cellulose that must be supplied in large amounts and are expensive to produce[28]. Once simple sugars are obtained, engineered microbes are used to produce fuels and chemicals.

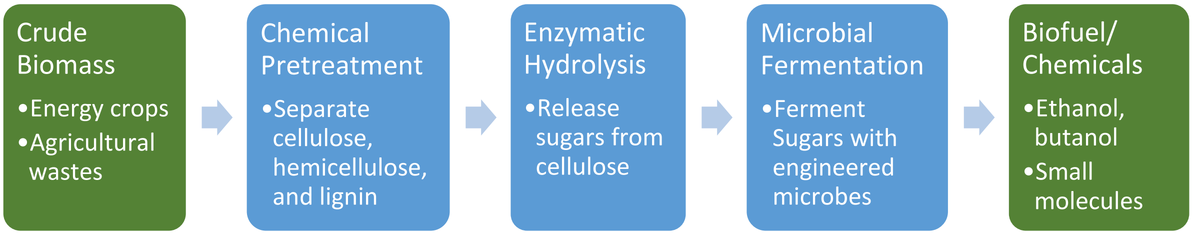


Figure 1.2. Typical process for biochemical conversion of biomass.

Biochemical conversion methods typically employ three main steps for conversion of crude biomass to value-added fuels and chemicals. Pretreatments separate the sugar rich cellulose and hemicellulose from lignin and include acid/base, steam explosion, ionic liquid, and other harsh treatments. Enzymatic hydrolysis of cellulose requires the action of many enzymes including endo- and exo-glucanases as well as β-glucosidases to produce glucose. Glucose can then be fed to microbes like *S. cerevisiae* and *E. coli* engineered for the production of fuels and chemicals. Flow chart adapted from Balan et al.[25]

Sustainable production of fuels began with the production of ethanol from sugar and starch rich sources and expanded into the use of lignocellulose as a feedstock to avoid using food sources[13]. Compared to gasoline, ethanol has a lower energy density with approximately 40% less energy per unit mass[29]. Longer chain fuels with higher energy densities can be made through bio-diesel production by transesterification of vegetable oils[30]. However, biodiesel's dependence on vegetable oil suffers from a competition with food sources for feedstock materials. Microbial fermentations offer another method for production of biodiesels by leveraging pathways for lipid and fatty acid biosynthesis that already exist in many microorganisms and engineering them to overproduce these compounds. These pathways can be modified to produce short chain fuels, fatty alcohols, and waxes from plant derived sugars[31-33].

In addition to fuels, petroleum is also an important source of many organic chemicals and polymers. As such, it is also necessary to develop methods to produce these chemicals from renewable biomass feedstocks. While bio-based fuel production is still expensive compared to fossil fuels, introduction of these chemicals as co-products provides an opportunity to offset the cost of fuel production[1]. In fact, a wide variety of chemicals can be produced through microbial fermentations that are derived from both sugar rich, starchy feedstocks and recalcitrant lignocellulose[34]. Common chemicals produced include lactic acid[35-37] and succinic acid[38-41], two chemicals that are natural byproducts of the energy generating metabolism of many microbes. These chemicals can be used as chemical feedstocks for the production of bio-polyesters such as polylactic acid (PLA) and poly(butylene succinate) (PBS)[34]. Aromatic compounds can also be produced from biomass, including cinnamic acid, phenyllactic acid, and caffeic acid from biomass hydrolysates[34]. Additional opportunities for cost reduction lie in developing methods that reduce the number of processing steps through consolidated bioprocessing.

1.3.2. Consolidated bioprocessing

The overall goal of consolidated bioprocessing (CBP) is to reduce the number of processing steps required to get from crude lignocellulosic plant material to valuable fuels and chemicals[5,42]. This consolidation is intended to improve the economics of bioprocessing largely by eliminating capital and operating costs associated with additional processing equipment[5]. Most CBP approaches aim to consolidate the hydrolysis of cellulose and hemicellulose into sugars and production by microbial fermentation into a single step rather than two separate steps. The primary approach is often referred to as the “superbug” approach: engineering a single organism to perform both hydrolysis and production (Figure 1.3.A). This can be accomplished either through native or recombinant strategies[5,42]. The recombinant strategy utilizes the wide array of genetic engineering tools available for typical industrially-friendly production organisms, such as *Saccharomyces cerevisiae* and *Escherichia coli*, to introduce cellulolytic capabilities through heterologous expression of cellulolytic enzymes and enzyme complexes[43-46]. However, there are many challenges associated with this approach, particularly in the expression of the cellulolytic enzymes in these host systems[47,48]. In particular, there is an added metabolic burden associated with the production of heterologous cellulases[49] that divert resources from other metabolic processes to produce these enzymes as well as secrete them at high titers. This added burden results in slower growth rates and therefore lower enzyme production rates.

Contrary to the recombinant approach, the native strategy takes advantage of the natural capability of cellulolytic organisms to hydrolyze cellulose and strives to engineer these organisms to also produce value-added fuels and chemicals[5,42]. This approach has used organisms such as *Clostridium*[50-53], *Caldicellulosiruptor*[54,55], and filamentous fungi[56] that already possess the capability to degrade cellulose and metabolically engineers them for production of fuels and chemicals. This strategy avoids the metabolic burden associated with engineering expression of many heterologous cellulases by using organisms with evolutionarily adapted cellulolytic activity. However, this approach suffers from a lack of genetic tools to engineer many of these organisms. While a variety of efficient tools have been developed to engineer typical production microbes like *S. cerevisiae* and *E. coli*[57] over the many years that they have been studied, many cellulolytic organisms were only isolated more recently due to advances in microbial isolation techniques. As such, many of these organisms lack the tools required to implement them for production, although rapid advances are being made in this field.

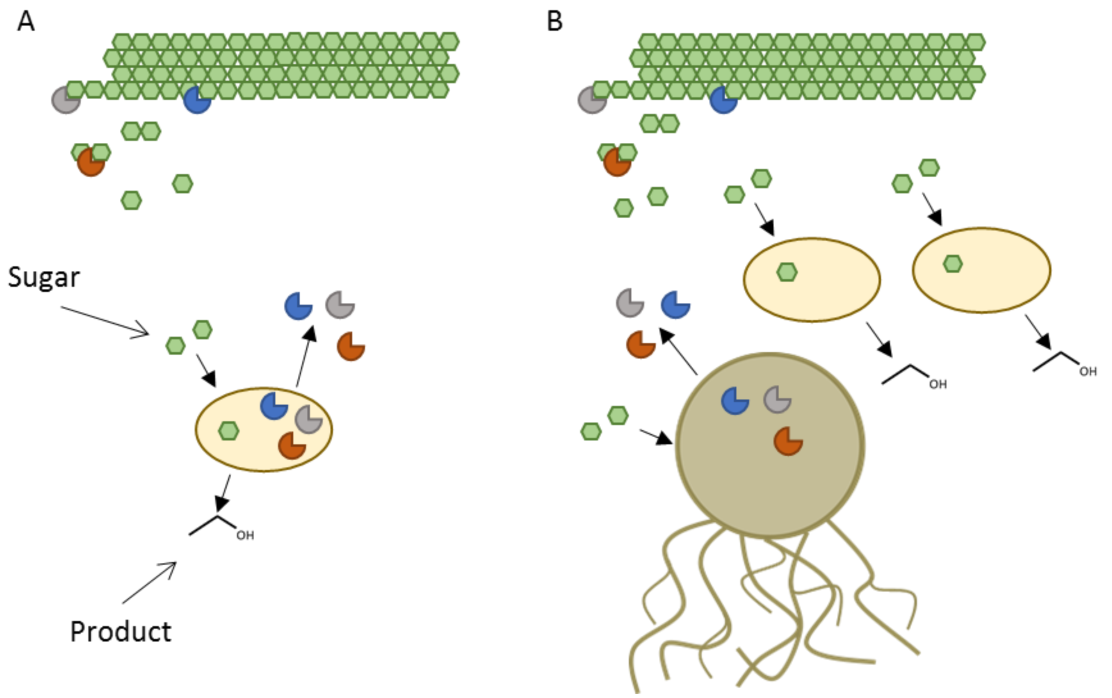


Figure 1.3. Consolidated bioprocessing with microorganisms.

Superbug based approaches to consolidated bioprocessing (A) use a single organism for production of cellulolytic enzymes and products. Consortia based approaches (B) distribute the responsibilities of cellulose degradation and production to the organisms better suited to each.

An alternative to this "superbug" approach is to use microbial consortia to combine different steps of bioprocessing (Figure 1.3.B). The consortia approach seeks to leverage the strengths of different organisms rather than attempt to engineer a single organism to possess all capabilities required for production from biomass. Microbial consortia for consolidated bioprocessing should contain multiple organisms that have complementary metabolic functions such that difficult tasks may be divided across all members[58]. Some recent examples have engineered symbiotic pairings of cellulolytic *Clostridium phytofermentans* with production ready *S. cerevisiae*[59], cellulolytic fungus *Trichoderma reesei* with production ready *E. coli*[60], and cellulolytic *Aspergillus oryzae* with production ready *S. cerevisiae*[61]. Many of these consortia still rely on extensive pretreatments of biomass to separate lignin from cellulose, except for the use of *A. oryzae* and *S. cerevisiae* to ferment waste brewer's grains for ethanol production. These approaches also limit the production conditions to those that accommodate both organisms thereby limiting the range of products that can be produced.

1.4. Anaerobic gut fungi are powerful degraders of crude plant biomass

While many cellulolytic organisms are limited in their ability to degrade crude lignocellulosic biomass without any pretreatment, anaerobic gut fungi thrive on untreated biomass. Gut fungi are found in the guts of large herbivores including ruminants (e.g. – cows, sheep, goats) and hindgut fermenters (e.g. – horses). Anaerobic gut fungi are part of a large microbial ecosystem responsible for the breakdown of plant material consumed by the animals supplying them with easily utilizable forms of carbon, energy and protein[62]. This microbiome is comprised of a large variety of cellulolytic and non-cellulolytic bacteria, anaerobic protozoa, archaeal methanogens, and anaerobic fungi[63-65].

Within this community, anaerobic gut fungi are considered the primary colonizers of plant biomass. Accounting for up to 8% of the microbial biomass in the rumen microbiome, gut fungi are capable of degrading the most recalcitrant lignocellulosic biomass due to their wide array of carbohydrate active enzymes[66]. Anaerobic gut fungi produce a wide variety of cellulases, hemicellulases, and accessory enzymes (e.g. – carbohydrate esterases, pectinases) that allow them to accomplish this difficult task[10]. In addition to enzymatic degradation of plant material, gut fungi also produce expansive networks of mycelia that apply physical force to aid in degradation of lignocellulose. As they degrade and ferment plant biomass, gut fungi produce formate, acetate, lactate, carbon dioxide, and molecular hydrogen as fermentation products. The production of carbon dioxide and hydrogen results in syntrophic pairings with archaeal methanogens also found in the rumen microbiome that convert carbon dioxide and hydrogen to methane[67]. This relationship has proven to enhance gut fungal degradation of biomass by removing inhibition of hydrogenases by molecular hydrogen[67-69]. Such metabolic linkages may be exploited for chemical production schemes.

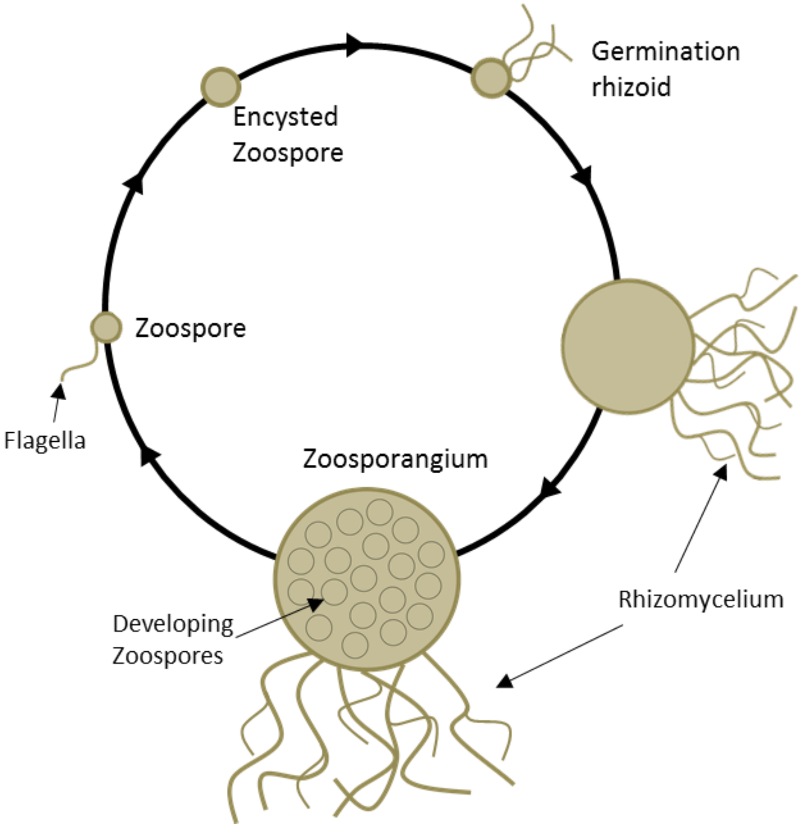


Figure 1.4. Life cycle of monocentric anaerobic gut fungus

The life cycle of anaerobic gut fungi starts with a motile zoospore that searches for a carbon source. Upon finding plant material or other carbon source, the zoospore encysts upon this material and begins to produce the rhizomycelium that root into plant material. The encysted zoospore grows into a zoosporangium and begins to produce more motile zoospores from within. The sporangium eventually ruptures, releasing zoospores and the cycle begins again. Figure adapted from Trinci et al.[62]

Despite their impressive ability to degrade biomass, gut fungi are largely understudied and have not been implemented in industrial processes. This is in part due to the difficulty of the isolation from animals as well as their sensitivity to oxygen and the need for strict anaerobic culture conditions[70]. Furthermore, gut fungi follow a unique life cycle (Figure 1.4) in which they start as motile zoospores until they find a food source at which point they encyst upon the surface of the plant material and develop a large sporangium. Inside this sporangium more motile zoospores are produced and eventually the large structure ruptures releasing tens to hundreds of zoospores. Upon first discovery, anaerobic gut fungi were classified as flagellated protozoa under the genus *Callimastix* and family Callimastigidae, as they were first discovered in their motile-zoospore growth phase. Later, they were identified as the zoospores of a primitive fungus[71] and placed into the fungal Family Neocallimastigaceae.

Gut fungi exist as two morphologically distinct subtypes: monocentric fungi form a single sporangium per vegetative mass, or thallus, and polycentric fungi are capable of forming multiple sporangia per thallus and are also characterized by the migration of nuclear material into their rhizomycelial root system[62]. Currently there are eight known genera of anaerobic gut fungi: *Piromyces*, *Caecomyces*, *Neocallimastix*, *Orpinomyces*, *Anaeromyces*[72], *Cyllamyces*[73], *Oontomyces*[74], and *Buwchfawromyces*[75]. Gut fungi were originally placed into genera and species through primarily morphological observation including sporangia size and structure, number of flagella on each zoospore, and rhizoid structure[76]. However, advances in molecular techniques as well as increases in the number of available gene sequences in databases has led to identification by short, conserved genetic sequences. For gut fungi these include the 18S ribosomal subunit and the internal transcribed spacer regions (ITS1 and ITS2) found between sequences for the different ribosomal subunits[77,78]. While these sequences are highly conserved to preserve the function of the ribosome, small changes can be used to differentiate and classify strains.

Compared to other microorganisms, there is a relative dearth of genetic information for anaerobic gut fungi. Until recent years, there were no complete genomes available. The first published genome of an anaerobic gut fungus was that of *Orpinomyces* sp. C1A, which was assembled using a combination of Pacific Biosciences long read sequencing and Illumina short read sequencing[79]. Prior to the *Orpinomyces* publication, the genome of *Piromyces* sp. E2 was assembled in March 2011 using Sanger DNA sequencing and is available on the Joint Genome Institute's genome portal[80,81], but was not published until 2017[82] due to a relatively poor assembly with many gaps. Subsequently, we sequenced the genomes of three fungal isolates *Piromyces finnis*, *Anaeromyces robustus*, and *Neocallimastix californiae*[82]. These latest genomes represent the best available genome assemblies for anaerobic gut fungi with the fewest gaps.

Isolation and sequencing of genomic DNA from anaerobic gut fungi suffers from several factors. In general, gDNA yield is very low, making it difficult to acquire enough material required as input for modern sequencing methods. This is in part due to the difficulty in lysing cells with thick cell walls and to relatively low amount of DNA compared to the amount of cellular material[83]. The genomes are also very AT-rich, with 80-85 mol% of the DNA comprised of adenosine and thymine bases[8,84,85]. The AT-rich nature of the genomes leads to complications in the manipulation genomic DNA and subsequent sequencing[86]. This high AT content is reflected in non-coding regions which can be above 97% AT and coding regions that tends toward AT-rich codons[8]. Furthermore, long repeat regions make assembly difficult using only short read, Illumina sequencing methods and the assembly of high quality genomes requires the use of Pacific Biosciences long read sequencing[87]. The rapid improvement of DNA sequencing technologies has provided a valuable opportunity to study genome organization of organisms like the anaerobic gut fungi.

1.5. Next Generation Sequencing and -Omics technologies

Next generation sequencing technologies and bioinformatics techniques have allowed for in depth study of the gut fungi's global metabolism and regulation mechanisms in a more complete, high throughput manner than is possible with other molecular biology techniques. Original sequencing methods, however, were very low throughput. Sanger sequencing, originally developed in 1977, involves the use of chain terminating dideoxy nucleoside analogues and gel electrophoresis to determine sequence based on the size of DNA fragments that incorporated these different dideoxy nucleosides to terminate DNA replication[88]. This method is the basis for most of the sequencing work conducted up to the present[89], including the initial sequencing of the human genome by the Human Genome Project[90] and the Craig Venter Institute[91], both employing a shotgun sequencing approach to sequence small DNA fragments and align them to obtain a complete genome[92]. Advancing from these technologies, newer sequencing techniques focus on a higher throughput approach where DNA fragments are sequenced in a highly parallel manner. Some techniques, such as Illumina sequencing focus on sequencing short reads ranging from 75 to 300 base pairs (bp) in length to reduce the error rate, and then align these short reads into full genomes. Others, like Pacific Biosciences and Oxford Nanopore sequencing technologies focus on obtaining longer reads up 20 kbp or 200 kbp, respectively. These latter techniques suffer from higher error rates on single pass reads, but use increased sequencing coverage to correct the errors[93]. The main driver in the development of new sequencing technologies has been reduction in the cost of the sequencing itself. The original sequencing of the human genome took between three and four years and cost approximately $300 million, but now companies are attempting to reduce that cost to just $1,000 to make it more accessible to the scientific and medical communities. Advances in next generation sequencing initially began rapidly dropping this price by drastically increasing the throughput with highly parallel methods[94]. Advances in sequencing technologies have made them not only more accessible for medical applications but for all research types, including the sequencing of microbes relevant for fuel and chemical production. In fact, the United States Department of Energy sponsored the 1000 Fungal Genomes project in order to address problems related to both energy and the environment[80]. While there are many different sequencing technologies that have become available, here we will discuss only Illumina and Pacific Biosciences technologies as they were implemented in the research discussed in this dissertation.

Illumina Inc. offers several different lines of sequencers that all operate on the same basic principle. These methods use a reversible terminator chemistry, advanced from the irreversible terminator chemistry used in Sanger sequencing. DNA sequencing is completed by repeated cycles of single base extensions with an engineered DNA polymerase using four reversible terminator variations of the four natural DNA nucleotides. After each incorporation, the identity of each base addition is determined by imaging the different fluorophores attached to each of the unnatural nucleotides. The fluorescent molecule and terminating side-arm to allow for the addition of another base. To enhance the signal from the fluorophores, the DNA templates are fixed to glass flow-cell and amplified into clusters of identical sequences[95]. While Illumina technologies tend to offer lower error rates (<1% in all sequencers offered, with some models at 0.1%)[93], these errors are not random. There are increased error rates towards the end of each read, likely due to accumulation of phase differences caused when a molecule fails to properly elongate on a given template, or advances faster than the other template strands in the cluster resulting in weaker fluorophore signal. This can cause difficulties for specific applications, such as amplicon sequencing as the length of each read is limited by this accumulation of errors[96]. For example, in metagenomic sequencing, the length of the variable regions used to classify the organisms in the consortia must be reduced to avoid error accumulation, although some techniques, such as paired-end sequencing can be used to alleviate some of these concerns[97]. There is also evidence to suggest that there is some sequence bias in the error of Illumina reads, with higher error rates in adenosine and cytosine nucleotides compared to guanosine and thymine[96], more errors among certain short motifs, and differences in sequencing coverage of GC- and AT-rich regions of the genome[98]. Despite these issues, Illumina sequencing technologies remains one of the most common for a variety of sequencing applications including metagenomics and *de novo* genome sequencing of relatively small genomes.

While Illumina employs short read sequencing, Pacific Biosciences (PacBio) single molecule, real-time (SMRT) sequencing technology aims to sequence extremely long strands of DNA to improve alignment and assembly of the DNA into complete genomes. PacBio sequencing employs polymerases fixed to the bottom of nanophotonic structures called the zero-mode waveguide (ZMW). These nanofabricated structures allow for the detection of a single fluorophore. Rather than employing modified nucleotides that terminate the based extensions, this technology uses nucleotides with a fluorophore linked to the terminal phosphate such that when the nucleotide is added to the DNA sequence, the fluorophore is removed along with the phosphate group and can be detected. The real-time nature of this design can be used for the determination of kinetics for each base addition allowing for the identification of DNA modifications, such as methylation[99]. Compared to Illumina sequencing SMRT sequencing offers much longer read lengths up to 20,000 base pairs. The single molecule approach avoids signal degradation over time as can be observed in Illumina sequencing. Longer read lengths provide less fragmented genome assemblies[87], but the error rates for PacBio sequencing are much higher: while Illumina can provide as low as 0.1% errors in a single read, SMRT sequencing yields error rates of 11-14%[87,93]. However, the error in PacBio sequencing is random, meaning that it does not suffer from the sequence bias nature of some Illumina errors. Since this means that the errors observed are unlikely to occur at the same locations in the DNA sequence, increasing the sequence coverage to at least 8x coverage can drastically reduce error rate to below 1%[87,93]. While there are advantages to SMRT sequencing compared to Illumina, the cost of Illumina sequencing is as low as $22.00 per gigabase (Gb) using an Illumina HiSeq 3000/4000 compared to a cost of approximately $1,000 per Gb on the PacBio RS II system[93]. SMRT sequencing allows for improved assembly of genomes containing a large amount of long repeat regions that would be more difficult to accurate assemble with shorter read lengths. Both of these technologies have demonstrated relevance in a variety of sequencing applications for analysis and discovery including both DNA and RNA sequencing.

1.5.1. Genomics

These different technologies have clear application to genome sequencing, but genomics encompasses much more than just obtaining complete DNA sequences for a given organism. Deciphering the human genome, for example, was carried out with the goals of understanding human evolution, identifying genetic causes for disease, and accelerating biomedical research[90,91]. Genome sequencing leads directly to the identification of individual genes as well as the regulatory elements that control their expression. Furthermore, each of these genes encodes for a specific protein that maintains an important function. The field of functional genomics seeks to utilize the copious amounts of data produced by both DNA and RNA sequencing projects to describe gene functions. The Encyclopedia of DNA Elements (ENCODE) Project was created with the purpose of compiling a catalog of all structural and functional components of the human genome[100]. This lead to the construction of databases containing annotated genetic information including gene sequences from different organisms for which the proteins and their function have been characterized, such as NCBI[101] and InterPro[102] databases. The Joint Genome Institute also is leading the 1000 Fungal Genomes Project[80] to expand the sequence search space for fungal genes and support the Fungal Nutritional ENCODE Project. Using these databases sequencing-based alignments can be performed to predict the function of a given gene based on the similarity of its sequence to another sequence of known function[103]. This bioinformatic analysis can provide valuable information for sequences that would otherwise be difficult or impossible to study in more detail.

Other branches of genomics include epigenomics and metagenomics. An epigenome is the complete set of all the epigenetic modifications contained within the genetic material of a cell[104]. Epigenetics describes features of the genetic material that affect expression of a given gene and therefore can change the phenotype observed without changing the genetic code itself. Epigenetic changes include DNA methylation, chemical modifications to histone proteins, and chromatin structure[105-107]. These changes can affect how easily a gene can be expressed by making them more, or less, accessible to the transcriptional machinery responsible for gene expression[104,108]. These features and their effect on gene expression is not fully understood, but is an active area of research to explain phenomena that cannot be explained by the genetic sequence alone.

Metagenomics seeks to profile the members of a dynamic microbial community. In nature, microorganisms are rarely found in isolation, but rather are part of consortia in which each organism plays a specific role such that it can benefit from and help the other members of the community. In fact, most of the organisms found in these communities have proven to be extremely difficult or nearly impossible to culture in isolation[109]. Metagenomic sequencing is commonly applied to profile these consortia by sequencing a highly-conserved region of DNA, typically regions associated with the ribosome: 16S rRNA gene in prokaryotes and archaea; 18S rRNA in eukaryotes[110,111]. For fungal specific identification, the internal transcribed spacer (ITS) regions located between ribosomal subunit genes in the genome has also been employed for profiling[77,78]. Metagenomic sequencing has also been employed to sequence the genomes of all members of a community in a full metagenome to identify functional capabilities of the consortia using gene annotation methods[63]. Metagenomics provides a powerful ability to sequence naturally existing communities and obtain an understanding of how they work together.

1.5.2. Transcriptomics

While Illumina and PacBio technologies can be used to sequence genomic information, they can also be applied to the sequencing of RNA for a variety of applications. Converting RNA strands to their complementary DNA (cDNA) using reverse transcriptase enzymes allows for the application of the same DNA sequencing methods[112]. A transcriptome is the full set of RNA molecules (mRNA, tRNA, rRNA, etc), often focusing specifically on mRNA molecules, expressed by a cell or population of cells, and is used to complement genomic sequence for gene identification[112,113]. While several methods exist for *ab initio* identification of genes using Hidden Markov Models (HMM) for prokaryotic genomes (e.g. - GLIMMER[114] and GeneMark[115]) and eukaryotic genomes (e.g. - SNAP[116] and GENSCAN[117]), transcriptomes provide some of the best gene annotations and are capable of identifying genes that the *ab initio* methods miss[118]. In this case, sequencing reads from RNA-seq experiments can be aligned directly to an assembled genome to more clearly delineate gene locations.

There are several common methods to assemble mRNA transcriptomes from RNA sequencing data. When a reference genome is already available, alignment based strategies are typically used to align reads from an RNA-seq experiment directly to the genome to assemble full transcripts and genes. For reference based alignments, the procedure commonly includes the use of TopHat[119] and Bowtie[120] to rapidly align the short reads obtained from the sequencing platform to the reference genome. Bowtie performs simple, rapid alignments, while TopHat allows for the discovery of splice sites when mapping mRNA sequences to genes containing introns[119,120]. The mapped reads are then fed into the Cufflinks package to obtain a final transcriptome assembly[121]. When there is no reference sequence available, *de novo* assembly methods must be used. The current standard for *de novo* transcriptome assembly is the Trinity platform that consists of three separate steps (Inchworm, Chrysalis, and Butterfly)[122]. These steps assemble reads into unique sequences, cluster overlapping sequences using de Bruijn graphs, and report all plausible transcript sequences including alternatively spliced isoforms[122]. Methods such as Trinity provide an important opportunity to study organisms for which genomic acquisition is difficult, like anaerobic gut fungi, on a sequence level.

RNA sequencing can also be used to obtain gene expression level information. Before the widespread availability of NGS equipment that exists today, these types of experiments relied on DNA microarrays that allow for the quantification of gene expression for a known subset of genes. In contrast, RNA-seq is not limited to existing genomic sequences, is not subject to high background noise, has a larger dynamic range, and requires lower RNA input[112]. Now with high-throughput NGS platforms, RNA-seq experiments can be performed to study the differential expression of genes in a given organism or cell type across a variety of conditions, with or without a reference genome and at a reasonable cost. Transcript quantification can be completed using RNA-seq data using a software package like RSEM (RNA-Seq by Expectation Maximization)[123]. This package can use reference transcriptomes or genomes to align RNA-seq reads and obtain quantitative information expressed in raw expected counts as well as normalized Reads Per Kilobase of transcript per Million reads (RPKM) and Transcripts Per Million (TPM)[123]. The normalized RPKM and TPM can be used for more direct comparison, but subsequent differential expression analysis packages that contain their own normalization methods, such as EdgeR[124] and DESeq[125]/DESeq2[126], require the raw expected counts as an input. The DESeq packages analyze RSEM expression data from different samples determining the $\log_2$-fold change in expression compared to a specified base condition using the expression levels across replicate samples and across all sequences in the transcriptome to determine the statistical significance of change in expression[125,126]. Altogether, these transcriptomic methods provide insight into the genes that are actively expressed and how these genes are regulated in response to changes in the cellular environment.

2. Isolation and characterization of novel gut fungal species

2.1. Introduction

Anaerobic gut fungi are a largely understudied class of microorganisms that have exciting potential for bio-based processing. Gut fungi possess a comprehensive array of biomass degrading enzymes that allow them to efficiently break down plant material[11]. It is this trait that makes them an asset in their native microbiome within the guts of ruminants such as cattle, goat, and sheep, allowing these animals subsist on a diet of crude plant material[69]. Unfortunately, there is very little sequencing data available for the anaerobic gut fungi is generally limited to short sequences used for phylogenetic analysis and genus identification[77]. Only recently the first full genome was released for an anaerobic gut fungus; the genome of *Orpinomyces* sp. C1A[79]. However, the *Orpinomyces* genome assembly is relatively poor quality, containing more than 30,000 scaffolds, indicating a fragmented assembly[79].

To enable deeper exploration into anaerobic gut fungal genes and genomes, we have isolated several additional strains from animals at the Santa Barbara Zoo. We have characterized their growth on a variety of substrates ranging in complexity from simple sugars such as glucose and fructose to cellulose and crude biomass (reed canary grass, switchgrass, corn stover, and alfalfa stems). We have also sequenced the transcriptomes and genomes of several strains of fungi, enabling identification of biomass degrading enzymes, gene characteristics, and regulatory elements such as antisense RNA and promoter sequences.

2.2. Results and Discussion

2.2.1. Isolation of novel microbes from fecal material

Several species of anaerobic gut fungi were isolated from the fecal material of herbivores at the Santa Barbara zoo. Cultures were started by mixing fecal samples in anaerobic culture media and isolated strains were obtained through a series of three single colony selections. Isolated strains include *Anaeromyces robustus*, *Neocallimastix californiae*, *Caecomyces churrovis*, and *Neocallimastix* sp. S1. *A. robustus*, *C. churrovis*, and *Neocallimastix* sp. S1 were isolated from Navajo churro sheep and *N. californiae* was isolated from a San Clemente Island goat at the Santa Barbara Zoo. These fungi were initially observed microscopically during the isolation procedure to maximize the diversity of characterized strains.

N. californiae is a monocentric fungus that forms only a single sporangium on each unit of vegetative growth (thallus) while *A. robustus* is polycentric, capable of forming multiple sporangia from a single center of growth[62] (Figure 2.1). While this results in a significant morphological difference between the two fungi, it is unclear what, if any, metabolic differences are correlated with this attribute. Figure 2.1 illustrates the vegetative growth of each fungus and their extensive network of ramifying, tapering rhizoids growing into particles of crude reed canary grass. This growth morphology was consistent with cultures grown in the absence of plant biomass on soluble substrates (Figure 2.1). Fungal rhizoids aid in plant breakdown via mechanical disruption and work in conjunction with secreted enzymes to deconstruct biomass[127] and likely increase the biomass surface area to enhance degradation by other cellulolytic bacteria[66]. Due to this capability, gut fungi are considered the primary colonizers of plant biomass in their microbiome despite comprising less than 8% of microbial biomass[66] in the rumen and less than 1.5% of the genes identified in rumen metagenomes[65].

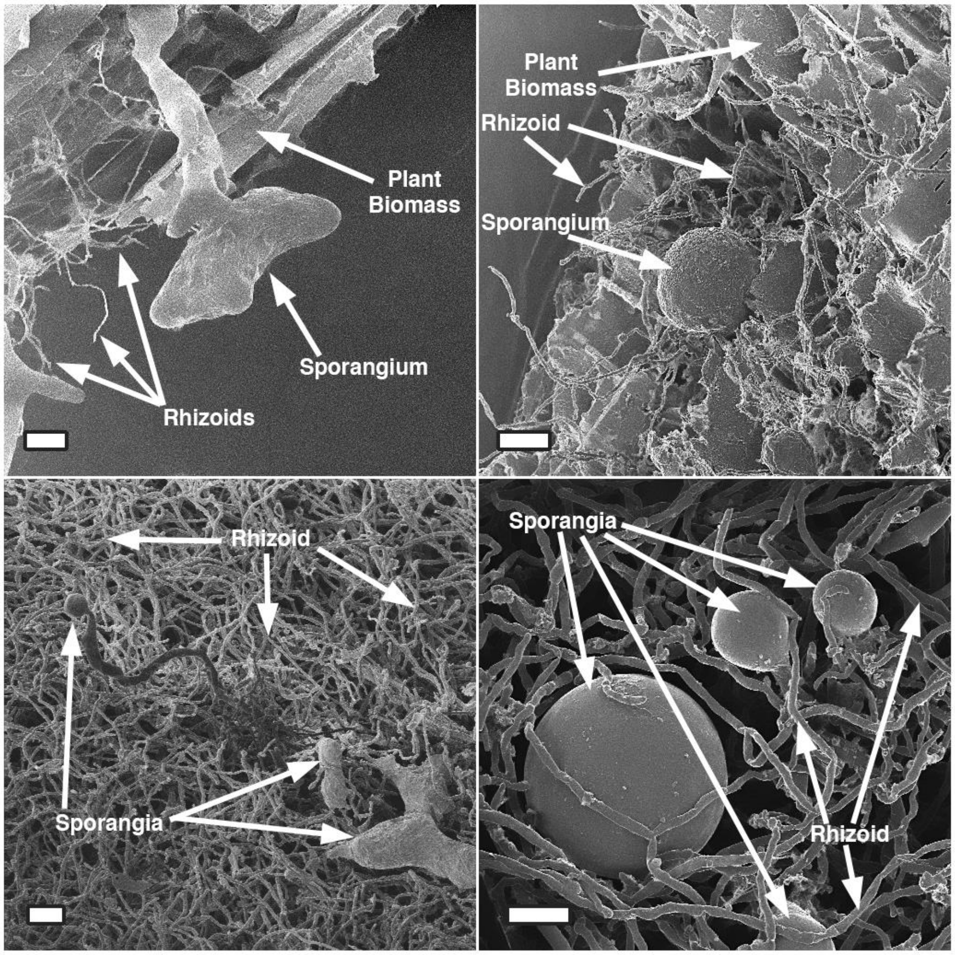


Figure 2.1. Gut fungi possess extensive rhizoidal network that penetrates into crude biomass

Helium ion micrographs of the sporangial structures of two recently classified gut fungal strains growing on lignocellulosic biomass. *Anaeromyces robustus* (top left) and *Neocallimastix californiae* (top right) grown on reed canary grass form root structures that penetrate the plant material. The same fungi grown on soluble a sugar, glucose, (*A. robustus* bottom left, *N. californiae* bottom right) still grow extensive root networks in the absence of plant biomass. All scale bars represent 10 micrometers.

C. churrovis is a monocentric fungus like *N. californiae*, however, unlike both *N. californiae* and *A. robustus*, it does not possess the extensive rhizoidal network that works to penetrate biomass. *C. churrovis* forms a large, spherical sporangium with minimal mycelia, just long enough to attach to plant biomass and other solid substrates (Figure 2.2). Microscopy shows that although there is not an extensive mycelial network formed by the fungi that aids in biomass disruption, these fungi still localize to and cover the surface of the plant biomass particles. Figure 2.2 shows sections of plant material nearly entirely covered in *C. churrovis* sporangia. Furthermore, the right image highlights the lytic life cycle of gut fungi, as a large sporangium has ruptured, releasing the cellular contents and motile zoospores. These zoospores will then go on to seek a carbon source, like biomass, and begin the process again, forming new sporangia colonies.

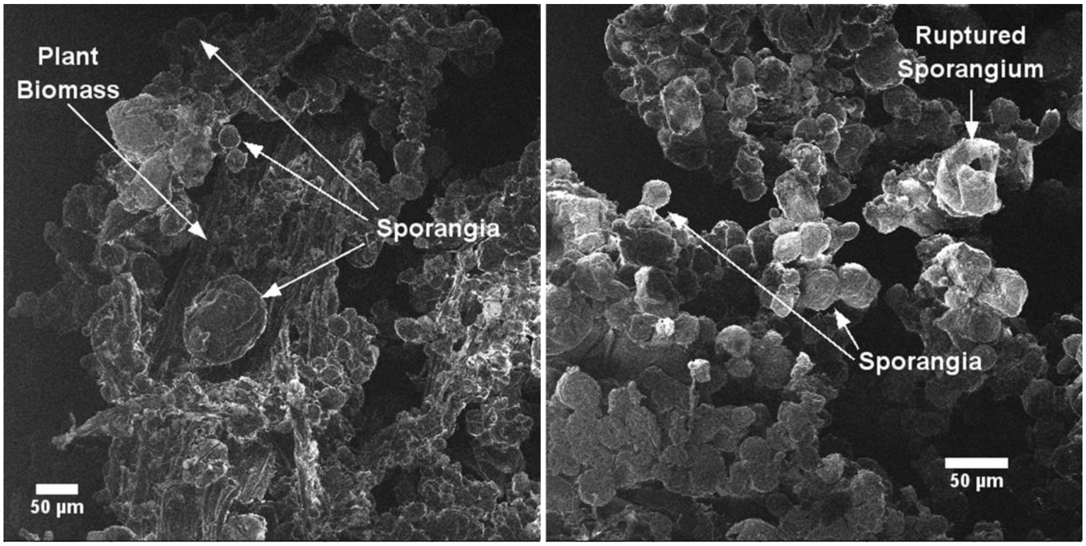


Figure 2.2. *Caecomyces churrovis* growth on reed canary grass

Culture of *C. churrovis* grown on crude plant material (reed canary grass) highlights the spherical sporangia and lack of extensive mycelial network. The fungus shows a wide range of size of sporangia, likely due to different phases of the growth cycle. In the image on the left, the reed canary grass is visible and *C. churrovis* sporangia are attached to it. The image on the right shows a section of plant material completely covered in sporangia such that the plant biomass is no longer visible. This image also shows a ruptured sporangium that has broken open to let out the motile zoospores as part of the gut fungal reproductive cycle.

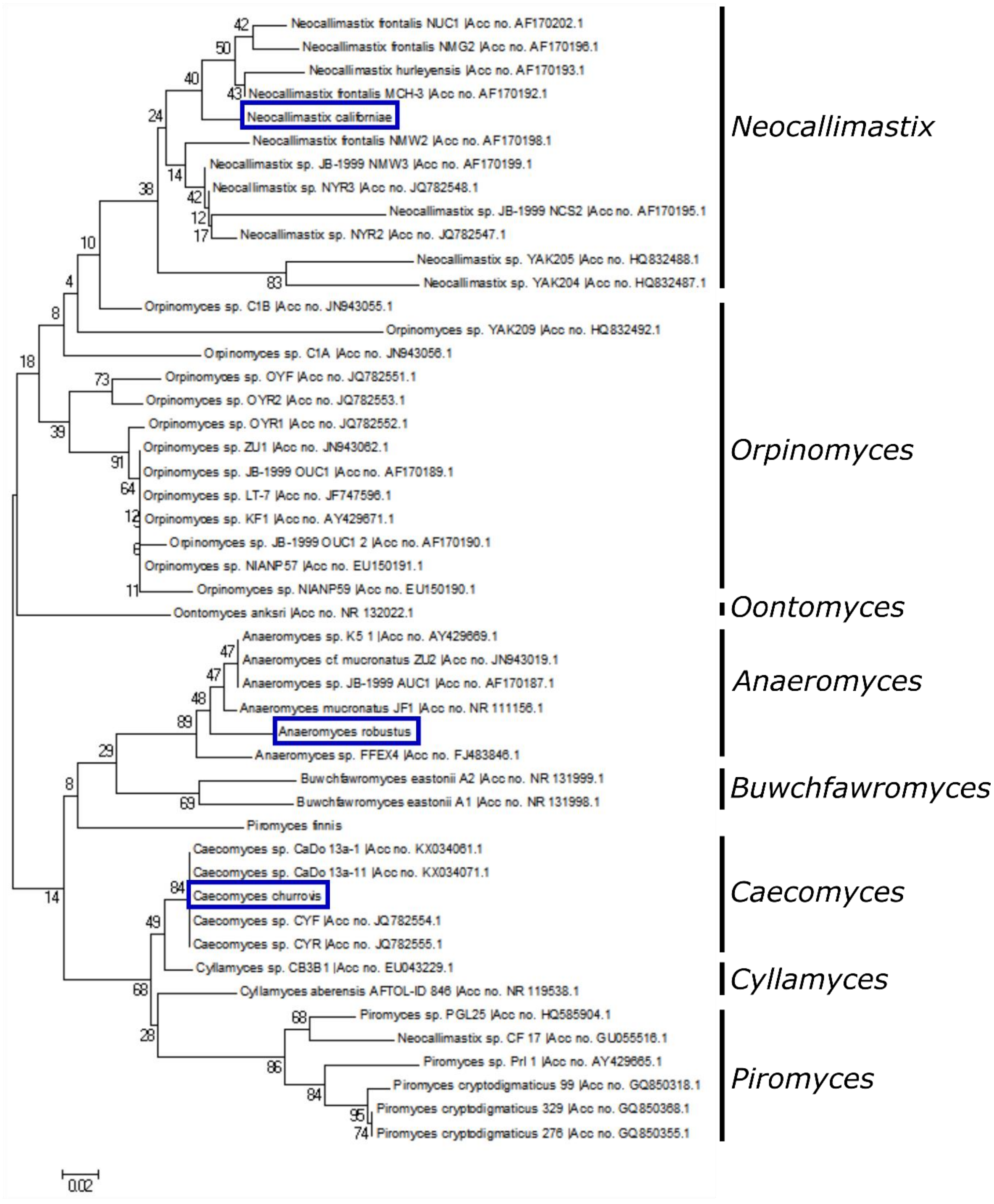

42
Neocallimastix frontalis NUC1 |Acc no. AF170202.1
50
Neocallimastix frontalis NMG2 |Acc no. AF170196.1
40
Neocallimastix hurleyensis |Acc no. AF170193.1
43
Neocallimastix frontalis MCH-3 |Acc no. AF170192.1
24
Neocallimastix californiae
Neocallimastix frontalis NMW2 |Acc no. AF170198.1
14
Neocallimastix sp. JB-1999 NMW3 |Acc no. AF170199.1
38
42
Neocallimastix sp. NYR3 |Acc no. JQ782548.1
12
Neocallimastix sp. JB-1999 NCS2 |Acc no. AF170195.1
17
Neocallimastix sp. NYR2 |Acc no. JQ782547.1
10
Neocallimastix sp. YAK205 |Acc no. HQ832488.1
4
83
Neocallimastix sp. YAK204 |Acc no. HQ832487.1
8
Orpinomyces sp. C1B |Acc no. JN943055.1
Orpinomyces sp. YAK209 |Acc no. HQ832492.1
18
Orpinomyces sp. C1A |Acc no. JN943056.1
73
Orpinomyces sp. OYF |Acc no. JQ782551.1
Orpinomyces sp. OYR2 |Acc no. JQ782553.1
39
Orpinomyces sp. OYR1 |Acc no. JQ782552.1
91
Orpinomyces sp. ZU1 |Acc no. JN943062.1
Orpinomyces sp. JB-1999 OUC1 |Acc no. AF170189.1
64
Orpinomyces sp. LT-7 |Acc no. JF747596.1
12
Orpinomyces sp. KF1 |Acc no. AY429671.1
6
Orpinomyces sp. JB-1999 OUC1 2 |Acc no. AF170190.1
Orpinomyces sp. NIANP57 |Acc no. EU150191.1
11
Orpinomyces sp. NIANP59 |Acc no. EU150190.1
Oontomyces anksri |Acc no. NR 132022.1
Anaeromyces sp. K5 1 |Acc no. AY429669.1
47
Anaeromyces cf. mucronatus ZU2 |Acc no. JN943019.1
47
Anaeromyces sp. JB-1999 AUC1 |Acc no. AF170187.1
48
Anaeromyces mucronatus JF1 |Acc no. NR 111156.1
89
Anaeromyces robustus
29
Anaeromyces sp. FFEX4 |Acc no. FJ483846.1
8
Buwchfawromyces eastonii A2 |Acc no. NR 131999.1
69
Buwchfawromyces eastonii A1 |Acc no. NR 131998.1
Piromyces finnis
Caecomyces sp. CaDo 13a-1 |Acc no. KX034061.1
Caecomyces sp. CaDo 13a-11 |Acc no. KX034071.1
14
84
Caecomyces churrovis
49
Caecomyces sp. CYF |Acc no. JQ782554.1
Caecomyces sp. CYR |Acc no. JQ782555.1
Cyllamyces sp. CB3B1 |Acc no. EU043229.1
68
Cyllamyces aberensis AFTOL-ID 846 |Acc no. NR 119538.1
68
Piromyces sp. PGL25 |Acc no. HQ585904.1
28
Neocallimastix sp. CF 17 |Acc no. GU055516.1
86
Piromyces sp. PrI 1 |Acc no. AY429665.1
84
Piromyces cryptodigmaticus 99 |Acc no. GQ850318.1
95
Piromyces cryptodigmaticus 329 |Acc no. GQ850368.1
74
Piromyces cryptodigmaticus 276 |Acc no. GQ850355.1
0.02
Neocallimastix
Orpinomyces
Oontomyces
Anaeromyces
Buwchfawromyces
Caecomyces
Cyllamyces
Piromyces

Figure 2.3. Gut fungal phylogeny using ITS1sequences

Alignment of gut fungal ITS1 sequences clearly groups our isolated strains *A. robustus*, *N. californiae*, and *C. churrovis* (boxed in blue) with strains of conserved genera, allowing for genus level identification.

Phylogenetic analysis for these strains of fungi was completed using a molecular bar coding approach that employs the sequencing of a highly-conserved region of DNA. In the case of anaerobic gut fungi, as well as other fungi, the internal transcribed spacer region is commonly used to determine the genus of a newly isolated fungus[77,78,128,129]. The three strains of gut fungi described here were aligned to other members of the Neocallimastigaceae family representing all known genera of anaerobic gut fungi using ITS1 sequences. From this analysis (Figure 2.3), *N. californiae* clearly clusters with ITS1 sequences from other *Neocallimastix* isolates, *A. robustus* clusters with other *Anaeromyces* fungi, and *C. churrovis* clusters with other *Caecomyces* fungi (Figure 2.3). Similarities in these sequences allows for the assignment of a putative genus for each of the fungi. Furthermore, the alignments revealed no identical matches among other cultured microbes, indicating that these strains of gut fungi were unique species not cultivated previously.

2.2.2. Gut fungi are powerful degraders of lignocellulose

Anaerobic gut fungi are a valuable, untapped resource for lignocellulosic bioprocessing due to their innate ability to degrade crude biomass through abundant secretion of diverse carbohydrate active enzymes[130]. However, they are immensely understudied compared to current industrial microbes, lacking genetic tools for metabolic engineering, and have not yet been adopted as biotechnology platforms. We have characterized the biomass-degrading activity of three unique anaerobic gut fungal isolates, classified as *Neocallimastix californiae* (IF551675), *Anaeromyces robustus* (IF551676)[131], and *Caecomyces churrovis* that are attractive for bio-based production applications. Each fungus was grown on a variety of carbon sources ranging from simple monosaccharides (i.e. – glucose, fructose, arabinose) to cellulose and complex biomass. The biomass substrates used are USDA energy crops reed canary grass, switchgrass, alfalfa stems, and corn stover.

Table 2.1. Effective net specific growth rates by substrate

	Substrate	***Neocallimastix californiae*** **(x10^{-2} hr^{-1})**	***Anaeromyces robustus*** **(x10^{-2} hr^{-1})**	***Caecomyces churrovis*** **(x10^{-2} hr^{-1})**
Hexose Sugars	***Glucose***	5.0 ± 1.6	11.0 ± 2.2	5.0 ± 0.29
	Galactose	ND	ND	ND
	Fructose	4.2 ± 0.72	9.8 ± 1.6	6.3 ± 0.84
	Mannose	ND*	ND	ND
Pentose Sugars	**Arabinose**	ND	ND	ND
	Xylose	ND	ND*	ND
Disaccharides	**Cellobiose**	5.9 ± 0.79	9.2 ± 1.3	4.3 ± 0.76
	Maltose	4.3 ± 0.73	9.6 ± 1.4	ND
	Sucrose	4.6 ± 0.39	ND	ND
Crystalline Cellulose	**Avicel**	8.4 ± 1.3	9.2 ± 2.4	1.9 ± 0.15
	Sigmacell	7.3 ± 1.1	9.2 ± 0.96	1.6 ± 0.30
	Carboxymethyl Cellulose	ND	ND	ND
Hemicellulose	**Xylan**	ND	ND	3.8 ± 3.0**
Lignocellulose	**Reed Canary Grass**	6.4 ± 0.72	7.2 ± 0.72	5.1 ± 0.84
	Corn Stover	4.6 ± 0.12	6.5 ± 0.47	4.9 ± 0.41
	Switchgrass	5.5 ± 0.91	2.2 ± 0.45	3.5 ± 0.27
	Alfalfa Stems	6.8 ± 1.8	4.7 ± 3.2	1.6 ± 0.05

ND: Growth Not Detected on substrate

***** Inconsistent replication on these substrates resulted in indeterminable growth rate, only 1/3 of the cultures tested demonstrated growth

** Xylan from corn stover used in *C. churrovis* growth experiment, xylan from beechwood used for other fungi

All three strains of gut fungi thrive on substrates ranging from simple sugars to cellobiose, cellulose, and lignocellulose. *N. californiae* maintained almost no change in net specific growth rate across all substrates, with the fastest growth rates measured on complex biomass and cellulose rather than simple sugars. *A. robustus* and *C. churrovis* demonstrated slightly faster growth on simple sugars compared to complex biomass (Table 2.1). Nonetheless, these results demonstrate that gut fungal growth is not largely inhibited by the complexity of plant biomass and the additional metabolic burden involved in the production and secretion of a wide array of biomass degrading enzymes. All three of the fungi were capable of growth on glucose, fructose, cellobiose, Avicel, and complex biomass. Although *Caecomyces* struggled to grow on alfalfa stems compared to the other two fungi. *N. californiae* and *A. robustus* also demonstrated growth on maltose while only *N. californiae* grew on sucrose. While gut fungi have been documented to grow on xylose in the past[132], *N. californiae* and *Caecomyces* displayed no growth while *A. robustus* displayed inconsistent growth on xylose in fungal batch culture, perhaps due to subtle environmental cues (e.g. pH) that may govern xylose assimilation. None of these fungal isolates grew on xylan or carboxymethyl cellulose (Table 2.1). Interestingly, while all fungi demonstrated growth on purified crystalline cellulose, *Caecomyces* struggled to grow on this substrate with long lag times and slow overall growth. Since *Caecomyces* fungi do not have rhizomycelia that aid in the penetration of biomass and other substrates, it is expected that the tight packing of crystalline cellulose particles at the bottom of the culture tubes limits the ability of the zoospores and sporangia to access the cellulose beyond the surface. Since the other two fungi have mycelial roots that can disrupt biomass structure, they are not inhibited in this way.

These results identify strengths and limitations in the carbohydrate utilization profile of each strain that could be exploited for consolidated bioprocessing purposes. For example, two sugar constituents of hemicellulose, galactose and arabinose, did not support growth of the gut fungi in isolation, but are expected to be liberated during lignocellulose digestion. These sugars may serve as metabolic links to second organism that can catabolize these substrates, discussed in more detail in Chapter 4.

2.2.3. Transcriptome sequencing and analysis

Transcriptomes for these anaerobic gut fungi were sequenced and assembled to develop an understanding of the genes that they express and those that are responsible for their biomass degrading capability. Since no genomic reference was available, the transcriptomes were each assembled *de novo*. For the transcriptome acquisition, total RNA was isolated from cultures grown on a variety of substrates ranging in complexity from simple sugars to cellulose and complex biomass to maximize the number of genes captured in the transcriptome. The transcriptomes for *N. californiae* and *A. robustus* were sequenced in collaboration with the Joint Genome Institute (JGI) using an Illumina HiSeq and the Rnnotator[133] algorithm for *de novo* assembly. The transcriptome for *C. churrovis* was sequenced on an Illumina NextSeq and assembled using the Trinity[122] algorithm. These efforts resulted in transcriptomes containing 29649, 17127, and 36595 transcripts and 27671, 16038, and 33437 predicted genes in *N californiae*, *A. robustus*, and *C. churrovis,* respectively (Table 2.2). The transcriptome assemblies include gene isoforms, and therefore, the predicted number of genes excludes isoforms identified by the *de novo* assembly algorithms.

Anaerobic gut fungi are well known for their AT-rich genomes that typically makes extraction and study of high quality DNA difficult[8,84,85]. After transcriptomes were obtained, the sequences were examined to determine the distribution of the four DNA nucleotides. This analysis identified AT-content of greater than 70% in each of the isolates (Table 2.3). While AT content typically affects the stability of DNA, with AT-rich DNA molecules resulting in lower melting, or strand dissociation, temperatures[134], there are additional implications for high AT-content in coding regions of DNA. For example, heterologous expression of genes in a model organism requires careful consideration of codon usage to minimize rare codon occurrence in the host organism that will express the gene. For example, due to the high AT-content, gut fungal transcripts contain more AT-rich codons compared to model organisms such as *Saccharomyces cerevisiae*.

Table 2.2. Transcriptome sequencing and annotation statistics for *de novo* assembly

	Neocallimastix californiae	***Anaeromyces robustus***	***Caecomyces churrovis***
Transcriptome size (bp)	36,250,970	21,955,935	30,884,864
# Transcripts	29,649	17,127	36,595
# Predicted genes	27,671	16,038	33,437
Average Length (bp)	1,222	1,281	843
# Reads	153,745,938	247,076,108	233,780,238
Read length (bp)	2 x 150	2 x 150	2 x 75
rRNA contamination (%)	5.19	22.1	Not determined
Coverage	1206	2630	567.7
% With EC number	6.23%	5.83%	7.55%
% With Blast hits	8.31%	10.04%	9.33%
% With Gene Ontology	24.9%	24.37%	33.22%
% With InterPro Scan	73.1%	76.58%	72.52%

Codons encoding for amino acids are highly redundant, meaning that an organism has the option of using more than one and, in some cases, up to six different codons for the same amino acid. For all highly represented amino acids gut fungal transcripts revealed a distinct bias toward the codons with higher AT representation (Figure 2.4). Comparison of the codon preference in gut fungi to codon representation in both highly and lowly expressed genes from

S. cerevisiae and highly expressed genes in *E. coli*[135] highlights important differences in codon usage. In some cases, like serine (Ser) gut fungi, *S. cerevisiae*, and *E. coli* have similar preferences for codons. However, in the case of lysine (Lys) all three gut fungi have a strong preference for the AAA codon, while highly expressed genes *S. cerevisiae* utilize primarily the AAG codon and lowly expressed genes use AAA. However, highly expressed genes in *E. coli* use the AAA codon so it may be necessary to optimize Lysine codons for heterologous expression in *S. cerevisiae*, but not in *E. coli*. For aspartic acid (Asp) and Leucine (Leu), gut fungi prefer different codons compared to both *S. cerevisiae* and *E. coli*, using GAT to express aspartic acid and TTA to express leucine compared to GAC and TTG in *S. cerevisiae* and *E. coli*, respectively.

Table 2.3. Transcriptome nucleotide frequencies across indicated strains

	N. californiae	***A. robustus***	***C. churrovis***
%A	0.410	0.419	0.389
%C	0.123	0.115	0.141
%G	0.141	0.135	0.155
%T	0.326	0.330	0.315
%AT	0.736	0.749	0.704
%GC	0.264	0.251	0.296

These variations in codon preference across different organisms highlight the importance of careful consideration if gut fungal carbohydrate active enzymes are to be produced heterologously by model organisms and isolated for use as enzyme cocktails for biomass degradation. While other considerations in protein expression, such as glycosylation are also important for expression of functional proteins, codon optimization can be a valuable first step towards improving heterologous expression.

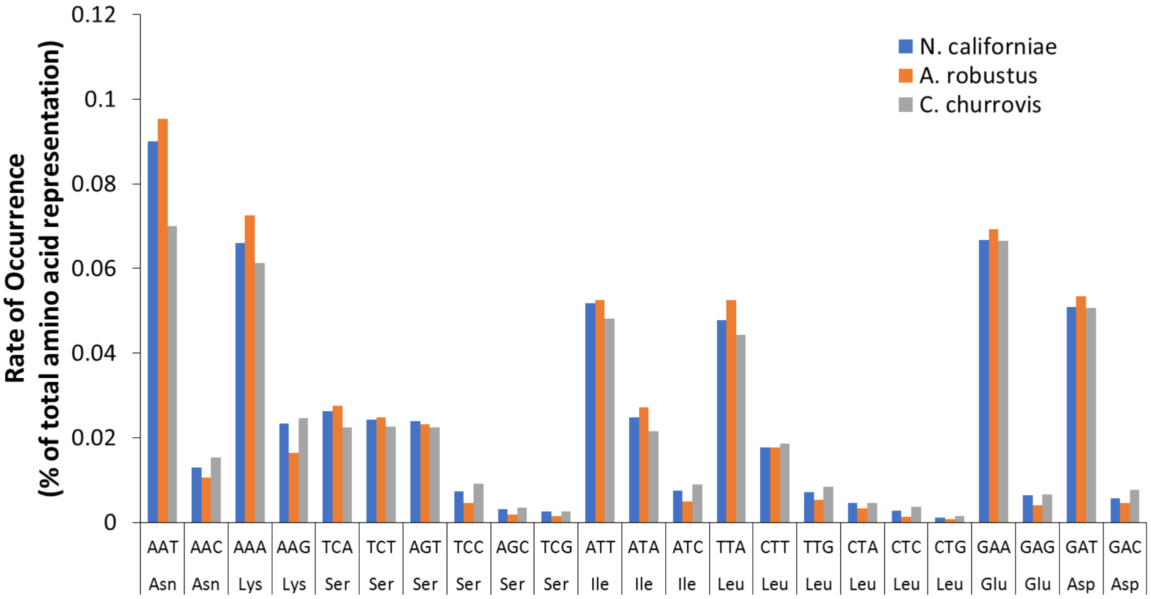


Figure 2.4. Codon usage for highly represented amino acids in gut fungal transcriptomes

Codon usage in anaerobic gut fungi reveals a clear preference for codons containing more A and T nucleic acids. The most highly represented amino acids in the coding regions of the transcriptome are represented in this graph. Clear biases for a single codon are present for asparagine (Asn), lysine (Lys), glutamic acid (Glu), and aspartic acid (Asp), but serine (Ser) uses three different codons approximately equally.

2.2.4. Transcriptome annotation reveals a wide array of biomass degrading enzymes

Functional annotation of all transcripts in the assembled fungal transcriptomes was completed using a variety of sequence alignment based techniques. These annotations allow for the building of metabolic pathways and identify important cellular functions, such as protein folding chaperones, membrane sensors and transporters, and biomass degrading enzymes. Alignments to full genes present in the NCBI database as well as to known protein domains within the InterPro database were completed to obtain a comprehensive set of functional predictions based on full sequence similarity and presence of specific protein domains, respectively. This analysis resulted in the annotation of 8-10% of each transcriptome via BLAST sequence alignment for full protein prediction, 6-8% assigned an enzyme commission number for specific enzymatic activity, 24-33% assigned gene ontology (GO) terms that classify broad function, and 72-77% annotated with protein domain functions by InterProScan (Table 2.2). Combining all this information, putative enzymatic functions present within the gut fungi can be identified.

A key feature of interest in the case of anaerobic gut fungi is their carbohydrate active enzymes (CAZymes)[136,137]. CAZymes are responsible for the gut fungi's unique ability to efficiently degrade crude biomass, but until this study the full repertoire of CAZymes in gut fungal genera was unknown. These enzymes were identified based on the protein domains using InterPro based annotations of known carbohydrate active protein domains already deposited in the NCBI and InterPro databases. These domains included glycoside hydrolases (GH) that are responsible for the hydrolysis of cellulose and hemicellulose as well as polysaccharide deacetylases, carbohydrate esterases, and pectin lyases that are responsible for accessory function necessary separate the sugar rich cellulose and hemicellulose from pectin and recalcitrant lignin. In each of these fungi, the highest represented CAZyme families fall under the hemicellulase and accessory function classes (Figure 2.5). This highlights the importance of these enzymes in the degradation of complex biomass. Most of the sugar within biomass will be released from cellulose, but to access cellulose the enzymes must first break though the outer layers of pectin, lignin, and hemicellulose. This requires the action of enzymes such as carbohydrate esterases and polysaccharide deacetylases to separate lignin from hemicellulose, pectinases and pectin lyases to break down pectin, and glycoside hydrolase families that are specialized in the variety of sugar-sugar bonds found in hemicellulose.

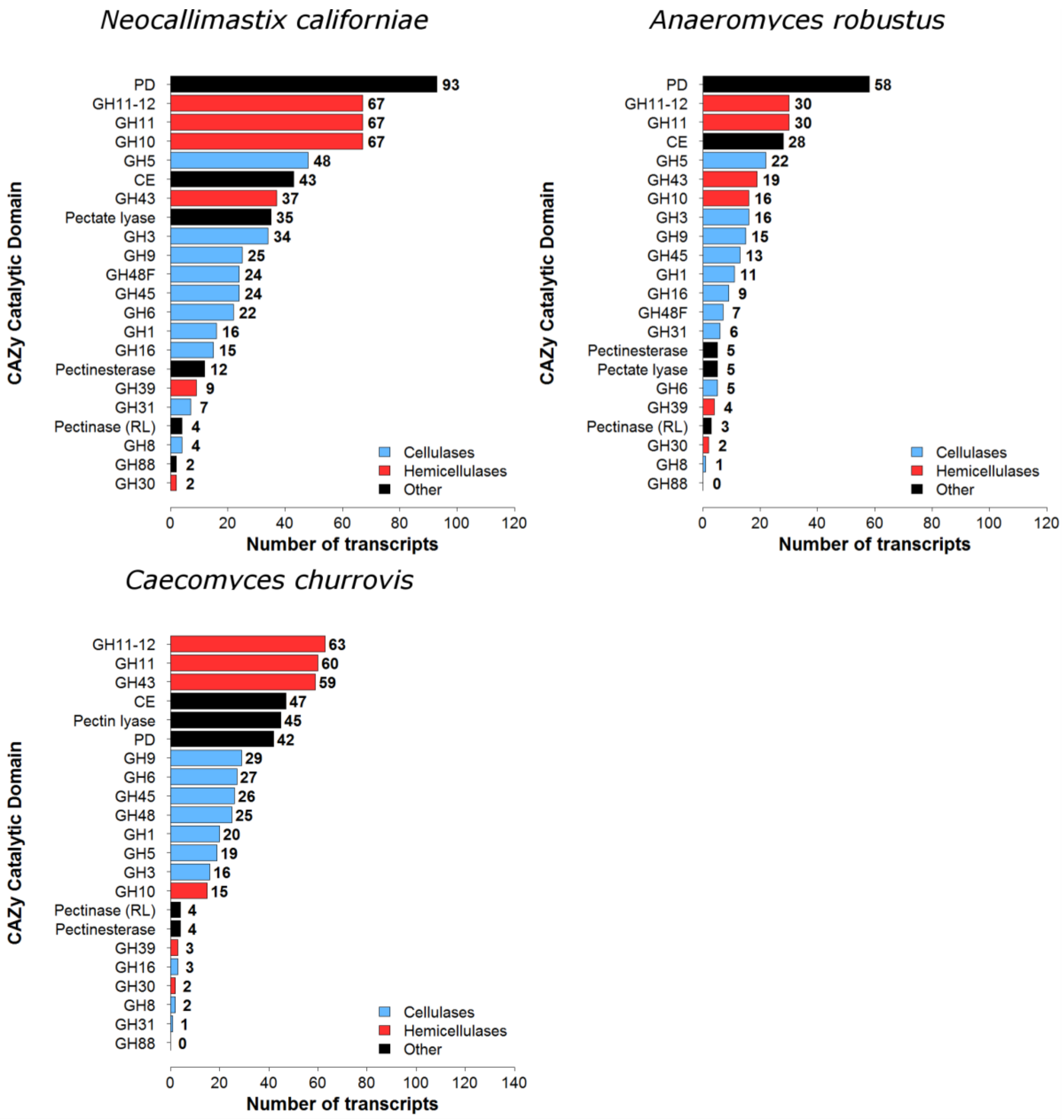


Figure 2.5. Breakdown of carbohydrate active enzymes in gut fungi.

Each of the three species of gut fungi sequenced here contain a wide array of enzymes required to break down complex biomass. These functions include cellulases (blue), hemicellulases (red), and accessory functions (black) involved in hydrolyzing sugars from plant material. The presence of all of these functions makes gut fungi fantastic degraders of biomass.

Cellulolytic function within all three of these fungal species is comprised of endoglucanase (GH 5, 6, 8, 9, 45), exoglucanase/cellobiohydrolase (GH48), and β-glucosidase (GH 1, 3) protein domains. Endoglucanases hydrolyze internal β-1,4-glucosidic bonds, cellobiohydrolases move processively acting on the ends of the cellulose chain to release cellobiose molecules, and β-glucosidases hydrolyze cellobiose into two molecules of glucose[3]. The primary hemicellulase domains identified were xylanase (GH 10, 11) and xylosidase (GH 39, 43). Though other activities may be present within these families – for example arabinofuranosidases are also commonly found in the GH 43 family – these are the typical functions identified. Other accessory enzymes identified that aid in hemicellulose digestion and separation from lignin included polysaccharide deacetylases, carbohydrate esterases, pectinases, and pectin esterases. Hemicelluloses are a heterogeneous biopolymer comprised of various sugars and bond types[19]. Therefore a greater diversity of enzymes are required to hydrolyze hemicellulose than are necessary for cellulose hydrolysis; this includes endoxylanases, xylosidases, arabinofuranosidases, glucuronidases, and a variety of esterases[138].

Additionally, these annotations can be used to identify a wide variety of additional enzyme activities and proteins and can be annotated and used to build metabolic pathways. This analysis is completed in Chapter 4 for the sugar catabolic pathways in *N. californiae* and *A. robustus* as they pertain to the use of gut fungi to supply sugars in a two-microbe fermentation scheme. Briefly, the analysis identified complete catabolic pathways for glucose, fructose, and xylose, but incomplete pathways for galactose and arabinose.

2.2.5. Genome sequencing of novel fungal isolates

The genomes of three anaerobic gut fungi (*Neocallimastix californiae*, *Anaeromyces robustus*, and *Piromyces finnis*) were sequenced to obtain the genome localization of transcripts in the transcriptomes leading to identification of regulatory DNA sequences, such as promoters. Genome sequencing can also identify additional genes that may not have been expressed under the growth conditions used for transcriptome acquisition. Due to the high AT-content and high percentage of repeat regions present in gut fungal genomes, assembly of short sequence reads, like those obtained from Illumina sequencing technologies is difficult. Therefore, Pacific Biosciences long-read SMRT (Single Molecule Real Time) sequencing was used almost exclusively.

Table 2.4. Genome assembly yields AT- and repeat-rich genomes

	A. robustus	***N. californiae***	***P. finnis***
Genome Assembly size (Mbp)	71.69	193.03	56.46
# of contigs	1,035	1,819	232
Contig N50/L50 (Mbp)	158/0.14	134/0.44	25/0.75
% GC content	16	22	21
% Repeats	56.8	65.8	51.4
Total gene number	12,939	20,393	11,477
# Transcripts mapped to genome	15,190	25,262	15,543
% Transcriptome mapped to genome	88.70%	85.20%	91.40%

The genome sequencing efforts resulted in the most complete gut fungal genomes sequenced to date with the fewest number of scaffolds. Previous efforts had yielded a genome for *Orpinomyces* sp. C1A of 100.95 Mbp (Mega base pairs) with 32,574 contigs (set of overlapping DNA segments)[79], and a genome for *Piromyces*sp. E2 of 71.02 Mbp with 17,217 contigs. The genomes of *A robustus*, *N. californiae*, and *P. finnis* yielded genomes of 71.69

Mbp and 1,035 contigs, 193.03 Mbp and 1,819 contigs, and 56.46 Mbp and 232 contigs, respectively. Of these, *A. robustus* had the lowest GC content with only 16% G and C nucleotides; *N. californiae* and *P. finnis* had 22% and 21% GC content, respectively. Furthermore, much of these genomes, 50-65%, were comprised of repeat regions. These two features highlight the challenges of sequencing and assembling the genomes of these organisms and why long read sequencing was necessary to align and assemble these highly AT and repeat-rich genomes.

When comparing the genome and transcriptome sequencing results, we identified that 85-91% of the transcriptomes obtained for each of these fungi were present in the genomes, indicating high quality *de novo* transcriptomes were created for each of these fungi (Table 2.4). With high quality genomes, development of genome-scale metabolic models is enabled. Genome localization can also be combined with transcript regulation data to identify potential promoters for control of the expression of heterologous proteins. Furthermore, they can be used to more effectively identify antisense transcripts that play a role in the regulation of protein expression.

2.2.6. Genomic and transcriptomic data reveal regulatory DNA sequences

The analysis of high-quality genomic assemblies allows for the identification of regulatory DNA sequences, such as promoters, that are responsible for controlling the transcription of genes into their corresponding mRNA. Transcript sequences were aligned to the genomes to provide genomic localization and subsequent searching of adjacent upstream regions presents a valuable starting point for the identification of putative promoter sequences. These promoter sequences can then be validated by molecular biology techniques such as cloning and expression studies. Sequencing techniques such as DNase-seq, ATAC-seq, FAIRE-seq, and

ChIP-seq that target specific DNA regions based on their accessibility to or interaction with regulatory proteins[107,139]. To support the prediction of promoter sequences, transcriptional regulation information was used to help identify conserved DNA sequences. We used substrate-based transcriptional regulation information (discussed in more detail in Chapter 4) to identify candidates for cellobiose triggered regulation as well as candidates for high level, constitutive expression in both fungi. Examples of each type of expression pattern for genes from *N. californiae* are shown in Figure 2.6.

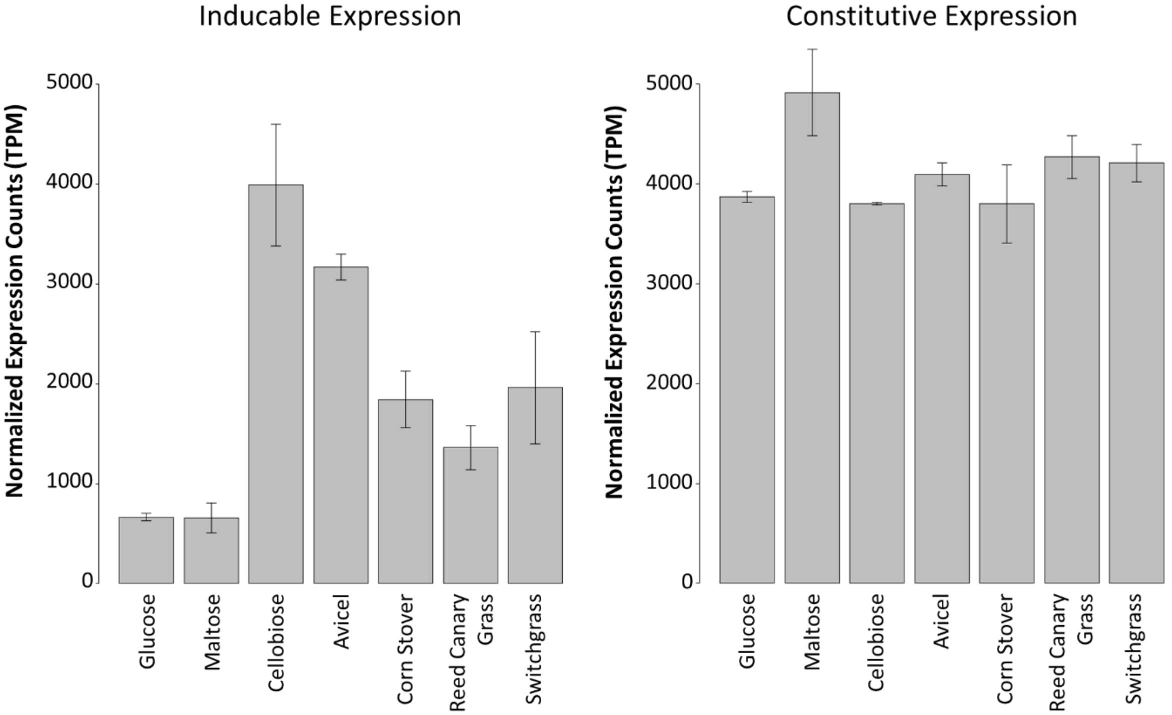


Figure 2.6. Expression of candidate induced and constitutive genes in *N. californiae*

Transcriptional regulation by varying substrates can identify regulation patterns typical of induced and constitutive expression. Here are shown candidate genes that are likely under inducible control (left) and constitutive control (right). The candidate for induced control shows low expression on glucose and maltose, with higher expression on other substrates, suggesting the growth conditions can tune expression. The regions of DNA upstream of these genes are likely to provide promoters for these expression strategies.

Using the genomic loci for transcripts that followed these regulation patterns indicative of cellobiose induction, we extracted the 2 kb of DNA sequence upstream of the gene. The upstream promoter regions were then grouped based on their level of regulation (by $\log_2$-fold change in regulation compared to expression on glucose). These promoter regions were then analyzed using motif finding algorithms in the MEME Suite[140], specifically Multiple Em for Motif Elicitation (MEME)[141] to find novel, gap-free motifs conserved among the promoter regions of similarly regulated genes. Alignment of 14 promoter sequences from *Anaeromyces robustus* for genes with $\log_2$-fold change between 3 and 4 resulted in the prediction of three possible motifs.

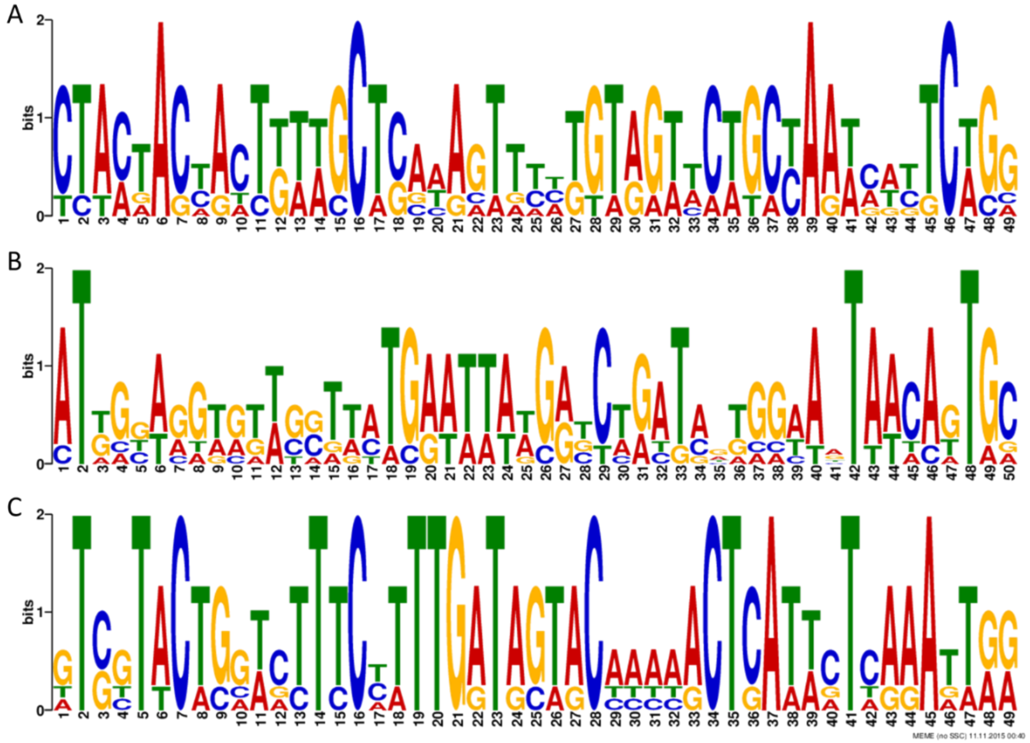


Figure 2.7. Motif identification in *A. robustus* GH promoters with similar regulation patterns

Promoter sequences identified for *A. robustus* GH transcripts that were regulated in response to growth cellobiose with a $\log_2$-fold change in expression between 3-4 were aligned to search for motifs. Three motifs were identified using the MEME motif finding tool[141]. These motifs were found in 6 (A), 7 (B), and 5 (C) of the 14 sequences used in the alignment.

While the sequences identified by the motif finder do not seem to show a high level of confidence in a conserved, continuous sequence, they were more highly populated by G and C nucleotides compared to the entire upstream regions. The upstream regions maintained a typical GC-content of approximately 15%, however, the conserved regions identified by the motif finder maintain GC-content of ranging from 30-50%. It is possible then that higher GC content is needed for effective promoters in the AT-rich genomes of anaerobic gut fungi. It is also possible, however, that the motif finding algorithm is less effective at the extreme AT-content found in the genomes of these fungi and is therefore biased towards relatively GC-rich regions of DNA. This analysis provides a valuable starting point for the identification of useful promoters for expression of heterologous genes in anaerobic gut fungi. In order to validate these sequences, however, it is necessary to develop tools to transform gut fungi such that they may be tested and a true minimal promoter region can be identified.

2.2.7. Antisense provides a mechanism for regulation in anaerobic gut fungi

Alignment of transcriptomic data to the genomes of newly isolated fungi has proven useful in the identification of promoter regions, but it can also be used to examine antisense RNA (asRNA) as a specific mechanism for post-transcriptional regulation of gene expression. Antisense transcripts are encoded on the antisense strand of DNA, or the strand opposite the sense strand that encodes for a protein coding gene. Expression of antisense sequences can inhibit expression of a protein-coding gene in several ways including inhibition of transcription initiation of the coding sequence, transcriptional inhibition via co-transcription with the coding sequence, and RNA:RNA duplex formation to induce instability in the coding mRNA[142]. Such antisense regulation mechanisms have been identified in all domains of life including filamentous fungi[143-147]. We identified natural antisense transcripts (NATs) as well as their predicted targets by aligning sequences from the transcriptomes of the fungi to the completed genomes of *Anaeromyces robustus*, *Neocallimastix californiae*, and *Piromyces finnis*. Antisense transcripts were identified if they aligned to the genome opposite of a protein encoding transcript. This process identified 439 (2.5% of transcriptome), 732 (2.5%), and 1586 (9.3%) NATs in *A. robustus*, *N. californiae*, and *P. finnis*, respectively. NATs targeted a variety of functions within the cell including protein expression, metabolism, and a small amount of lignocellulose hydrolysis genes. Differential expression analysis was then used to study the regulation of NATs and their targets.

Expression data from cultures grown on glucose, cellobiose, crystalline cellulose (Avicel®), and reed canary grass were used to identify mechanisms of regulation (Figure 2.8). The mechanisms identified were: antisense regulated – the target expression level is consistent across conditions tests, but the antisense expression level changes; transcriptionally regulated – antisense expression remains unchanged and the target is transcriptionally regulated; and coregulated – antisense and target expression are both regulated either in the same direction, or opposite directions. These mechanisms are depicted in heat maps that demonstrate how expression changes across the different substrate conditions tested (Figure 2.8). The presence of these three modes of regulation were well conserved among all three fungal isolates tested.

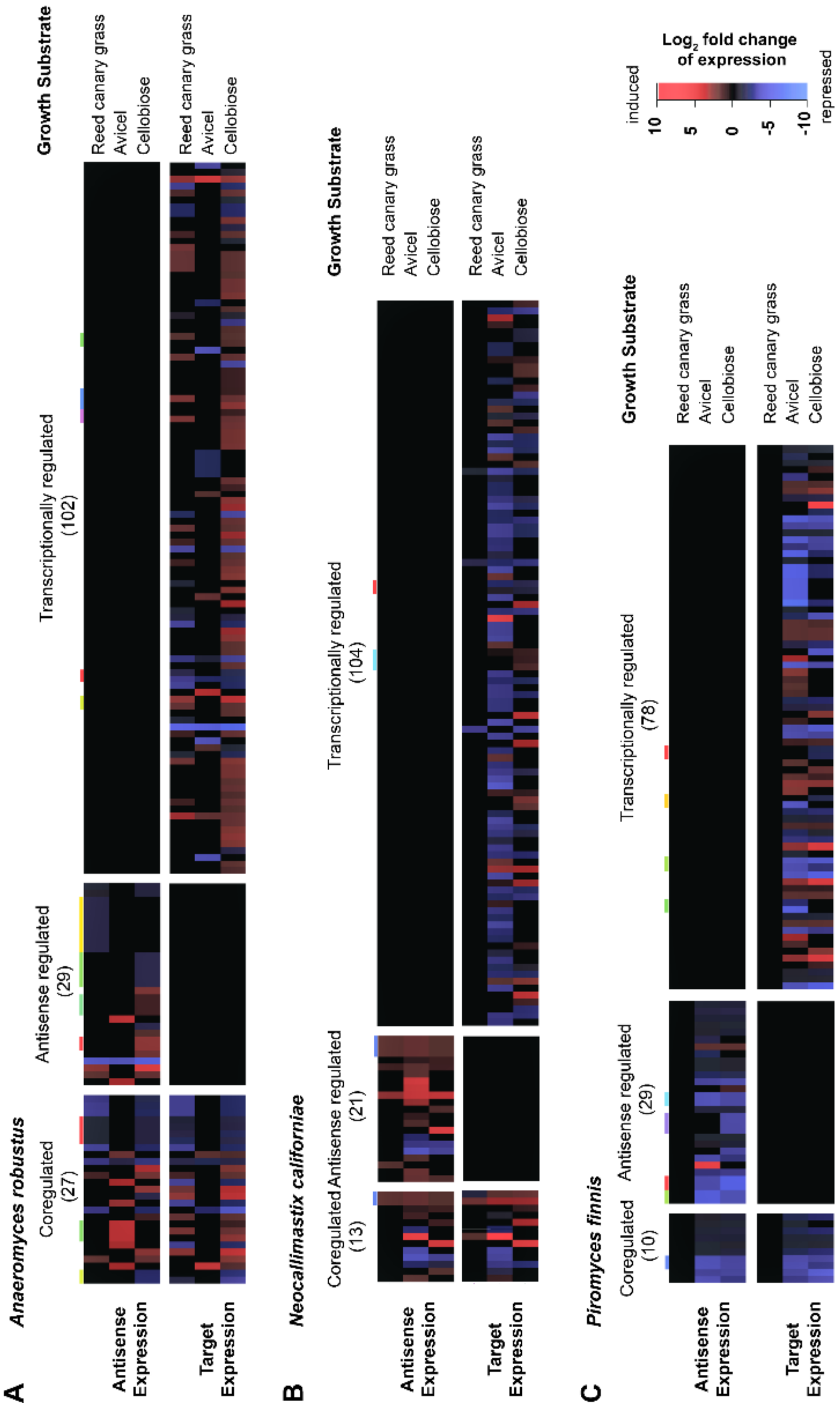


Figure 2.8. Antisense transcripts and their targets are regulated to control expression

Analysis of differentially expressed genes (log$_2$-fold change ≥ 1; $p \leq 0.01$) were binned by whether or not the target and antisense transcription were regulated. Results for *Anaeromyces robustus* (A), *Neocallimastix californiae* (B), and *Piromyces finnis* (C) revealed three unique modes of regulation – coregulated, antisense regulated, and transcriptionally regulated target.

Analysis of the relative expression level of NATs and their targets compared between glucose and reed canary grass growth conditions identified clusters of NAT and target pairs that follow the same regulation mechanism (Figure 2.9). When NAT expression was greater than that of its target under both growth conditions (lower left quadrant of the plots) regulation was dominated by changes in antisense expression rather than target. When NAT expression was much lower than the target under both conditions (top right quadrant), the regulation was dominated by transcriptional regulation of the target gene itself rather than the NAT. Co-regulation of NAT and target occurred primarily in cases where the expression of NAT and target were similar under both growth conditions. These regulation patterns are consistent with a model where NATs are used to fine tune gene expression. Cells maximize the dynamic range of expression outputs by varying the expression of whichever transcript is dominant[142].

While overall abundance of NATs in these organisms was relatively low (< 10% in all isolates and 2.5% in two), they yielded the same mechanisms of control across all three isolates. We identified three classes of mechanisms using NATs to fine tune gene expression: NAT expression dominated, target expression dominated, and co-regulation. This analysis highlights the importance of genomic information for NAT and target identification and profiling to develop a full understanding of gene regulation that may not be explained by simple transcript expression profiling in anaerobic gut fungi.

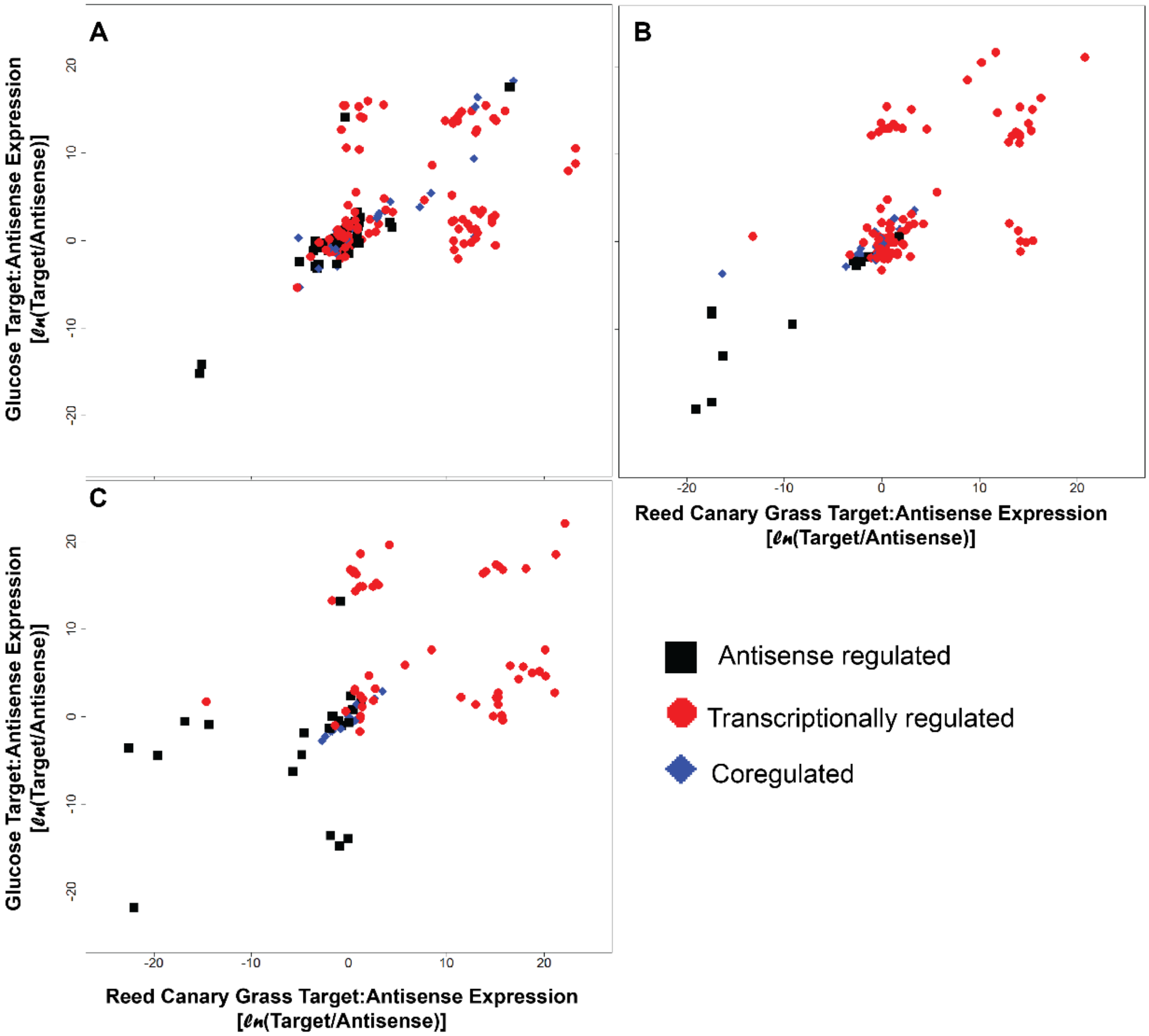


Figure 2.9. Relative expression of regulated NAT and target gene pairs

Relative expression levels of NAT and target in transcripts per million (TPM) were measured and compared between glucose and reed canary grass growth conditions for *A. robustus* (A), *N. californiae* (B), and *P. finnis* (C). NAT and target pairs are colored by method of gene regulation as determined in Figure 2.8.

2.3. Conclusions

Using classical anaerobic microbiological techniques, novel species of anaerobic gut fungi were isolated from the fecal material of mammalian herbivores. In this case, we have isolated three novel species of gut fungi from goats (*N. californiae*) and sheep (*A. robustus* and *C. churrovis*) housed at the Santa Barbara Zoo. We used phylogenetic analysis of the ribosomal internal transcribed spacer (ITS) region as a first step to molecular characterization. This identified the genus of each of these species based on similarity to ITS sequences from other gut fungi and identified that they were significantly different from other characterized species to date. Microscopic analysis highlighted the significant morphological differences between each of these strains. *C. churrovis* represents the largest divergence from the morphology of the other two isolates as there is no mycelial root network to aid in plant biomass breakdown. Based on our growth characterization, it seems that this missing characteristic inhibits effective digestion of purified crystalline cellulose. However, all species of gut fungi examined here grow well on crude biomass with growth rates comparable to growth on simple carbon sources, highlighting their effectiveness at hydrolyzing biomass.

Extensive transcriptomic characterization provided functional annotations to identify key enzymes involved in biomass breakdown. The diversity of carbohydrate active enzymes (CAZymes) including cellulases, hemicellulases, deacetylases, and esterases produced by the fungi allow them to efficiently hydrolyze biomass without any pretreatment to separate lignin from cellulose and hemicellulose. Acquisition of complete genomes provided an opportunity for more in depth study of these microbes. We identified putative regulatory elements and possible conserved promoter sequences by combining genomic localization information with regulation data. We also used alignments of transcriptomes and genomes to identify putative antisense RNA that plays a role in the regulation of protein expression. These genomes also present an opportunity to identify novel proteins[82] and build genome-scale metabolic models. Overall, this work highlights the potential of these unique microbes for exploitation in industrial bioprocesses and builds the foundation for future application of these unique microbes.

2.4. Materials and Methods

2.4.1. Isolation and culture maintenance

Gut fungi were isolated from the fecal material of animals at the Santa Barbara Zoo. Fresh fecal material was collected, ground, and suspended into culture Medium C[148]. Next five serial dilutions were performed. From each serial dilution, triplicate cultures were inoculated and monitored for growth signified by accumulation of fermentation gases in the head space of the sealed culture tubes. Cultures that demonstrated growth were sustained through routine transfers into culture media. To obtain an isolated strain of fungus, 25 mL tubes coated with 5 mL of solid Medium C containing 2% agar were inoculated with 0.1 mL of growing culture. These roll tubes were grown for 2-3 days after which single colonies were selected by cutting colonies out of the agar and transferring to a new liquid culture tube in a procedure performed in a box under a constant flow of CO_2 to maintain anaerobic conditions. This process was completed three times for each strain of gut fungus to ensure selection of a single, isolated strain.

Anaerobic gut fungi were routinely grown in 10 mL batch cultures of Medium C[148] containing ground reed canary grass (4 mm particle size) in 15 mL Hungate tubes. The tube headspace was filled with 100% CO_z and cultures were grown at 39°C. Cultures were transferred to new media every 3-5 days to continue growth.

2.4.2. Phylogenetic analysis

Phylogenetic analysis was completed by sequencing the internal transcribed spacer region for each of the isolated fungi. ITS sequences were PCR amplified using the JB206 (GGA AGT AAA AGT CGT AAC AAG G) and JB205 (TCC TCC GCT TAT TAA TAT GC) primers[78] that amplify fragments start in the small rRNA subunit (18S) gene, ending in the large rRNA subunit (28S) gene, and spanning ITS1, 5.8S, and ITS2 regions. The amplified DNA was sequenced and the ITS1 region was employed in phylogenetic analysis. ITS1 sequences were obtained for other anaerobic gut fungi across all known genera. The phylogenetic tree was created using Molecular Evolutionary Genetic Analysis (MEGA) software version 6.0[149]. Sequences were aligned using the Clustal Omega multiple sequence alignment method[150,151], and the alignment was used to construct phylogeny using the neighbor-joining statistical method. To test the confidence of the phylogeny, a bootstrap method was used with 1000 replications.

2.4.3. Helium Ion Microscopy

Helium ion microscopy was completed at the Pacific Northwest National Laboratory (PNNL) Environmental and Molecular Sciences Laboratory (EMSL) by James Evans, Chuck Smallwood, and Vaithiyalingam Shutthanandan. Fungi grown on various substrates were chemically fixed with 2% glutaraldehyde (Sigma Aldrich) and dehydrated through a series of 10 mL step-gradients from 0% to 70% ethanol then centrifuged at 4°C (3000Xg for 2 mins). The biomass was washed twice more with 10mL of 100% ethanol for 15 mins, then centrifuged and finally resuspended in 5mL of 100% ethanol to remove any residual water. Fungal and/or plant biomass suspensions in 100% ethanol were gently extracted by wide-mouth pipet and placed onto stainless steel carriers for automatic critical point drying (CPD) using an Autosamdri-815 (Tousimis, Rockville, MD), with CO_2 as a transitional fluid. The CPD-processed biomass was mounted onto aluminum stubs and sputter coated with approximately 10 to 20nm of conductive carbon to preserve the sample surface information and minimize charge effects. Secondary electron images of the samples were obtained using Orion helium ion microscope (HIM) (Carl Zeiss Microscopy, Peabody, MA) at 25 or 30 keV beam energy, with a probe current range of 0.1 to 1 pA. Prepared samples were transferred into the HIM via load-lock system and were maintained at ~3×10^–7 Torr during imaging. Use of a low energy electron flood gun (~ 500 eV) was applied briefly interlaced with the helium ion beam that enabled charge control to be maintained from sample to sample. The image signal was acquired in line-averaging mode, with 16 lines integrated into each line in the final image with a dwell time of 1µs at a working distance range of 7 to 8 mm. Charge neutralization was applied to the sample after each individual line pass of the helium ion beam, which displaced charges on the surface minimizing charging effects in the final image. No post-processing procedures were applied to the digital images besides standard noise reduction, brightness and contrast adjustment using Photoshop plugins.

2.4.4. Growth curve generation

Growth curves were generated by measuring the pressure of fermentation gases during growth. Accumulation of pressure in the headspace of the closed Hungate tubes is correlated to fungal growth and inversely correlated to substrate loss[152]. Soluble substrates were present at a concentration of 5g/L and insoluble substrates were present at a concentration of 10 g/L. Cultures that accumulated pressure significantly more than the blank control (10 mL Medium C culture containing no carbon source, but inoculated with fungi) were considered positive for growth. Effective net specific growth rates were determined from the pressure accumulation data of 3 x replicate cultures during the phase of exponential gas accumulation.

2.4.5. RNA Isolation

RNA was isolated from growing fungal cultures during the exponential growth phase using the Qiagen RNeasy Mini Kit (Qiagen, Valencia, CA). The protocol for plants and fungi was followed, including a liquid nitrogen grinding step to disrupt cell walls and an on-column DNase digest. The RNA quality was determined through measurement on an Agilent Tapestation 2200 (Agilent, Santa Clara, CA) to obtain RINe scores. The total RNA quantity was determined by using Qubit Fluorometric Quantitation (Qubit, New York, NY) using the high sensitivity RNA reagents.

2.4.6. RNA sequencing and transcriptome assembly

The transcriptome of each organism was obtained using RNA isolated from cultures grown on a variety of substrates, including glucose, cellobiose, cellulose, and reed canary grass. RNA was pooled prior to generation of the sequencing library using equal quantities of RNA from each growth condition. The transcriptome for *Piromyces finnis* was sequenced by our collaborators at the Broad Institute on an Illumina HiSeq, the transcriptomes for *Anaeromyces robustus* and *Neocallimastix californiae* were sequenced by our collaborators at the Joint Genome Institute on an Illumina HiSeq, and the transcriptome of *Caecomyces* sp. A was sequenced using the Biological Nanostructures Laboratory core sequencing facility's Illumina NextSeq. After pooling libraries were created using an Illumina Truseq Stranded mRNA library prep kit (Illumina Inc., San Diego, CA) following the kit protocol. Transcriptomes for *N. californiae* and *A. robustus* were sequenced with greater than 1000X coverage and assembled *de novo* using Rnnotator[133].*Caecomyces* sp A transcriptome was sequenced with greater than 500X coverage and assembled *de novo* using Trinity[122].

2.4.7. Transcriptome annotation

The transcriptomes were annotated using the automated BLAST2GO package[153]. First, transcripts were analyzed for sequence homology using the blastx program against the NCBI non-redundant database with an E-value cutoff of 10^{-3}. Transcripts were then analyzed for protein domains using alignment to sequences in the EMBL-EBI InterPro database before gene ontology[154] terms and enzyme commission[155] numbers were assigned. Due to strand specificity of the library, transcripts with BLAST hits in a reverse orientation (reading frames -1, -2, -3) were non-coding and flagged as antisense transcripts (asRNA). All transcripts were examined for orthology by comparing all possible open reading frames to the OrthoMCL database using a BLAST-based alignment against genomes from all domains of life[156]. Sequences with significant hits across taxa were assigned as orthologs and grouped into ortholog groups.

2.4.8. DNA Isolation

Genomic DNA was isolated from cultures grown for 5-7 days to allow for accumulation of a larger amount of cellular material. Culture were grown on glucose to reduce the interference of plant material during cell lysis. DNA was extracted using the MoBio PowerPlant Pro kit, which proved to be the optimal method for isolation of high molecular weight DNA (see Chapter 3 for more information). To obtain the required quantity of DNA (>12 µg) for submission to the Joint Genome Institute for sequencing with Pacific BioSciences single molecule real-time (SMRT) sequencing DNA was isolated from 5-10 cultures grown in 40 mL volumes and pooled together by collecting the DNA in the same silica column. This process was repeated until the total amount of DNA isolated was greater than 12 µg.

2.4.9. DNA sequencing and genome assembly

Genomes for *A. robustus*, *P. finnis*, and *N. californiae* were sequenced at the Department of Energy Joint Genome Institute using the Pacific Biosciences platform. To prepare PacBio libraries, gDNA was treated with DNA damage repair mix followed by end repair and ligation of SMRT adapters using the PacBio SMRTbell Template Kit (Pacific Bioscience of California Inc., Menlo Park, CA). DNA was sheared to 10kb fragments using the g-TUBE™ (Covaris) or templates were size selected using a Sage Science BluePippin instrument with a 10kb minimum cut off. PacBio Sequencing primer was then annealed to the SMRTbell template libraries and the sequencing polymerase was bound to them. The prepared libraries were then sequenced on a PacBio RSII sequencer using 4-hour sequencing movie run times. Genomes were assembled with Falcon (Pacific Biosciences) and improved with FinisherSC[157] except for *N. californiae* that was polished with Quiver[158].

2.4.10. Promoter analysis

Promoter sequences were determined by aligning the transcriptome to the genome for each fungus to identify the location of each gene. Then the 2kb region of DNA upstream of the gene was identified as the putative promoter region and extracted based on the scaffold location of the gene. To identify motifs, promoter sequences were grouped based on the regulation patterns for their corresponding genes. These groups of promoter regions were then fed into the MEME motif finding algorithm[141] to identify conserved nucleic acid sequences.

2.4.11. Antisense analysis

Antisense transcripts were first identified based on the orientation of their alignment to BLAST hits during the transcriptome annotation process. The transcriptomes were obtained using a strand specific library and any annotation in a negative reading frame (-1, -2, -3) flagged a transcript as a candidate for antisense RNA. The transcriptomes were then mapped to their corresponding genomes using GMAP[159] with a strict cutoff of > 80% complementarity and mappring length no greater than 3 times the length of the transcript. From this list of verified mappings, antisense candidates were validated if they mapped to a target transcript as cis-natural antisense transcripts (NATs) whose function are given by the annotations of their target mRNA[130].

2.4.12. Differential expression analysis

Counts of transcripts were quantified by using the RSEM analysis[123] present within the Trinity[122] programming package. Transcriptomes previously obtained[130] were used as reference templates to obtain count data. Expected counts from this analysis were then fed into the DESeq2 package[126] in the R programming language to determine statistically significant changes in expression with a minimum of one $\log_2$ fold change in expression and p-value ≤ 0.01. Results from all substrates were compared to the base case of glucose to determine fold change in expression of all transcripts.

Heat maps were made using the $\log_2$-fold change values for expression changes of each transcript. Scatter plots of relative expression levels of NAT and target were made using the raw transcripts per million (TPM)[123] output from the RSEM analysis.

3. Robust methodologies for cryogenic storage and DNA extraction

Adapted from *Anaerobe*, Vol 38, John K. Henske, Kevin V. Solomon, Michael K. Theodorou, Michelle A. O'Malley, Robust and effective methodologies for cryopreservation and DNA extraction from anaerobic gut fungi, 39-46, Copyright 2015, with permission from Elsevier.

3.1. Introduction

Anaerobic gut fungi, of the class Neocallimastigomycetes, are a promising group of underexplored organisms that efficiently degrade cellulose, hemicellulose, and pectin in crude plant biomass into their constituent sugars[11]. While the increasing demands of renewable biotechnology have renewed interest in non-model microbes with unusual properties, such as those that degrade atypical substrates or make natural products[63,160-162], many of their attributes hinder their application as industrial strains. Gut fungi exist ubiquitously in the digestive tracts of large herbivores[11] and are major contributors to the degradation of ingested plant material through their invasive, rhizoidal growth and secretion of an array of powerful enzymes that efficiently degrade biomass. However, they are exceedingly difficult to lyse, genetically manipulate[163], and to preserve in traditional culture collections.

The rigid cell walls that allow gut fungi to effectively penetrate fibrous plant biomass is also a challenge for preservation and manipulation of cell strains. Cell wall rigidity makes gut fungal cells more susceptible to damage from expansion of ice crystals formed during cryopreservation leading to poor viability beyond a few months of storage[70,164]. This complication has led to complex cryopreservation procedures that require hazardous cryoprotectants, numerous reagents, and multistage protocols that can take up to a full day to complete[165,166]. The exceptionally low oxygen tolerance of gut fungi further complicates these procedures, making even basic laboratory manipulations non-trivial. Current methods to preserve isolated strains of gut fungi are not robust and as few as 40% of all stocks retain viability after less than one year[165] and as a result gut fungal cultures are typically maintained with continual passage of cultures. Thus, only a handful of researchers that are equipped to isolate and routinely sub-culture gut fungal specimens have been able to explore their lignocellulolytic capabilities[11].

The rigid cell wall of the fungus also prevents efficient cell lysis and acts as a tough barrier against the recovery of the cell's genomic contents[167]. The chitin-rich composition of their cell wall and abundance of intracellular polysaccharides also leads to co-purification of carbohydrate contaminants that render genomic DNA unsuitable for next generation sequencing platforms[69,167-169]. Molecular characterization is further limited by low natural abundance of DNA content by cell weight[167], a consequence of the elaborate invasive growth of a single fungal thallus, and the inherent fragility of their AT-rich genomes[84]. As a result, genomic characterization of the Neocallimastigomycota has been limited with only a few published genomes[79,82].

Current state of the art techniques for working with gut fungi require the repeated passage of liquid cultures whose phenotype, and genome, likely adapt and drift over time. Therefore, development of simple, rapid, and reliable methods to both "preserve" and "break-through" cellular integrity would enable future efforts to develop promising anaerobic strains for biotechnology. Using four unique strains including one each of *Piromyces* and *Anaeromyces*, and two of *Neocallimastix*[130], we developed, modified, and compared simple, new methods for genomic DNA isolation and cryopreservation of gut fungi.

3.2. Results and Discussion

3.2.1. Robust and reliable long-term storage of gut fungi via cryopreservation

A common consequence of culturing gut fungi is their repeated transfer to fresh medium at regular 2-5 day intervals[11]. Failure to passage results in loss of culture viability due to accumulation of toxic bioproducts in the culture vessel. Thus, the use of cryopreservation is essential for the long-term storage of these strains. Cryopreservation not only safeguards culture viability but also prevents genetic drift due to selection during repeated sub-culture. To address these issues, we sought to develop a simple, robust protocol that did not require specialized equipment, and used the inexpensive and non-hazardous cryoprotectant glycerol. Unlike previous methods that (a) preserve fungi grown on soluble substrates (b) use centrifugation to pellet fungal biomass (c) cool fungal stocks in stages over several hours, and (d) demonstrate fungal viability within a range of 3 months to one year[164-166], our protocol avoids the use of pelleted biomass and preserves fungi *'in situ'* on their preferred particulate growth substrates. In essence, our protocol differs from previous methods in that it requires fewer steps, thereby reducing the risk of oxygen exposure during the preservation process. After two days growth on reed canary grass, we use anaerobic procedures to simply replace the liquid growth medium with glycerol-containing cryopreservation medium. Glycerol-incubated strains are quickly aliquoted into cryovials under a stream of CO_2, which are flash frozen in liquid nitrogen, and stored at -80°C (Methods: Chapter 3.4.6).

We tested three different glycerol concentrations (10%, 15%, and 25%) and four different fungal isolates to identify the optimal composition of cryoprotectant to promote cellular viability (Table 3.1). The four isolates represented three genera of anaerobic gut fungi: *Neocallimastix*, *Anaeromyces*, and *Piromyces* isolated from sheep, goats, and horses fed a fiber-rich diet. For each concentration, 12 cryovials in total were prepared and revived to quickly verify the feasibility of any cryopreservation. Cryovials were stored for 1-2 weeks at -80°C, thawed, inoculated in fresh media, and assessed for viability via the generation of fungal fermentation gases on reed canary grass. Of these concentrations, 10% and 15% glycerol had slightly higher success rates in short term storage (1-2 weeks) with 8 out of the 12 vials (67% of cryostocks) retaining fermentative viability (Table 3.1). In contrast, only 7 of the 12 vials tested (58% of cryostocks) were preserved in 25% glycerol (Table 3.1). Consequently, subsequent studies were conducted with 15% glycerol medium as a cryopreservant.

Table 3.1. Viability of cryopreserved fungi as a function of preservation medium

Fungal Isolate	10% Glycerol	15% Glycerol	25% Glycerol
Neocallimastix sp. S1	4/4	3/4	4/4
Anaeromyces robustus	3/3	3/3	3/3
Neocallimastix californiae	1/2	2/2	0/2

Fraction of cryostocks successfully revived for various monocentric and polycentric fungi after 2 weeks at -80°C. Cryovials were stored in one of three different glycerol concentrations. X/Y = #revived strains/#frozen strains.

Of the fungal isolates tested, *Anaeromyces* was most robust as it retained viability in all the glycerol concentrations tested (Table 3.1). Similarly, both *Neocallimastix* strains were successfully revived after storage in 10% and 15% glycerol. However, one *Neocallimastix* isolate was non-viable after storage in 25% glycerol while the *Piromyces* isolate was not successfully revived at any concentration of glycerol tested. These differences in storage stability may arise from inherent variations in the morphology and physiology of the various gut fungal genera. The preservation of the *Neocallimastix* and *Anaeromyces* isolates, but not *Piromyces*, may suggest that these two genera are more capable of forming a resistant survival structure with thicker cell walls that allows them to maintain viability in adverse conditions. Such a structure has been suggested to play a role in long term survival of fungus in liquid culture[170] and may also be important for successful cryopreservation, though further investigation is necessary to verify as these structures have never been isolated. Other contributing factors to cryopreservation survival may include their sensitivity to the microaerobic conditions formed during aliquoting, and the ability of these fungi to quickly exchange the water in their cytoplasm with cryoprotecting glycerol prior to freezing.

Given the robustness of *Anaeromyces* to cryopreservation, we employed this isolate as a model species to determine the stability of gut fungal cryostocks over the course of multiple years in storage. A culture bank of 100 cryovials were created and stored at -80°C. At periodic intervals, 5 cryovials were revived and tested for viability (Figure 3.1.A). In all intervals tested, from 1 month to 23 months of storage, all cryovials led to a vibrant culture that could be repeatedly passaged. That is, this simple and safe storage protocol with 15% glycerol as a cryoprotectant proved to be robust with 100% survival of *Anaeromyces* cultures revived between 1-23 months.

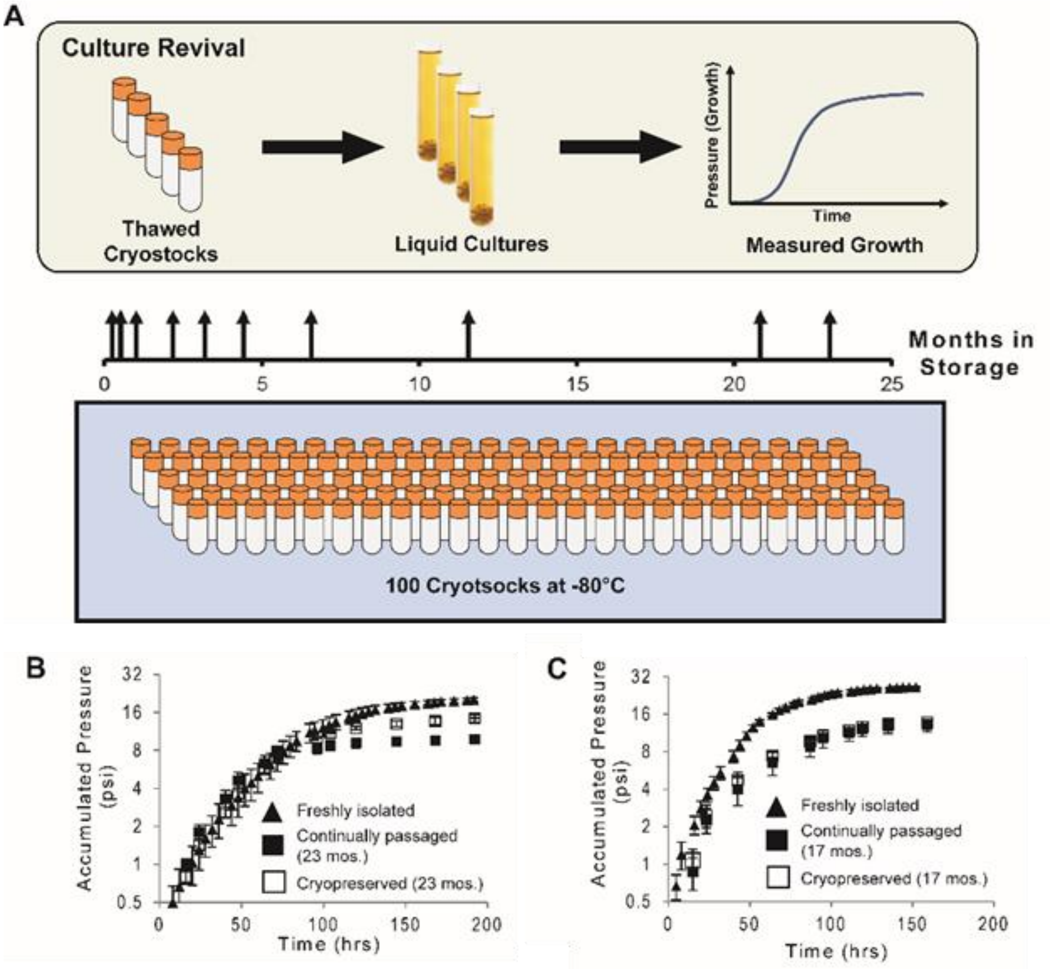


Figure 3.1. Cryopreservation of anaerobic fungal cultures promotes long-term culture viability.

A) Protocol to test long-term viability of *Anaeromyces* cryostocks. From a repository of identical cryostocks, samples were periodically thawed at the indicated intervals, inoculated, and assessed for growth. Growth curves of initial, cryopreserved and continually passaged cultures were generated for species **B)** *Anaeromyces robustus* and **C)** *Neocallimasitx californiae*.

In parallel with long-term storage at -80°C, the fungal isolate was serially passaged in liquid culture and used as a baseline to benchmark the performance of revived cryostocks. Samples of the polycentric fungus *Anaeromyces robustus* that were cryopreserved for 23 months were used to seed new liquid cultures whose growth patterns were characterized and directly compared to the continually passaged cultures (Figure 3.1.B). Comparisons were made using pressure accumulation from fungal fermentations gases[11] to calculate an effective net specific growth rate. These studies revealed no significant difference in the growth rates between the initially isolated strains, the cryopreserved strains, and cultures continually passaged every 4-5 days for 23 months (Figure 3.1). We also examined the health of cryo-stocks of a monocentric fungus, *Neocallimastix californiae*, after 17 months of storage at -80°C to demonstrate the broad applicability of this method to successfully store monocentric gut fungi long term. These cultures also demonstrated no difference in specific growth rate between the cryopreserved strains, and cultures continually passaged every 4-5 days (Figure 3.1.C). Using inexpensive and safe glycerol, our method robustly stored anaerobic gut fungi at -80°C for up to 23 months, the longest period currently reported[164-166], with a 100% survival rate for several isolates. Additionally, we have demonstrated that this method has no deleterious impact on the rate of growth of cryopreserved cultures on crude biomass when compared to cultures that were continually passaged over the same time period (Figure 3.1).

3.2.2. Preparation of high quality, high molecular weight genomic DNA

Since their initial isolation in the 1970s[71], a number of nucleic extraction methods for gut fungi have emerged to allow for basic molecular characterization (e.g. PCR, molecular cloning, taxonomic classification)[77,79,84,130,167]. However, unlike RNA that can be readily isolated from a cellular lysate with minimal degradation or contamination using commercial kits[79,130] due to its small, unmodified nature, DNA must be unpacked from chromatin, and separated from the protein/carbohydrate modifiers that mediate its activity, while not being sheared and degraded during the lysis of the tough cell wall and removal of cellular debris. As a result, DNA extractions typically rely on slow overnight precipitations, and toxic reagents that produce impure samples unsuitable for modern next generation sequencing pipelines[169]. To address this, we assessed a number of common kits and protocols for their ability to produce high molecular weight DNA with sufficient purity and integrity for next generation sequencing platforms; i.e. DNA with minimal degradation to <10 kb fragments, negligible protein contaminants (A_{260}/A_{280} range of 1.7 – 2.0), and low carbohydrate contamination (A_{260}/A_{230} range of ≥ 1.0).

One commonly used approach to isolate genomic DNA from gut fungal isolates is the FastDNA™ SPIN KIT for Soil[171], which relies on adsorption to a silica slurry (Glassmilk®) for DNA isolation and purification. Following manufacturer instructions, cells are first lysed with detergents (sodium dodecyl sulfate – SDS) and mechanical bead disruption to release the DNA for purification. Under exponential growth with a glucose substrate, gut fungal genomic DNA yields were typically 200 ng DNA/mg fungal biomass as measured by absorbance at 260 nm (Table 3.2). This DNA displayed minimal degradation to <10 kb fragments (Figure 3.2) and had an average fragment size of 15-20 kb as revealed by pulsed field gel electrophoresis (Figure 3.1.B). Samples prepared in this manner contained minimal protein contamination ($A_{260}/A_{280} = 1.8$, Table 3.2) and were suitable for routine PCR amplification[130]. However, due to the high degree of co-purifying carbohydrates ($A_{260}/A_{230} = 0.1$, Table 3.2)[167] these samples did not meet the purity standards for next generation sequencing platforms.

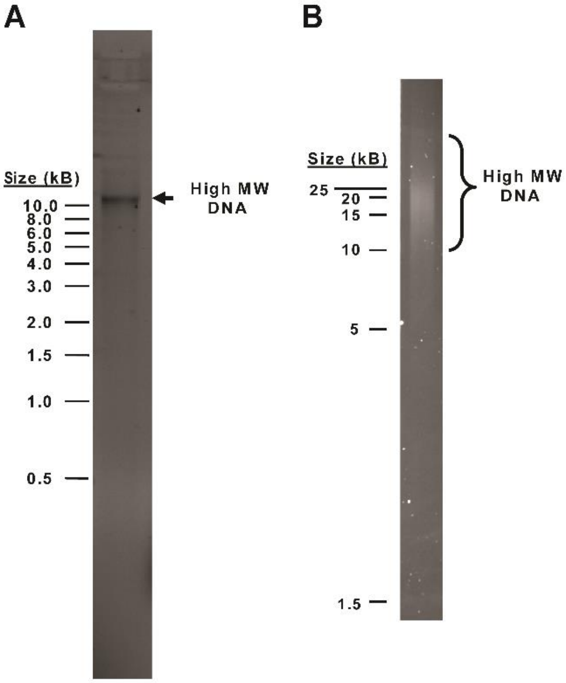


Figure 3.2. Integrity and size distribution of fungal DNA isolated from a monocentric fungus with the silican slurry adsorption method.

A) DNA as separated on native agarose gel electrophoresis at 90 V for 60 min with 1X Tris-Acetate-EDTA B) DNA via pulsed field gel electrophoresis on 0.5X Tris-Borate-EDTA. See Methods for detailed conditions. High MW DNA = High molecular weight (>10 kb) DNA.

We hypothesized that inefficient cell lysis is a critical barrier to achieving high quantities of pure genomic DNA in anaerobic fungi. While aggressive cell lysis may improve DNA yield and remove co-purifying contaminants, it likely increases DNA shearing and degradation. Conversely, inefficient lysis could reduce yield and DNA purity while improving DNA quality. Thus, we assessed the effect of cell lysis choice on DNA yield and purity (Figure 3.3). While the lysis methods tested had no significant effect on the resulting DNA purity, there was a marked impact on the amount of DNA recovered (Table 3.2). Membrane solubilization with the lysis reagent Y-PER™ greatly reduced DNA yields, likely due to an inability of the propriety detergent formulation to effectively break the tough chitin-rich cell wall of the gut fungi. Conversely, enzymatic digestion of the cell wall with lyticase, or mechanical disruption with beads was able to produce yields that were more than 2.5-fold greater than Y-PER™ lysis alone. However, combining enzymatic digestion and mechanical disruption did not greatly improve DNA yield suggesting that the standard protocol of mechanical disruption was able to recover the majority of the fungal DNA (Figure 3.1). Moreover, the elevated temperatures and buffers required for enzymatic digestion (30 °C) and Y-PER™ (65 °C) increased DNA damage with Y-PER™ producing ≤100 bp fragments and lyticase producing ~3kb fragments (data not shown). Thus, the recommended cell lysis protocol of cell lysis with bead beating and SDS solubilization was optimal for DNA yield and quality.

To improve the sample purity and reduce carbohydrate contamination, we attempted post-extraction DNA cleanup by precipitating with polyethylene glycol (PEG 8000) or ethanol. PEG precipitation was particularly attractive as it is able to selectively precipitate DNA fragments of high molecular weight (>10 kb)[172] and remove contaminating coprecipitants[167]. However, we were able to recover less than 5% of the initial DNA when precipitated with two-thirds volume of 30% PEG 8000 and 1.5 M sodium chloride. Ethanol precipitation of samples with 0.3 M sodium acetate was marginally more effective with little more than 10% of the DNA recovered and carbohydrate contamination being reduced by an order of magnitude (A_{260}/A_{230} increased to 0.70 from 0.06). To account for this loss, we assayed our samples with a DNA specific dye (PicoGreen dsDNA Assay Kit) to reveal that approximately only 25% of our starting samples were intact double stranded DNA. That is, spectrophotometric quantification greatly overestimated the abundance of quality DNA due to the presence of co-purifying contaminants[173]. More importantly, these contaminants were inhibiting the ability of the DNA to precipitate with high efficiency, potentially causing damage to individual DNA strands.

Table 3.2. Overview of DNA isolation methods tested

Method	Silica Slurry Adsorption[a]	Isopropanol Precipitation[b]	PEG Precipitation[167]	CTAB Extraction [171]	Sarkosyl/ CTAB Extraction	Silica Column Adsorption[c]
Cell Lysis	Bead beating → SDS	LN_2 Grinding → bead beating → Proteinase K	LN_2 Grinding → Proteinase K	LN_2 Grinding → CTAB	LN_2 Grinding → Sarkosyl → CTAB	Bead beating → SDS + proprietary polymer (Phenolics Separation Solution™)
Protein Removal	Acetic Acid Precipitation	Chloroform	Phenol: Cholorform: Isoamyl Alcohol (25:24:1)	Phenol: Cholorform: Isoamyl Alcohol (25:24:1)	Phenol: Cholorform: Isoamyl Alcohol (25:24:1)	Acetic Acid Precipitation
RNA Removal	N/A	RNAse A	N/A	+/- RNAse A	RNAse A	RNAse A
DNA Isolation	Glassmilk/ Silica adsorption	Isopropanol precipitation	Methoxyethanol + PEG 8000 precipitation	Isopropanol precipitation	Isopropanol precipitation	Silica adsorption
DNA Cleanup	Ethanol Wash	Ethanol Wash	Ethanol Wash	Ethanol Wash	Ethanol Wash	Ethanol Wash
Typical Yields (ng DNA/ mg biomass)	200	200	34	200	150	75
Typical A_{260}/A_{280}	1.8	1.7	2.1	1.6	1.9	1.7
Typical A_{260}/A_{230}	0.1	0.5	0.1	1.0	1.3	1.1

Typical DNA yields and quality are from the unmodified protocol. [a]FastDNA™ SPIN Kit for Soil; [b]OmniPrep™ for Fungi; [c]PowerPlant® Pro DNA Isolation Kit. A_{260}/A_{280} estimates protein contamination. Target values are 1.7-2.0. A_{260}/A_{230} estimates carbohydrate contamination, target range ≥ 1.0

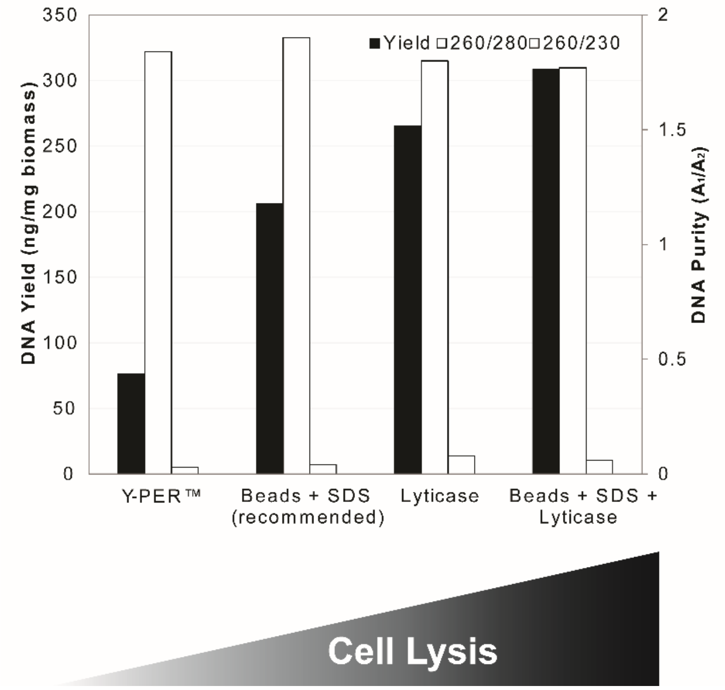


Figure 3.3. Cell lysis technique effects genomic DNA yield and quality with silica slurry adsorption.

DNA yield calculated from absorbance at 260 nm and normalized by estimated fungal mass. DNA purity estimates protein contamination (A_{260}/A_{280}, 1.7 – 2.0 is target range) and carbohydrate contamination (A_{260}/A_{230}, ≥1.0 is target range) levels from spectrophotometric absorbance. DNA isolated from *P. finnis*.

In addition to the silica slurry adsorption protocol, we tested and evaluated a number of alternate commercial kits and gut fungal genomic DNA isolation protocols for their ability to produce high quality DNA with minimal contaminants (Figure 3.4). All protocols tested, with the exception of the PEG precipitation which yielded marginal amounts of DNA, produced DNA with comparable yields and purity (Table 3.2). However, DNA quality varied tremendously. The isopropanol precipitation kit and CTAB extractions appeared spectrophotometrically to produce samples with high yields and superior purity to that of the silica slurry adsorption. However, these samples manifested as a high molecular weight band that was unable to even enter an agarose gel in a standard electrophoresis experiment (Figure

3.4, isopropanol precipitation not shown). This band was resistant to shearing with a vortexer and digestion with EcoRI suggesting that it was contaminant rich and not high molecular weight chromosomal DNA. In contrast, the Sarkosyl/CTAB protocol was able to produce DNA at high yields and purity that was able to effectively enter the gel. The dramatic improvement in DNA performance on the gel electrophoresis between the CTAB and Sarkosyl/CTAB protocols supports the earlier finding that the CTAB protocol product was primarily contamination as Sarkosyl was needed in the lysis steps to release usable DNA. Nonetheless, all CTAB-based protocols were slow, relied on toxic deproteinizing reagents such as chloroform, and produced significant amounts of RNA contamination which were not adequately removed with RNAse A (Figure 3.4).

All extraction methods applied to gut fungi in the literature were able to produce genomic DNA of varying quality. However, these samples contained persistent carbohydrate impurities, likely arising from the abundance of storage polysaccharide energy reserves and the chitin-rich cell wall of gut fungi[69,168]. Thus, we evaluated a silica spin column based method, PowerPlant® Pro DNA Isolation Kit, which was designed to remove these contaminants from tough, hardy samples such as seeds and pine needles. Using proprietary buffers, this kit was able to readily produce DNA of moderate yield at 75 ng/mg biomass with minimal carbohydrate and protein contamination (Table 3.2). More importantly, this genomic DNA was mostly intact double stranded DNA with minimal degradation products (Figure 3.4) for all isolates tested (*Neocallimastix californiae*, *Anaeromyces robustus*, *Piromyces finnis*). Similarly, the high degree of purity allowed these extracts to readily precipitate in ethanol with minimal loss of sample. This material could be amplified, concentrated and ultimately made into DNA fragment libraries for genomic sequencing with both next generation Illumina and PacBio platforms. The high quality of these preps have culminated in the most intact gut fungal genomes sequenced to date, 232 scaffolds vs. the published 32, 574[79], from organisms with genomes of at least 50 Mb and GC content as low as 17%[80].

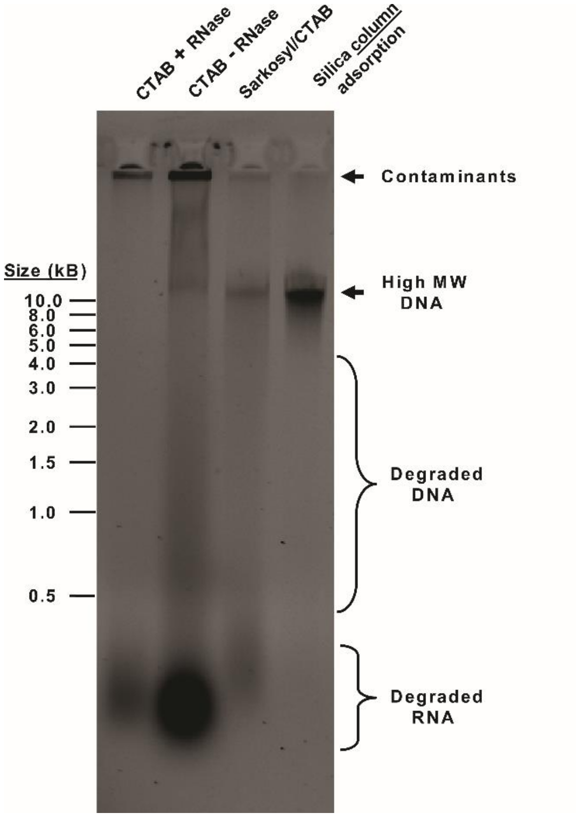


Figure 3.4. Yield and integrity of gut fungal DNA varies as a function of genome preparation.

Each lane loaded with 2 µL of anaerobic fungal genomic prep run on a 0.7% (w/v) agarose gel at 90 V, 60 min. Overview of each preparation method provided in Table 2. High MW DNA = High molecular weight (>10 kb) DNA. DNA isolated from *P. finnis*.

3.3. Conclusions

In these studies, we developed rapid, robust and inexpensive methods of cryopreserving gut fungal cultures and extracting high quality genomic DNA. We established methods to cryopreserve gut fungi long term at -80°C with safe and inexpensive glycerol. In contrast to previous methods that are complex, slow, and unreliable, our methods are robust, do not require specialized equipment, and generate cryostocks with a two-step protocol that can be completed in less than 2 minutes per vial. More importantly, cultures were viable for at least 2 years, the longest reported for any gut fungi, with no obvious impact on fungal viability. Simultaneously, we identified a commercial kit, PowerPlant® Pro DNA Isolation Kit, which was able to quickly isolate genomic DNA without contaminants common in traditional preparations. This DNA could be produced at moderate yields with minimal degradation, and of high enough purity for sequencing with next generation platforms. The resulting preps have yielded the most intact gut fungal genomes sequenced to date and will enable a wealth of new molecular level studies.

Taken together, these improved techniques catalyze future opportunities for research and development of gut fungi in biotechnology. Our ability to reliably cryopreserve gut fungal cultures long term facilitates the development of strain repositories that foster scientific collaboration between groups, and enables the development of industrial processes that can meet stringent quality control requirements. Similarly, the isolation of contaminant-free genomes ushers in a new age of research for gut fungi that can leverage the latest advances in sequencing technology to reveal new enzymes and chemistries for biochemical production. More importantly, however, we anticipate that the simplicity and ease of implementation of these techniques will increase the accessibility of gut fungi and lead to their development as interesting new model organisms for biotechnology.

3.4. Materials and Methods

3.4.1. Strains Used

Gut fungal cultures used in this study were previously isolated[130] and are listed in Table 3.3.

Table 3.3. Gut fungal strains used for DNA isolation and cryopreservation

Gut Fungal Isolates	NCBI Taxonomic ID	Isolated from	Source
Anaeromyces robustus	105135	Sheep	Santa Barbara Zoo, Santa Barbara, CA
Neocallimastix californiae	1550276	Goat	Santa Barbara Zoo, Santa Barbara, CA
Neocallimastix sp. S1	--	Sheep	Santa Barbara Zoo, Santa Barbara, CA
Piromyces finnis	45796	Horse	Verrill Farm, Concord, MA

3.4.2. Culture maintenance

Gut fungal cultures were continually passaged anaerobically in Medium C containing 15% bovine rumen fluid (Bar Diamond, Parma, ID) and supplemented with up to 0.5% of a soluble or insoluble carbon source under a 100% CO_2 headspace[70]. Fresh rumen fluid was centrifuged to remove particulates, and frozen at -20 °C in single use 75 mL aliquots until media preparation. Prepared media was dispensed in 10 mL volumes, autoclaved, and stored at 4 °C until use. Cultures (10 mL) were grown for 3-5 days in 15 mL Hungate tubes at 39 °C before passaging. During passaging, 1 mL culture was transferred to fresh medium using a sterile needle and disposable syringe. Insoluble substrates were dried and ground (1 mm dry mesh screen) prior to inclusion in Medium C.

3.4.3. Growth quantification and fungal biomass estimation

To quantify growth, periodic head space pressure measurements were taken with a pressure transducer[11]. The accumulated pressure was then used as a proxy to generate the fungal growth curve, as done previously[174]. The effective specific growth rates were calculated as the slope of the linear regime of the log-linear plot of accumulated pressure vs. culture time in hours, i.e. during exponential growth.

Fungal biomass was estimated using the correlations between culture gas volume at atmospheric pressure and biomass production established by Theodorou et al[152]. Culture gas volumes at atmospheric pressure were estimated from the measured culture pressure and headspace volume (5 mL) using Boyle’s Law.

3.4.4. DNA isolation methods

Piromyces, Anaeromyces, and *Neocallimastix* cultures were grown for 3-4 days in Medium C supplemented with 0.5% glucose before the cultures were harvested by centrifugation. The resulting cell pellets were then processed to isolate the genomic DNA.

Silica slurry adsorbed samples were prepared using the FastDNA™ SPIN KIT for Soil (MP Biomedicals, Santa Ana, CA) according to manufacturer instructions except where an alternative lysis method is noted. Y-PER™ lysed cells were resuspended in 1 mL Y-PER™ Yeast Protein Extraction Reagent (Thermo Scientific, formerly Pierce Biotechnology, Rockford, IL) and incubated at 65 °C for 10 min before being pelleted. The pellet was then resuspended in 400 µL of the FastDNA™ MT buffer and the manufacturer’s protocol continued from Step 5. Cells treated with lyticase (Lyticase from *Arthrobacter luteus* L4025, Sigma-Aldrich, St. Louis, MO) were resuspended in 978 µL of the provided sodium phosphate, 122 µL MT buffer and digested with 200 U lyticase for 10 minutes at 30 °C.

Depending on treatment, the included beads (Lysing Matrix E) were added and the manufacturer's protocol continued from Step 4 (with beads) or Step 5 (without beads).

Isopropanol precipitated and silica column adsorbed samples were prepared with Omniprep™ for Fungi (G-Biosciences, St. Louis, MO) and the PowerPlant® Pro DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA), respectively, according to manufacturer instructions. PEG precipitation was completed as described by Brownlee[167] with polyethylene glycol (PEG) 8000 (Sigma-Aldrich, St. Louis, MO). CTAB – RNAse A was performed as described by Brookman and Nicholson[171]. CTAB + RNAse A used a modified protocol where cells were resuspended and lysed in a CTAB DNA isolation buffer supplemented with 0.2% β-mercaptoethanol (Sigma-Aldrich, St. Louis, MO), and 0.1 mg/ml proteinase K (New England Biolabs, Ipswich, MA) at 60 °C for 1 h. The lysate was treated with 0.8 ml phenol:chloroform:isoamyl alcohol (25:24:1) (Sigma-Aldrich, St. Louis, MO) for 2 minutes before being centrifuged. The aqueous layer was removed and then treated with 1 µL of 300 U/ml RNAse A (G Biosciences, St Louis, MO) for 30 minutes at 37 °C before being precipitated in isopropanol (Sigma-Aldrich, St. Louis, MO) and 70% ethanol (Fisher Scientific, Pittsburgh, PA)[57]. DNA was resolubilized in TE buffer. The Sarkosyl/CTAB protocol was provided by Prof. Mostafa Elshahed at Oklahoma State (personal communication). Briefly, fungal biomass was ground under liquid nitrogen before being resuspended in 10 mL TE buffer (100 mM Tris-HCl, pH 8.0; 100 mM EDTA; Sigma-Aldrich, St. Louis, MO) supplemented with 250 mM NaCl (Sigma-Aldrich, St. Louis, MO) and 0.45 mg/ml Proteinase K (G-Biosciences, St. Louis, MO). To this, 1 mL of 10% sodium lauroylsarcosine (Sarkosyl – MP Biomedicals, Santa Ana, CA) was added before overnight incubation at 50 °C with gentle agitation. The lysate was then incubated with 2 ml of 5M NaCl and 1.6 ml 10% CTAB (hexadecyltrimethylammonium bromide – Sigma Aldrich, St. Louis, MO) in 0.7M NaCl at 65 °C. The samples were then deproteinized with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1) and centrifugation for 10 minutes at 9.7×10^3 g, 4 °C. The aqueous layer was then precipitated in isopropanol and then anhydrous molecular biology grade ethanol (Fisher Scientific, Pittsburgh, PA) before being resuspended in 500 µL 10 mM Tris-HCl, pH 8.5 buffer. This prep was then treated with 0.1 mg RNAse A (G-Biosciences, St. Louis, MO) at 37 °C for 1 h before being precipitated in ethanol, washed and resolubilized in 10 mM Tris-HCl, pH 8.5 buffer (Thermo Scientific, Grand Island, NY).

3.4.5. Assessment of DNA quality, quantity, and integrity

DNA yield and quality were evaluated spectrophotmetrically using a NanoDrop 2000 (Thermo Scientific, Waltham, MA). Intact double stranded DNA was assessed using the Quant-iT™PicoGreen dsDNA Assay Kit (Invitrogen, formerly Molecular Probes Inc, Eugene, OR). DNA integrity was evaluated with gel electrophoresis using 0.7 % (w/v) agarose at 90V, 60 min[57]. DNA fragment size was quantified using a Bio-Rad CHEF-DR III System (Bio-Rad Laboratories, Hercules, CA) using a 1% SeaKem Agarose (Lonza, Basel, Switzerland) gel in 0.5X TBE (45 mM Tris-HCl, 45 mM borate, 1.0 mm EDTA, pH 8.3) with electrophoresis parameters set at 6 V/cm^2, an initial switch time of 1 s, a final switch time of 7s, included angle of 120° and run time of 17 h. DNA preps were compared against either the 1 kb ladder (NEB, Ipswich, MA) or the 5 kb Ladder (Bio-Rad Laboratories, Hercules, CA).

3.4.6. Cryopreservation stock creation and revival

Gut fungi were grown in Hungate tubes as described above, supplemented with 0.1 grams of reed canary grass (graciously provided by Paul Weimer, US Department of Agriculture) for 2 days prior to preservation. Cultures were stored in a preservation medium of sterile, anaerobic Medium C supplemented with 15% rumen fluid[70] and glycerol (Fisher Scientific, Pittsburgh, PA) at 10, 15 or 25% (v/v). After two days of growth, the culture supernatant was removed anaerobically with a syringe needle, leaving only insoluble substrate and colonizing fungal growth. Preservation medium was then added anaerobically and mixed by inversion to disperse the colonized reed canary grass. These tubes were then uncapped under a stream of 100% CO_2 gas (Praxair, Oxnard, CA) and the contents transferred into a 2-mL polypropylene cryovial (Corning Part # 430488, Corning, NY) using a wide bore pipette, also under a stream of CO_2. These vials were capped and rapidly frozen in liquid nitrogen before being stored at -80°C. To revive preserved cultures, cryovials were removed from -80°C and placed into an incubator at 39°C for 15 minutes to quickly thaw the vials and minimize ice crystal damage. Once thawed, the cryovials were opened under a stream of CO_2, the glycerol containing media was removed, and the fungal biomass and particles of reed canary grass resuspended in sterile growth medium. This material was transferred to a fresh 10 mL culture supplemented with reed canary grass and allowed to grow at 39°C. Successfully revived cultures produced fermentation gases within 2 days and displayed visible indicators of fungal growth such as the appearance of rising bubbles and the formation of a buoyant plug of plant material and fungal biomass[11].

4. Application of gut fungi for consolidated bioprocessing

4.1. Introduction

New approaches to harness lignocellulosic feedstocks for energy and chemical production are needed to grow a sustainable bio-based economy [12]. However, most fermentation processes utilize microbes that require simple sugars as feedstocks. In industry, lengthy, expensive, and often harsh pretreatments are used to separate lignin in crude biomass from carbohydrate fractions [23] that must then be hydrolyzed into fermentable sugars by large cocktails of 40-50 enzymes isolated from a variety of microbial species [25]. Combining lignocellulose hydrolysis and biocatalysis in a single bioprocess would improve the efficiency of bio-based chemical production and reduce overall costs. Common consolidated bioprocessing (CBP) approaches rely on endowing model organisms with cellulolytic activity or engineering natively cellulolytic organisms for bioproduction [5]. Similarly, the ability to compartmentalize breakdown and production steps within different microbes offers a third path forward, and capitalizes on the strengths of each microbe[58,60,175-177]. However, existing consortia-enabled technologies still require extensive pretreatment to remove lignin from biomass prior to breakdown and conversion.

The use of microbes that natively degrade crude biomass greatly reduces (or even removes) the need for these pretreatment steps. For this purpose, anaerobic gut fungi are members of a natural community found in the guts of ruminants and large monogastric herbivores that evolved to break down plant material [6-8]. They effectively degrade biomass [9] through the secretion of cellulases, hemicellulases, and other hydrolytic enzymes required for lignocellulose breakdown via the activity of extracellular fungal cellulosomes [10,11]. Gut fungi are known to form syntrophic relationships with rumen methanogens that convert the carbon dioxide and hydrogen they produce into methane [9,67] We hypothesize that these fungi liberate additional valuable nutrients during lignocellulose hydrolysis that benefit other microbes. However, due to a lack of genetic information as well as a detailed understanding of their metabolism, gut fungi have not been utilized as a method for industrial lignocellulose digestion or product conversion.

Here, we evaluated the potential of two recently classified [131] strains of anaerobic gut fungi, *Neocallimastix californiae* and *Anaeromyces robustus,* for their use in a CBP co-culture strategy with the model production yeast *Saccharomyces cerevisiae*. Through transcriptomic analysis we established the catabolic pathways of biomass derived sugars to predict the carbohydrates utilized by gut fungi and those left behind for potential microbial partners. Differential expression analysis identified how key carbohydrates regulate fungal biomass degrading enzymes and highlighted the culture conditions required to elevate their production in each fungus. Batch fermentation experiments revealed that high production of fungal enzymes led to the release and accumulation of excess sugars, enabling biphasic fermentation opportunities that harness the excess sugars to support growth of non-cellulolytic, industrially relevant organisms like *S. cerevisiae*. Overall, this work shows that gut fungi can consolidate pretreatment and hydrolysis steps, providing sugar rich hydrolysate to support growth of model microbes for bioproduction from lignocellulose.

4.2. Results

4.2.1. Gut fungi release excess sugars during hydrolysis of biomass

Given the exceptional biomass degrading capability of gut fungi and their natural presence in a competitive microbial community, we hypothesized that fungal enzymes hydrolyze more sugars from biomass than are necessary to support their own growth. In isolation, gut fungi have no competition for sugars and other resources and their extracellular cellulolytic enzymes are not subject to extensive proteolytic degradation, leading to more extensive hydrolysis and accumulation of sugars in the culture broth. To evaluate this hypothesis, the concentration of glucose was quantified in isolated cultures of *N. californiae* and *A. robustus* grown on crystalline cellulose (Figure 4.1.A). From 100 milligrams of crystalline cellulose in a 10-mL culture, *A. robustus* yielded 49.1 ± 2 milligrams of excess glucose with a maximum rate of 0.303 mg glucose/hr and *N. californiae* yielded 49.3 ± 4 milligrams with a maximum rate of 0.287 mg/hr with the bulk of glucose released after fungal growth had ceased (Figure 4.1.A). The maximum rate of glucose consumption (Figure 4.1.B), 1.470 mg/hr and 0.590 mg/hr for *A. robustus and N. californiae*, respectively, was greater than the rates of glucose release. This suggests that the fungal enzymes remained active and stable well beyond fungal death with continued hydrolysis. This excess hydrolytic capacity was highlighted when cellulose loading was increased to 200 mg in 10 mL of media and resulted in nearly doubling the amount of excess glucose released by *A. robustus*, although it had no significant effect on sugar release by *N. californiae* (Figure 4.2).

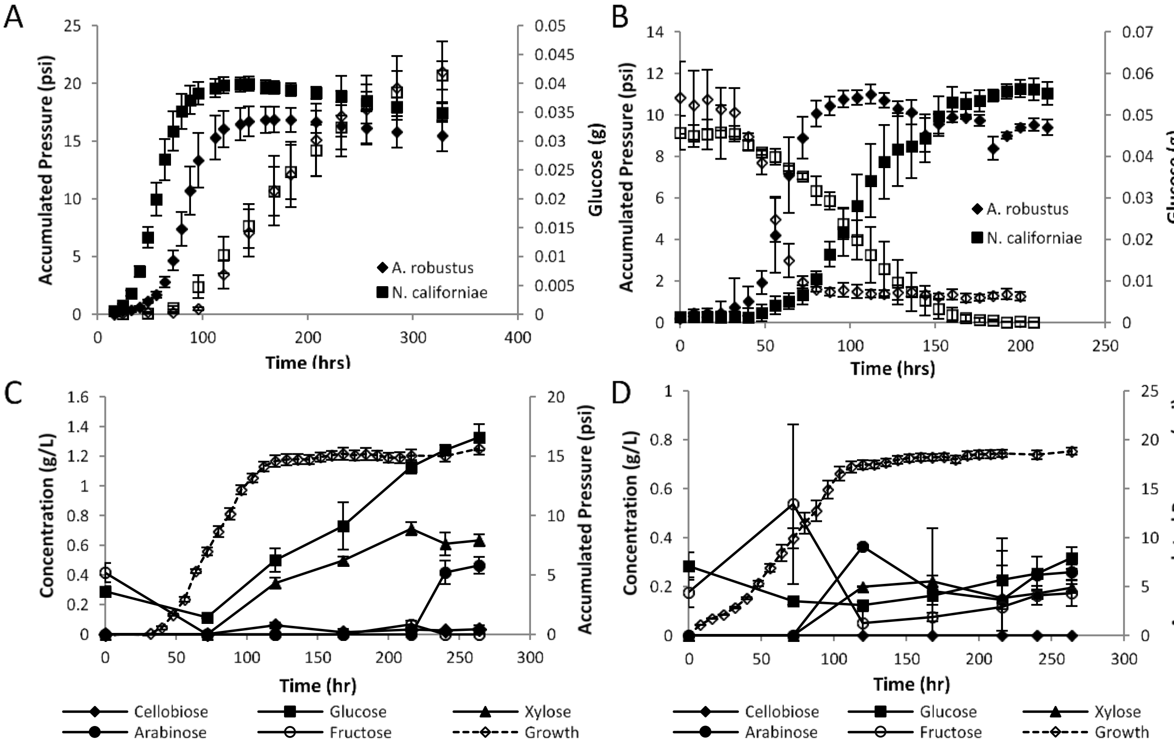


Figure 4.1. Excess sugars are released from cellulosic and lignocellulosic substrates

A) Growth of anaerobic gut fungi on crystalline cellulose. Accumulated pressure of fermentation gases (filled symbols) tracks growth and glucose concentration (empty symbols) tracks release of excess sugar from cellulose. B) Glucose consumption by *A. robustus* and *N. californiae* when grown on glucose as a sole carbon source. C) Growth of *A. robustus* on 0.5g of reed canary grass in 10 mL culture and sugar released from biomass. Growth (pressure) data is shown in empty symbols and sugar data in solid symbols. D) Growth of *N. californiae* on 0.5g of reed canary grass in 10 mL culture and sugar released from biomass. Growth (pressure) data is shown in empty symbols and sugar release data in solid symbols.

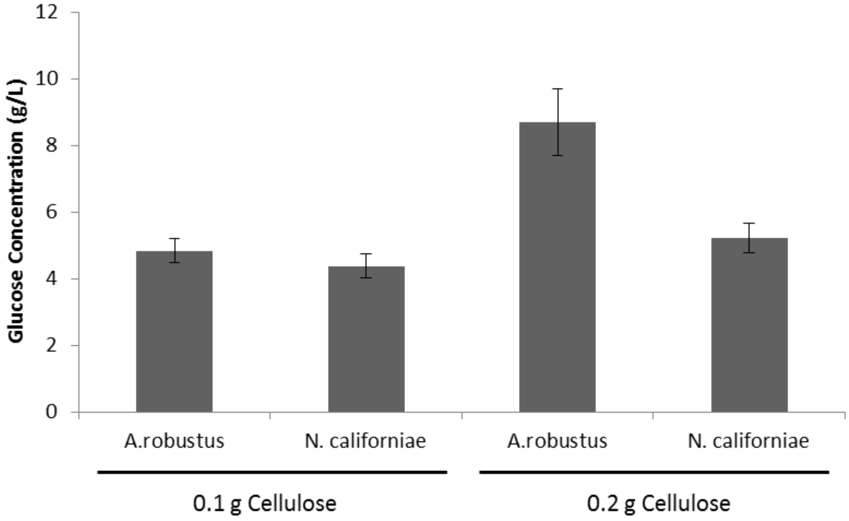


Figure 4.2. Concentration of excess glucose increases with cellulose loading

Increasing the ratio of cellulose mass to culture volume from 100 mg in 10 mL to 200 mg in 10 mL resulted in a drastic increase in the amount of excess glucose released by *A. robustus* and a small increase in the amount of free glucose released by *N. californiae*.

Subsequently, fungi were grown on reed canary grass to determine if excess sugars were available following hydrolysis of more industrially-relevant biomass substrates (Figure 4.1.C-D). When grown on 500 mg of reed canary grass, *A. robustus* yielded 16.4 ± 1.2 mg of excess glucose and *N californiae* yielded 7.1 ± 0.5 mg glucose in a 10-mL batch culture. Considering the reed canary grass cell wall composition consists of approximately 21% glucose from cellulose [178], this indicates that *A. robustus* released at least 16% of the total cellulose in the reed canary grass as excess free glucose. While this yield was significantly lower than the 49% released from pure cellulose and is likely due to the increased complexity of plant material, additional sugars derived from hemicellulose were also present in the hydrolysate. *A. robustus* also released 8.2 ± 0.7 mg xylose, 8.1 ± 0.3 mg arabinose, and 4.3 ±0.8 mg fructose, while *N. californiae* released 3.2 ± 0.3 mg xylose, 6.0 ± 1.0 mg arabinose, and 16.6

± 3.3 mg fructose. A summary of final concentrations for each carbohydrate breakout product are presented in Figure 4.7.B. *A. robustus* and *N. californiae* yielded a total accumulated sugar concentration of 4.5 ± 0.4 and 4.0 ± 0.6 g/L, respectively. Low concentrations of cellobiose were measured in the culture broth of both fungi; however, we expect that cellobiose is primarily hydrolyzed to glucose or directly taken up by the fungi due to a wealth of putative cellobiose transporters [179]. We note that a small amount of sugar was released from the reed canary grass upon autoclaving the media - these are likely soluble sugar components or easily hydrolyzed components of hemicellulose. However, these sugars were immediately consumed by the fungi (Figure 4.1C-D), with additional quantities released at later times due to high enzyme activity.

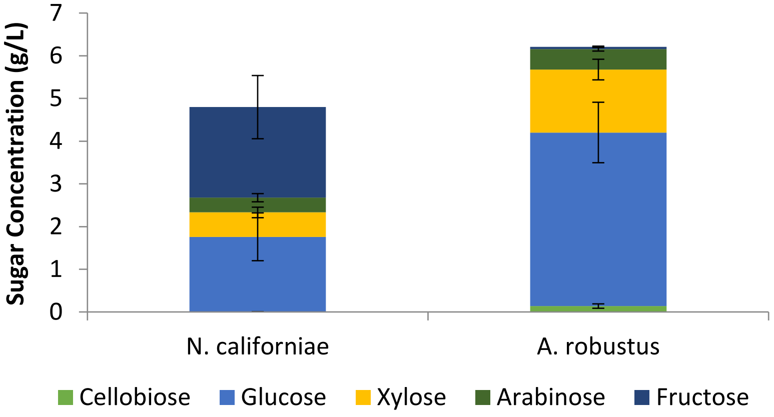


Figure 4.3. End-point sugar concentrations for cultures dosed with Hygromycin B at 72 hrs

Cultures killed by Hygromycin B during exponential growth showed an increase in the concentration of sugar released compared to wild-type cultures. This further demonstrates the capability of the fungal enzymes alone to hydrolyze biomass without the physical deconstruction by the active growth of the fungal rhizoidal network.

Similar to cultures grown in crystalline cellulose, the bulk of the excess sugar release was observed after fungal growth was depleted (Figure 4.1.C-D). Excess xylose and arabinose were expected to accumulate because the fungi did not demonstrate an ability to grow on these sugars in isolation (Table 2.1). However, glucose is likely present in large quantities because it is the most abundant sugar present in the biomass such that there is more than enough to support fungal growth. Additional fungal cultures grown on 500 mg of reed canary grass were killed with the antimicrobial hygromycin B during exponential growth at 72 hours post-inoculation to evaluate the capability of fungal enzymes alone to hydrolyze biomass. These cultures yielded greater amounts of overall sugars, with the largest increases in the amounts of glucose released (Figure 4.3). Sugar yields in *A. robustus* killed with hygromycin were 32.5 ± 5.7 mg glucose, 11.8 ±1.9 mg xylose, 3.9 ± 0.4 mg arabinose, and 0.4 mg ± 0.1 mg fructose. *N. californiae* yielded 14.1 ± 4.5 mg glucose, 4.5 ± 1.0 mg xylose, 2.8 ± 0.8 mg arabinose, and 17.0 ± 5.9 mg fructose. These hygromycin dosed cultures yielded increased amounts of total free sugars in the culture broth, when the enzymes could act on biomass in the absence of fungal growth and consumption of sugar. These results highlighted the capability of gut fungal enzymes alone to hydrolyze biomass and present gut fungi as a source for improved enzyme cocktails to hydrolyze crude lignocellulose. Because sugars generally do not begin to accumulate until after fungal growth has ceased (Figure 4.1.A,C,D) the most feasible application of a co-culture system is a two-stage approach, whereby biomass or cellulose is first incubated with gut fungi to produce excess sugar that can then be fed to a second, model organism for direct production of a value-added product

4.2.2. Biomass degrading enzymes are regulated by substrate availability

Anaerobic gut fungi possess a large and diverse suite of biomass degrading enzymes[10,79,130,180] that allow them to easily break down complex plant polysaccharides. However, very few studies have explored how these genes are regulated in response to changing environmental conditions, such as addition of a catabolite repressor[130] or general substrate availability[181]. We sought to understand the conditions that optimized biomass degrading enzyme production in *N. californiae* and *A. robustus* for their potential application to consolidated bioprocessing. Based on their varied growth and metabolic capabilities, we hypothesized that different gut fungi rely on species-specific mechanisms to regulate their biomass degrading enzymes in response to substrate availability. Due to the importance of CAZymes to consolidated bioprocessing our analysis focused on the regulation of these genes in both strains of gut fungi using RNA-Seq to evaluate their expression during growth on different substrates.

Table 4.1. Breakdown of carbohydrate active enzymes

	CAZyme	**# Transcripts**	
		N. californiae	***A. robustus***
Cellulase	GH1	16	11
	GH3	34	16
	GH5	48	22
	GH6	22	6
	GH8	4	1
	GH9	25	15
	GH16	15	9
	GH31	7	6
	GH45	24	13
	GH48F	24	7
	Total	**219**	**106**
Hemicellulase	GH10	67	16
	GH11	67	30
	GH11-12	67	30
	GH30	2	2
	GH39	9	4
	GH43	37	19
	Total	**249**	**101**
Accessory Enzyme	Polysaccharide deacetylase	93	58
	Carbohydrate Esterase	43	28
	Pectinesterase	12	5
	Pectate lyase	35	5
	Rhamnogalcturonate lyase	4	3
	GH88	2	0
	Total	**189**	**99**
Binding	CBM10	534	271

Table 4.2. Summary of up- and down-regulated CAZyme transcripts under different growth conditions compared to growth on glucose

	A. robustus		***N. californiae***	
Growth Condition	**Down Regulated**	**Up Regulated**	**Down Regulated**	**Up Regulated**
Maltose	0	3	0	10
Cellobiose	9	84	36	87
Avicel	4	86	122	124
Corn Stover	11	97	36	168
Reed Canary Grass	19	122	65	177
Switchgrass	34	108	46	168

Overall, the transcriptome of *N. californiae* contained more than twice as many carbohydrate active enzyme (CAZy) domain containing transcripts compared to *A. robustus* (657 compared to 306 CAZymes), an observation that aligns with the sizes of the genomes for each of these fungi[82]. However, the relative functional distribution of these CAZymes is conserved across both species with cellulases, hemicellulases, and accessory enzymes each comprising roughly one third of all CAZymes (Table 4.1). This conserved balance of functional activities suggests that each function is required in equal proportion to efficiently degrade biomass. We analyzed transcript abundance with RSEM[123] to obtain expression counts for all transcripts during growth on glucose, maltose, cellobiose, cellulose, corn stover, reed canary grass, and switchgrass. RNA from cultures grown on each substrate was isolated in triplicate samples that were sequenced with greater than 50X coverage (Table 4.3 and Table 4.4).

Table 4.3. *Anaeromyces robustus* substrate regulation sequencing summary and RNA quantity and quality data.

Substrate	Sample	# Clusters	# Reads	Read Length	Coverage	RINe	Qubit Conc. (ng/uL)
Glucose	1	18,257,400	36,514,800	75	124.73	9.5	190
	2	13,869,993	27,739,986	75	94.757	9	122
	3	17,627,885	35,255,770	75	120.43	9.6	200
Maltose	1	14,420,112	28,840,224	75	98.516	8.9	232
	2	16,589,476	33,178,952	75	113.33	8.5	200
	3	15,962,547	31,925,094	75	109.05	9.1	84.8
Cellobiose	1	20,725,704	41,451,408	75	141.59	9.7	112
	2	15,650,596	31,301,192	75	106.92	8.8	62
	3	17,689,401	35,378,802	75	120.85	9.7	198
Avicel	1	29,722,731	59,445,462	75	203.06	9.6	112
	2	18,871,814	37,743,628	75	128.93	9.9	128
	3	19,271,995	38,543,990	75	131.66	9.9	138
Corn Stover	1	21,722,361	43,444,722	75	148.40	9.8	57.6
	2	21,958,993	43,917,986	75	150.02	9.4	18.4
	3*	N/A	N/A	N/A	N/A	9.6	38.8
Reed Canary Grass	1	17,299,145	34,598,290	75	118.19	9.4	34.4
	2	18,759,427	37,518,854	75	128.16	9.6	43
	3	17,914,503	35,829,006	75	122.39	7.7	9.48
Switchgrass	1	16,840,549	33,681,098	75	115.05	9.5	27.2
	2	16,440,220	32,880,440	75	112.32	8.6	34
	3	37,893,909	75,787,818	75	258.89	8.9	14.1
Total		387,488,761	774,977,522		2647.27		

*Did not hydridize to flow cell properly

Table 4.4. *Neocallimastix californiae* substrate regulation sequencing summary and RNA quantity and quality data

Substrate	**Sample**	**# Clusters**	**# Reads**	**Read Length**	**Coverage**	**RINe**	**Qubit Conc. (ng/uL)**
Glucose	1	15,425,937	30,851,874	75	63.829	9.8	166
	2	50,215,392	100,430,784	75	207.78	10	114
	3	21,825,155	43,650,310	75	90.308	9.8	124
Maltose	1	16,490,554	32,981,108	75	68.234	9.9	134
	2	14,655,979	29,311,958	75	60.643	9.7	156
	3	27,461,449	54,922,898	75	113.63	9.9	79.2
Cellobiose	1	24,211,289	48,422,578	75	100.18	9.6	98.4
	2	23,233,084	46,466,168	75	96.134	9.6	89.2
	3	41,734,749	83,469,498	75	172.69	9.3	200
Avicel	1	23,806,234	47,612,468	75	98.506	9.6	172
	2	22,377,016	44,754,032	75	92.592	9.7	134
	3	30,770,853	61,541,706	75	127.32	9.8	99.8
Corn Stover	1	39,890,138	79,780,276	75	165.06	9.9	53.2
	2	19,060,236	38,120,472	75	78.868	9.8	97.2
	3	14,100,732	28,201,464	75	58.346	9.7	65.2
Reed Canary Grass	1	13,769,947	27,539,894	75	56.978	9.4	47.4
	2	20,948,842	41,897,684	75	86.683	9.5	35.8
	3	20,341,423	40,682,846	75	84.169	9.5	75.8
Switchgrass	1*	N/A	N/A	N/A	N/A	9.6	47
	2	21,691,394	43,382,788	75	89.755	9.5	74.8
	3	15,045,190	30,090,380	75	62.254	9.6	35
Alfalfa Stems	1	22,452,251	44,904,502	75	92.903	8.8	53.6
	2	12,266,278	24,532,556	75	50.756	8.9	28.2
	3	14,923,672	29,847,344	75	61.751	9.1	26
Total		439,231,310	878,462,620	75	3000.77		

*Did not hybridize to the flow cell properly

Differential expression analysis identified a total of 350 unique CAZymes in *N. californiae* (53% of all CAZymes) and 202 (66%) in *A. robustus* that were significantly regulated (greater than 2-fold change, $p \leq 0.01$) in response to growth on differing substrates compared to glucose. These transcripts were primarily upregulated as substrate complexity increases, though there was some downregulation observed (Figure 4.4). Down regulation was likely the result of transitioning to more effective CAZymes required to break down more complex substrates. Growth on cellobiose, cellulose, and plant biomass triggered large changes in expression of CAZymes, with primarily upregulation of transcripts (Table 4.2). Only growth of *N. californiae* on Avicel resulted in the downregulation of many CAZyme transcripts, nearly equal to the number upregulated under that condition. There are also many regulated transcripts that contain fungal dockerin (CBM10) domains without any other assigned CAZy functionality; 230 in *N. californiae* and 137 in *A. robustus*. Fungal biomass degrading enzymes are predicted to form multienzyme complexes facilitated by the interaction of the fungal dockerin domains and a cohesin domain present on a large, non-catalytic scaffoldin protein[82]. While these transcripts cannot be designated as CAZymes, they may play an unknown role in biomass degradation, representing unclassified carbohydrate active enzymes, or alternate functions involved in improving lignocellulolytic activity of fungal cellulosome complexes.

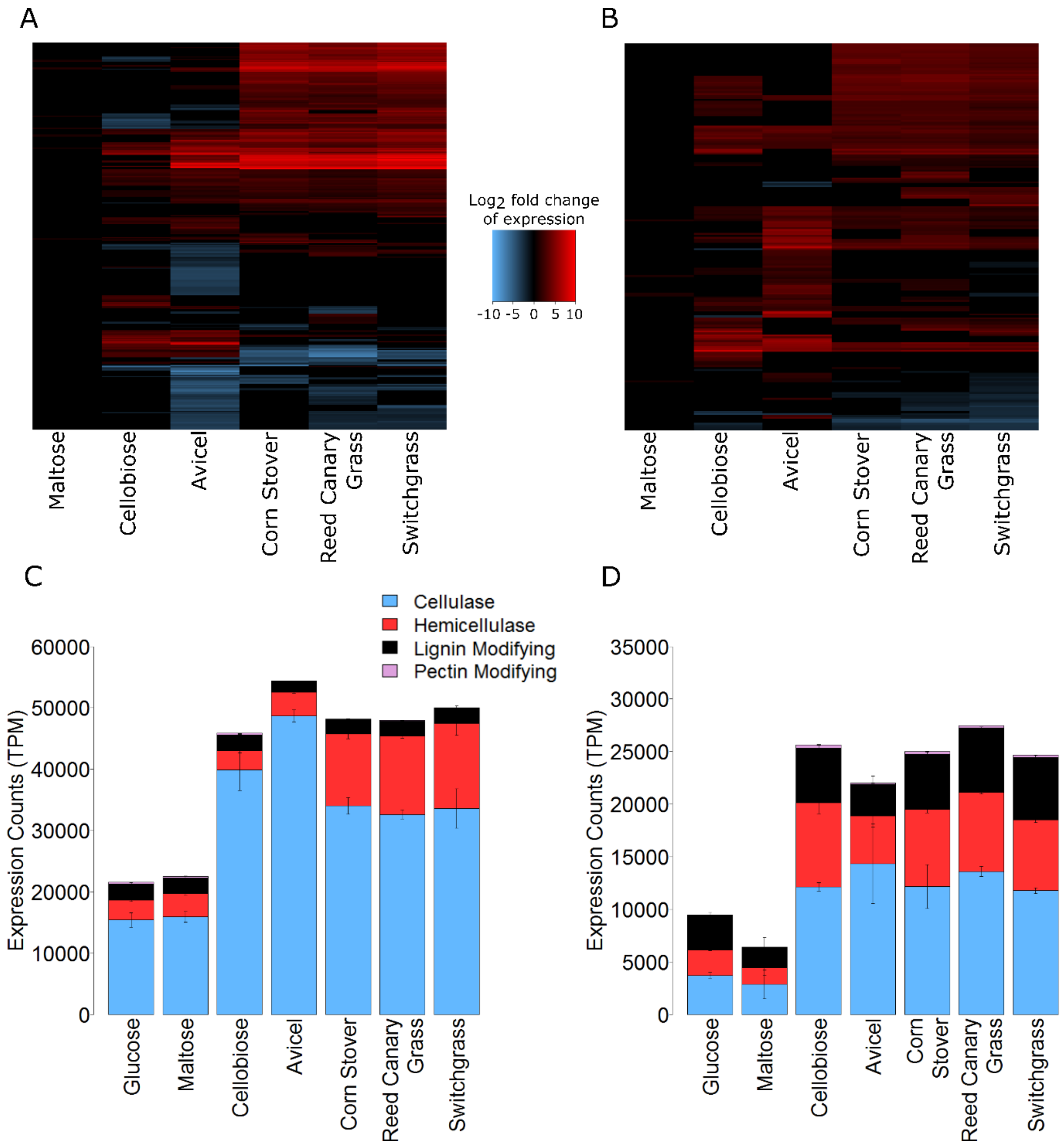


Figure 4.4. Biomass degrading enzymes of anaerobic fungi are tuned to substrate availability

A and B: Heat maps of the $\log_2$ fold change in expression of biomass degrading enzymes on a variety of substrates compared to expression on glucose for *N. californiae* and *A. robustus*, respectively. These genes are primarily upregulated on more complex substrates. C and D: Normalized expression counts in transcripts per million (TPM) of biomass degrading enzymes under all growth conditions for *N. californiae* and *A. robustus*, respectively. There is a basal level of expression on glucose, but higher expression levels are triggered by more complex substrates. In *A. robustus* cellobiose triggers increased expression of all biomass degrading enzyme types, but in *N. californiae* only cellulases demonstrated increased expression on cellobiose, while the expression of hemicellulases only increased on biomass substrates.

We further hypothesized that the overall expression of cellulases, hemicellulases, and accessory enzymes would increase only when their activity was necessary to degrade a given substrate. For example, hemicellulases would only be expressed when hemicellulose was present and available to the fungus. This was the case for *N. californiae* with a drastic increase in expression of cellulases on cellobiose and Avicel, but no change in expression of hemicellulases until hemicellulose was present in biomass substrates. Overall hemicellulase expression was increased almost 3-fold on reed canary grass as compared to Avicel (Figure 4.4.C). This suggests separate mechanisms that rely on different breakout product trigger molecules to control the expression of cellulases and hemicellulases in *N. californiae*. Alternatively, growth on cellobiose and cellulose, as well as biomass, triggered increased expression of cellulases, hemicellulases, and accessory enzymes in *A. robustus* (Figure 4.4.D). This suggests that *A. robustus* utilizes a single activator to regulate all biomass degrading enzymes, a pattern very different than that observed in *N. californiae*.

It is important to note that both organisms demonstrated a significant basal expression level of biomass degrading enzymes on glucose, approximately 21,500 and 10,500 TPM (2.15% and 1.05% of total transcriptome expression) in *N. californiae* and *A. robustus*, respectively. This basal activity likely releases break out carbohydrates from lignocellulose, such as cellobiose, that can then promote increased expression of enzymes required to hydrolyze plant material. In fact, overall expression of CAZymes in both *N. californiae* and *A. robustus* increased most drastically (by greater than 200%) when grown on cellobiose, a low molecular weight cellodextrin, compared to glucose (Figure 4.4.C-D). This effect revealed that growth of *A. robustus* on cellobiose will induce production of the entire suite of enzymes required to break down crude biomass. Considering that many of these enzymes contain carbohydrate binding domains that keep them tightly bound to lignocellulose, this would allow for simpler purification of enzymes that does not require the separation of enzymes from the substrates they act on. Conversely, *N. californiae* requires growth on complex biomass to produce all necessary enzymes, making enzyme purification more difficult. For the isolation of enzymes for a lignocellulose hydrolysis cocktail, *A. robustus* presents the best path forward.

Insight into the regulatory mechanisms of gut fungi can be used to optimize enzyme production and achieve maximum lignocellulolytic activity and sugar handoff to model microorganisms. For example, identifying candidates for knockout can lead to increased expression of lignocellulolytic enzymes and enhancement of biomass breakdown. Possible regulators of biomass degrading enzymes in these gut fungi were previously identified by Solomon et al.[130] by searching for transcripts orthologous to conserved transcription factors, Cre1/CreABC, ACE1-2, ClbR, Clr1-2, and Xyr-1/XlnR that regulate hemicellulase and cellulase production in *Trichoderma reesei*, *Neurospora crassa*, and *Aspergillus niger*[182]. Solomon et al. identified orthologs to the CreABC regulator family from *A. niger* in both *A. robustus* and *N. californiae*, specifically *creB* and *creC*. With the growing amounts of sequencing data that are regularly updated to bioinformatics databases, we now have sufficient evidence to confidently identify orthologs to *creA* as well as the Cre-1 regulators from both *T. reesei* and *N. crassa* (Table 4.5 and Table 4.6). Though these sequence alignments were not as strong, alignment against the OrthoMCL database[156] resulted in placement of the transcripts in the same ortholog group as the aerobic fungal regulators (Table 4.5 and Table 4.6).

Table 4.5. Comparison of fungal lignocellulolytic regulators from *Trichoderma reesei*, *Neurospora crassa*, and *Aspergillus niger* to *A. robustus* transcripts

Regulator Gene	Response	Query Accession Number	Organism	Best BLAST hit	Bit Score	E_{value}	Similarity (Coverage)	Orthologous to query?
cre-1	Repress ligno-cellulolytic enzymes, XlnR, Ace2 on glucose	589100213	*T. reesei*	Locus12200v2rpkm0.63	75.09	9.07E-16	38% (23%)	Y
cre-1		67476474	*N. crassa*	Locus12200v1rpkm0.64	78.18	2.43E-16	36% (25%)	Y
creA		544095	*A. niger*	Locus12200v1rpkm0.64	71.63	3.67E-14	52% (13%)	Y
creB		317025538	*A. niger*	Locus5673v1rpkm4.49	300.8	1.43E-90	46% (43%)	Y
creC		300680900	*A. niger*	Locus5906v1rpkm4.09	278.9	6.42E-85	36% (70%)	Y
ace1	Represses cellulases	32699313	*T. reesei*	Locus5676v1rpkm4.49	44.28	1.90E-04	30% (15%)	N
ace2	Induces cellulases	340518224	*T. reesei*	Locus7291v1rpkm2.52	41.97	2.93E-04	42% (13%)	N
clbr2	Induces ligno-cellulytic enzymes in response to cellulose/ cellobiose	399769775	*A. aculeatus*	Locus8550v1rpkm1.68	48.14	9.41E-06	31% (14%)	N
xyr-1	Induce hemi-cellulases in presence of xylan	340517797	*T. reesei*	Locus8550v1rpkm1.68	49.29	5.87E-06	35% (6%)	N
xlnR		85108643	*N. crassa*	Locus7291v1rpkm2.52	50.45	2.91E-06	27% (13%)	N
xlnR		292495047	*A. niger*	Locus7291v1rpkm2.52	51.99	9.25E-07	43% (5%)	N
clr-1	Induces ligno-cellulolytic enzymes on cellobiose	553136585	*N. crassa*	Locus8645v1rpkm1.64	53.14	2.75E-08	35% (10%)	N
clr-2		553136900	*N. crassa*	Locus8550v1rpkm1.68	50.44	1.86E-06	30% (20%)	N

The above findings suggest that the gut fungi possess a similar genetic response system for glucose-based regulation, indicating an early evolutionary origin of the CreABC regulatory network. However, only differential expression results for *A. robustus* are consistent with a lack of hemicellulase specific regulators, Xyr-1/XlnR. The results for *N. californiae* suggest a similar hemicellulase regulatory system despite missing orthologs, which may indicate parallel evolution of this function in gut fungi. Glucose concentrations as small as 0.5 g/L (0.05% w/v) can trigger carbon catabolite repression in gut fungi[130]. The CreABC regulators are likely candidates for the source of this regulation and knocking them out may alleviate catabolite repression of CAZymes as sugars accumulate during active growth of gut fungi.

Table 4.6. Comparison of fungal lignocellulolytic regulators from *Trichoderma reesei*, *Neurospora crassa*, and *Aspergillus niger* to *N. californiae* transcripts

Regulator Gene	**Response**	**Query Accession Number**	**Organism**	**Best BLAST hit**	**Bit Score**	**E_{value}**	**Similarity (Coverage)**	**Orthologous to query?**
cre-1	Repress ligno-cellulolytic enzymes, XlnR, Ace2 on glucose	589100213	*T. reesei*	Locus22410v1rpkm0.33	75.49	3.86E-15	49% (14%)	Y
cre-1		67476474	*N. crassa*	Locus22410v1rpkm0.33	75.49	4.41E-15	49% (13%)	Y
creA		544095	*A. niger*	Locus22410v1rpkm0.33	6.26	2.21E-15	51% (14%)	Y
creB		317025538	*A. niger*	Locus6300v1rpkm7.52	301.9	2.75E-90	48% (43%)	Y
creC		300680900	*A. niger*	Locus4513v1rpkm13.19	304.3	7.32E-93	39% (75%)	Y
ace1	Represses cellulases	32699313	*T. reesei*	Locus9020v1rpkm3.85	46.21	7.16E-05	34% (13%)	N
ace2	Induces cellulases	340518224	*T. reesei*	Locus15611v1rpkm1.01	41.59	6.68E-04	40% (13%)	N
clbr2	Induces ligno-cellulytic enzymes in response to cellulose/ cellobiose	399769775	*A. aculeatus*	Locus15611v1rpkm1.01	51.22	2.43E-06	41% (7%)	N
xyr-1	Induce hemi-cellulases in presence of xylan	340517797	*T. reesei*	Locus15611v1rpkm1.01	55.45	1.57E-07	48% (5%)	N
xlnR		85108643	*N. crassa*	Locus15611v1rpkm1.01	58.92	1.32E-08	28% (13%)	N
xlnR		292495047	*A. niger*	Locus15611v1rpkm1.01	61.23	2.56E-09	28% (15%)	N
clr-1	Induces ligno-cellulolytic enzymes on cellobiose	553136585	*N. crassa*	Locus11145v1rpkm2.40	51.99	1.08E-06	45% (5%)	N
clr-2		553136900	*N. crassa*	Locus15611v1rpkm1.01	51.22	2.79E-06	29% (10%)	N

4.2.3. Metabolic maps reveal opportunities for consolidated bioprocessing

Anaerobic gut fungi are capable of releasing sugars from both cellulose and hemicellulose due to the wide array of CAZymes that they possess (Figure 4.5.A), yet batch growth experiments (Table 2.1) revealed that they did not metabolize some of these sugars in monoculture. Metabolic maps were built from the transcriptomes to highlight gaps in sugar catabolism pathways that may provide opportunities for microbial co-culturing via sugar exchange. Enzyme Commission (EC) numbers were assigned to transcripts during transcriptome annotation and were used to generate metabolic maps based on the KEGG databases [183,184]. We sought to identify sugars that each of the isolated strains were capable of metabolizing based on the enzymes and metabolic routes they possess (Figure 4.5.B).

Analysis of glycolysis (KEGG path 00010) revealed a complete catabolic pathway for glucose (Figure 4.5.B). While all the glycolytic enzymes were identified, two enzymes necessary for complete gluconeogenesis, fructose bisphosphatase (EC:3.1.3.11) and glucose-6-phosphatase (EC:3.1.3.9), were missing via EC annotation, though fructose bisphosphatase was identified by BLAST annotation (>70% similarity). This corroborates previous observations from other gut fungal genera suggesting that gluconeogenesis is incomplete in gut fungi[79]. Our analysis of xylose metabolism revealed the xylose isomerase pathway typical of prokaryotes[185] in both *N. californiae* and *A. robustus*; an observation consistent with previous findings for the gut fungus *Piromyces* sp. E2[132]. The xylose isomerase pathway may have arisen from horizontal gene transfer in the rumen microbiome and lead to increased fitness over the eukaryotic oxido-reductase pathway that would suffer from poor activity under anaerobic conditions. The oxido-reductase pathway requires the oxidation of NAD(P)H to $NAD(P)^+$ to convert xylose to xylitol and the reduction of NAD^+ to NADH to convert xylitol to D-xylulose[185]. The anaerobic, reducing environment of the gut is likely to upset the redox balance of this pathway reducing its effectiveness and resulting in accumulation of xylitol, while the xylose isomerase is less affected by anaerobic conditions[186]. Though metabolic maps indicate that the fungi are capable of xylose catabolism, growth experiments revealed that they do not thrive on the pentose sugar in isolated culture (Table 2.1). This discrepancy between transcriptomic and growth experiment observation suggests that another limitation is responsible for lack of xylose utilization in these gut fungi. This may be the result of inefficient transport of xylose into the cell, or other influences from the native environment that were not present under laboratory conditions. Xylose catabolism may be triggered by other plant components or even organisms in the environment that were not present in these experiments.

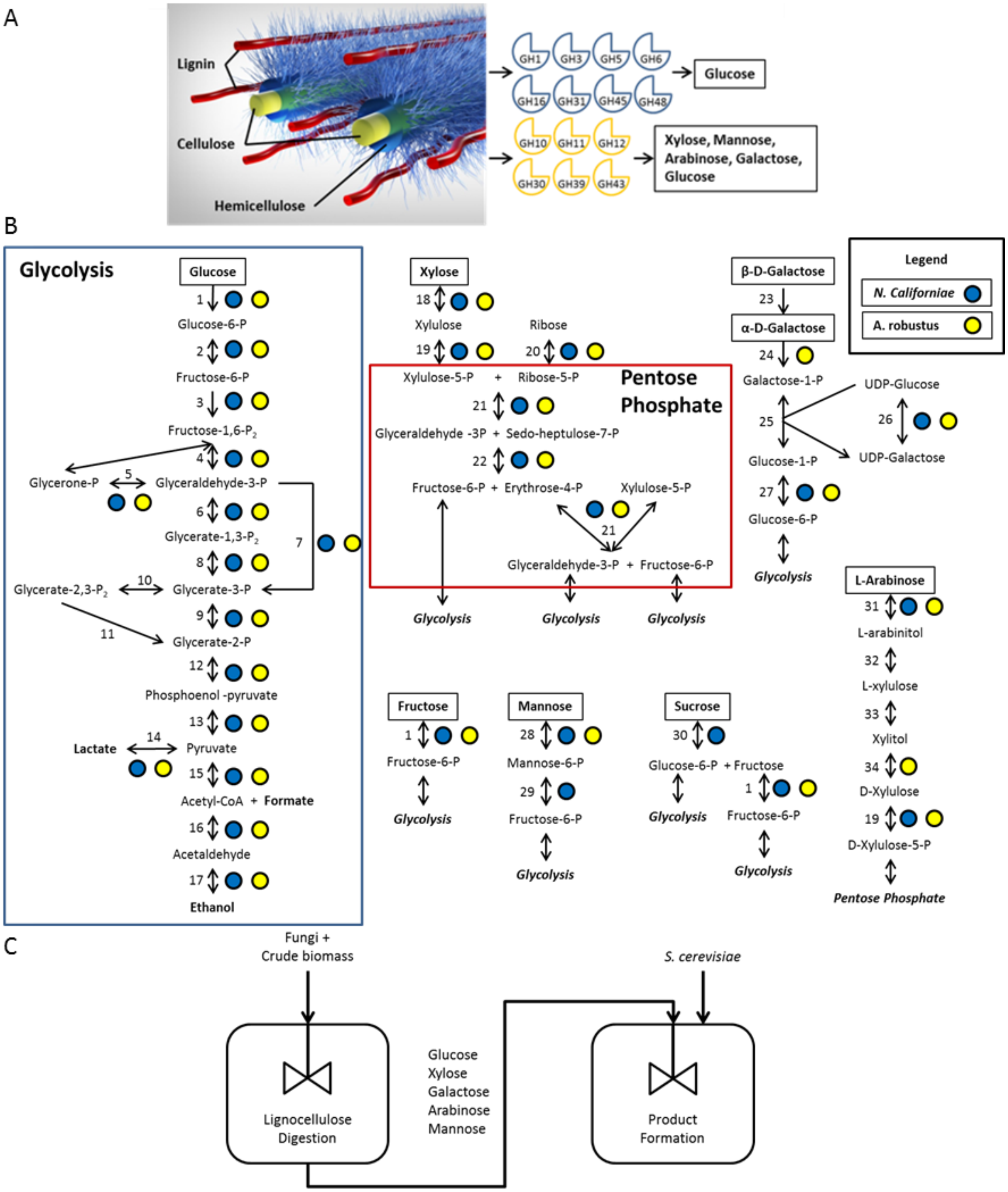


Figure 4.5. Metabolic reconstruction of sugar catabolic pathways in gut fungi

A) Cellulases and hemicellulases release sugar-rich hydrolysates from lignocellulose. B) Enzymatic steps in the pathway are identified as present in each of the fungi. Dots indicate enzymes identified in the transcriptomes of *N. californiae* (blue) and *A. robustus* (yellow) – details in Table 4.7. C) Two-stage culture system where fungi are used to release sugar from biomass that can be fed to a production organism, such as *S. cerevisiae*, in a second step.

Table 4.7. Sugar catabolism pathway enzyme summary for Figure 4.5. Information obtained from BRENDA[187].

Reaction #	E.C. Number	Enzyme Description	Reaction Description (BRENDA)
1	2.7.1.1	Hexokinase	ATP + D-hexose = ADP + D-hexose 6-phosphate
2	5.3.1.9	Glucose-6-phosphate isomerase	D-Glucose 6-phosphate = D-fructose 6-phosphate
3	2.7.1.11	6-phosphofructokinase	ATP + D-fructose 6-phosphate = ADP + D-fructose 1,6-bisphosphate
4	4.1.2.13	fructose-bisphosphate aldolase	D-fructose 1,6-bisphosphate = glycerone phosphate + D-glyceraldehyde 3-phosphate
5	5.3.1.1	Triose-phosphate isomerase	D-Glyceraldehyde 3-phosphate = glycerone phosphate
6	1.2.1.12	glyceraldehyde-3-phosphate dehydrogenase	D-glyceraldehyde 3-phosphate + phosphate + NAD+ = 3-phospho-D-glyceroyl phosphate + NADH + H+
7	1.2.1.9	glyceraldehyde-3-phosphate dehydrogenase (NADP+)	D-glyceraldehyde 3-phosphate + NADP+ + H2O = 3-phospho-D-glycerate + NADPH + 2 H+
8	2.7.2.3	phosphoglycerate kinase	ATP + 3-phospho-D-glycerate = ADP + 3-phospho-D-glyceroyl phosphate
9	5.4.2.12	phosphoglycerate mutase (2,3-diphosphoglycerate-independent)	2-phospho-D-glycerate = 3-phospho-D-glycerate
10	3.1.3.13	bisphosphoglycerate phosphatase	2,3-bisphospho-D-glycerate + H2O = 3-phospho-D-glycerate + phosphate
11	3.1.3.80	2,3-bisphosphoglycerate 3-phosphatase	2,3-bisphospho-D-glycerate + H2O = 2-phospho-D-glycerate + phosphate
12	4.2.1.11	phosphopyruvate hydratase	2-phospho-D-glycerate = phosphoenolpyruvate + H2O
13	2.7.1.40	pyruvate kinase	ATP + pyruvate = ADP + phosphoenolpyruvate
14	1.1.1.28	D-lactate dehydrogenase	(R)-lactate + NAD+ = pyruvate + NADH + H+
15	2.3.1.54	formate C-acetyltransferase	acetyl-CoA + formate = CoA + pyruvate
16	1.2.1.10	acetaldehyde dehydrogenase (acetylating)	acetaldehyde + CoA + NAD+ = acetyl-CoA + NADH + H+

17	1.1.1.1	alcohol dehydrogenase	a primary alcohol + NAD+ = an aldehyde + NADH + H+
18	5.3.1.5	Xylose isomerase	D-xylopyranose = D-xylulose
19	2.7.1.17	xylulokinase	ATP + D-xylulose = ADP + D-xylulose 5-phosphate
20	2.7.1.15	ribokinase	ATP + D-ribose = ADP + D-ribose 5-phosphate
21	2.2.1.1	transketolase	sedoheptulose 7-phosphate + D-glyceraldehyde 3-phosphate = D-ribose 5-phosphate + D-xylulose 5-phosphate
22	2.2.1.2	transaldolase	sedoheptulose 7-phosphate + D-glyceraldehyde 3-phosphate = D-erythrose 4-phosphate + D-fructose 6-phosphate
23	5.1.3.3	Aldose 1-epimerase	alpha-D-Glucose = beta-D-glucose
24	2.7.1.6	galactokinase	ATP + alpha-D-galactose = ADP + alpha-D-galactose 1-phosphate
25	2.7.7.12	UDP-glucose-hexose-1-phosphate uridylyltransferase	UDP-alpha-D-glucose + alpha-D-galactose 1-phosphate = alpha-D-glucose 1-phosphate + UDP-alpha-D-galactose
26	5.1.3.2	UDP-glucose 4-epimerase	UDP-alpha-D-glucose = UDP-alpha-D-galactose
27	5.4.2.2	phosphoglucomutase (alpha-D-glucose-1,6-bisphosphate-dependent)	alpha-D-glucose 1-phosphate = D-glucose 6-phosphate
28	2.7.1.7	mannokinase	ATP + D-mannose = ADP + D-mannose 6-phosphate
29	5.3.1.8	Mannose-6-phosphate isomerase	D-Mannose 6-phosphate = D-fructose 6-phosphate
30	3.2.1.20	alpha-glucosidase	Hydrolysis of terminal, non-reducing (1->4)-linked alpha-D-glucose residues with release of D-glucose
31	1.1.1.21	aldehyde reductase	alditol + NAD(P)+ = aldose + NAD(P)H + H+
32	1.1.1.12	L-arabinitol 4-dehydrogenase	L-arabinitol + NAD+ = L-xylulose + NADH + H+
33	1.1.1.10	L-xylulose reductase	xylitol + NADP+ = L-xylulose + NADPH + H+
34	1.1.1.9	D-xylulose reductase	xylitol + NAD+ = D-xylulose + NADH + H+

Characterization of fructose and mannose metabolism (KEGG path 00051) as well as starch and sucrose metabolism (KEGG path 00500) identified that both fungi metabolize fructose. However, only *N. californiae* metabolizes mannose and sucrose, and both fungi lack enzymes required for galactose and arabinose metabolism. These predictions are corroborated by results from growth experiments (Table 2.1) and identify these sugars as candidates for hand off to another organism without interference from fungal growth.

Table 4.8. Transcriptome alignment to NuoF and NuoE proteins from *Trichomonas vaginalis*

	***T. vaginalis* sequence**	**Transcript ID**	**Bit Score**	**E-value**
Anaeromyces robustus	**NuoE**	Locus433v1rpkm290.85	64.7583	4.61238e-30
	NuoF	Locus297v1rpkm532.54	336.934	1.82717e-155
Neocallimastix californiae	**NuoE**	Locus485v1rpkm376.26	65.2165	4.04388e-30
		Locus449v1rpkm432.50	65.2165	4.8221e-29
	NuoF	Locus415v1rpkm484.94	339.683	3.61833e-157
		Locus850v1rpkm157.37	336.476	4.95987e-157

Downstream, the enzymes required for ethanol production from pyruvate were identified in both organisms (Figure 4.5.B), yielding formate as a side product. Energy generation in anaerobic gut fungi also relies on the hydrogenosome organelle[188] that is also found in members of the *Trichomonas* genus and several other anaerobic protists[189]. This organelle performs a similar function to the mitochondria commonly found in eukaryotes, but generates energy in the absence of oxygen by substrate level phosphorylation[190,191]. In the hydrogenosome, we identified malate dehydrogenase that produces pyruvate from the oxidative decarboxylation of malate, derived from phosphoenolpyruvate to produce 1 ATP and recycle a molecule of NADH. We also identified soluble components of mitochondrial complex I, NADH:ubiquinone oxidoreducatase (EC:1.6.5.3), which were also strong homologs (E ≤ 10^{-150}; Table S8) of the *Trichomonas vaginalis* enzymes NuoF and NuoE (Table 4.8) that regenerate NAD^+ for this step by transferring the electrons to ferredoxin [192]. Pyruvate is then converted to acetyl-CoA and formate using ferredoxin as an electron acceptor and coenzyme A is then transferred from acetyl-CoA to a succinate molecule to form succinyl-CoA and acetate. Succinate and CoA are regenerated in a step that generates a molecule of ATP. Reduced ferredoxin is oxidized by a hydrogenase (EC:1.12.7.2) in a reaction that also yields molecular hydrogen (Figure 4.6).

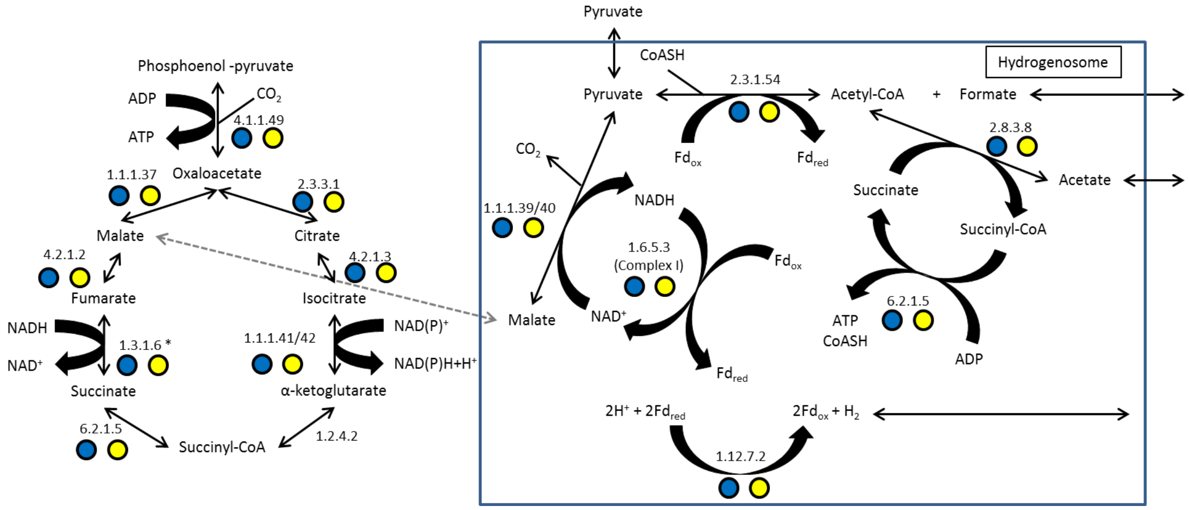


Figure 4.6. Energy generation through the anaerobic fungal hydrogenosome.

In both *N californiae* and *A. robustus* only one enzyme, oxoglutarate dehydrogenase, was not identified in the tricarboxylic acid (TCA) cycle. In the hydrogenosome, malate or pyruvate are transported into the organelle and are converted to formate and acetate. ATP is generated by substrate level phosphorylation by succinyl-CoA synthetase.

While the most abundant sugars remaining after biomass hydrolysis were glucose and fructose, both metabolized by the gut fungi, other sugars such as xylose and arabinose also accumulated after fungal growth (Figure 4.1.C-D). This accumulation is likely due to an inability to metabolize these sugars due to either a lack of enzymes (Figure 4.5.B) or another limitation (e.g. transport). Glucose and fructose can readily be used to support the growth of many other organisms, such as *S. cerevisiae*. As these sugars primarily accumulated in the culture broth after fungal growth was completed, we tested a two-stage production system where fungi digest biomass in the first step and the hydrolysate supports the growth of *S. cerevisiae* in a second bioreactor (Figure 4.5.C).

4.2.4. Two-step co-culture reveals potential for gut fungi in bio-based production

Following growth of fungi, the sugar-rich "spent" fungal media was inoculated with *Saccharomyces cerevisiae* (Figure 4.7.A) to determine if the fungal hydrolysate was capable of supporting yeast proliferation. The spent media containing 6-7 g/L of glucose released by gut fungi from crystalline cellulose supported growth of *S. cerevisiae* to saturation, with an OD_{600} of 14 while fresh media containing no fungal hydrolysate grew to a negligible OD_{600} (Figure 4.7.A). This not only demonstrates that the fungi were capable of hydrolyzing enough excess sugar to support growth of *S. cerevisiae*, but also that they did not produce any compounds that inhibited yeast growth. *Escherichia coli* was also tested on media from fungal cultures on cellulose resulting in a small increase in optical density compared to the control case, again indicating no inhibitory compounds were produced by the fungi (Figure 4.8). Biomass hydrolysate from fungal growth on reed canary grass without any pretreatment was then tested for support of *S. cerevisiae*. While the amount of glucose released from reed canary grass was much lower compared to that released from cellulose (Figure 4.1.C-D), the yeast reached a similar optical density (Figure 4.7.A) when grown on this media. Measurements of sugar concentrations before and after yeast growth (Figure 4.7.B) revealed that the yeast consumed primarily glucose and fructose present in the fungal media, but also small amounts of xylose and arabinose to support growth. There was a reduction in overall sugars of 79% and 73% after yeast growth in *N. californiae* and *A. robustus* media, respectively, leaving primarily xylose and arabinose and a small amount of glucose behind.

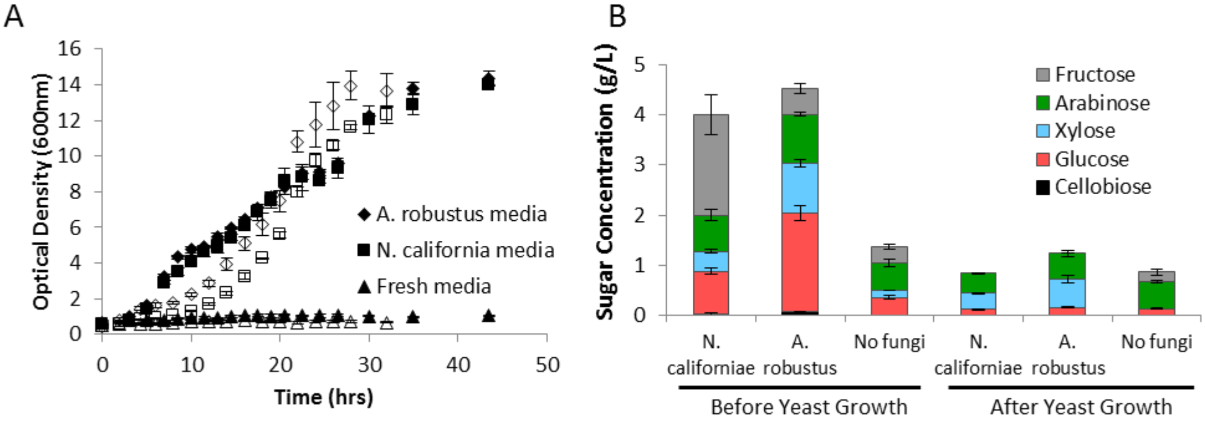


Figure 4.7. Fungal biomass hydrolysate supports growth of *S. cerevisiae*

A) Growth of *S. cerevisiae* on fungal spent media. Spent media containing crystalline cellulose broken down by the fungi into glucose (filled symbols) or reed canary grass broken down into glucose and other sugars (empty symbols). B) End-point sugar concentrations produced after fungal growth on reed canary grass and sugar concentration after yeast growth in spent fungal media.

The above results demonstrate that there was a wealth of sugars released from biomass by anaerobic gut fungi that may be handed off to another organism for production, and that a two-stage process is feasible. Further, the extent to which the yeast can remove the excess sugars suggest that presence of another organism may alleviate catabolite repression of biomass degrading enzymes in gut fungi during a simultaneous co-culture, increasing overall production of enzymes while improving enzyme efficiency by removing sugar-based inhibition of cellulases. Previous studies on microbial co-cultures and consortia for production have paired cellulolytic organisms, such as *Clostridium phytofermentans*[59], with production organisms, requiring cellulose as an input rather than biomass. *Trichoderma reesei* and *E. coli* have also been paired for production of isobutanol from biomass, but still rely on the use of pretreated biomass [60]. Gut fungi are capable of supplying sugars directly from crude biomass without any pretreatments. Furthermore, pairing to growth of *T. reesei* limits production to aerobic conditions, while the system proposed here is amenable to anaerobic and aerobic conditions, tailoring the process to the desired product.

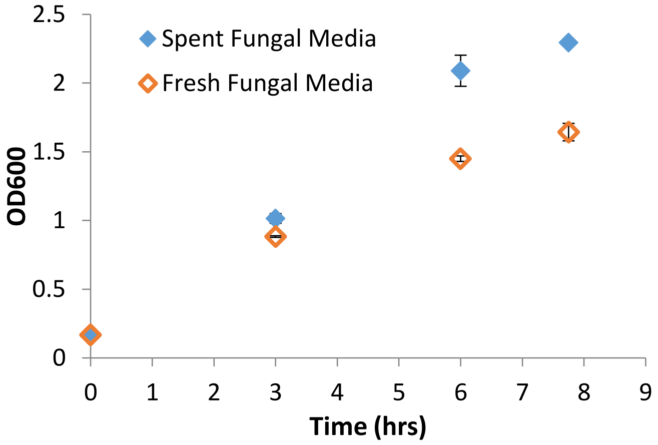


Figure 4.8. Growth of *Escherichia coli* on fungal cellulose hydrolysate.

A small increase in the growth of *E. coli* was observed in the fungal hydrolysate, or “spent” fungal media. However, *E. coli* are still able to grow on the contents of the complex fungal media, making sugar-dependent growth difficult to assess.

4.3. Conclusions

Anaerobic gut fungi efficiently hydrolyze crude biomass through a combination of mechanical disruption and enzymatic activity from a wide array of biomass degrading enzymes. They release excess amounts of sugars such as glucose, fructose, xylose, and arabinose during growth on crude biomass. Reconstruction of metabolic maps both validated growth experiment results and identified sugars that are more likely to accumulate from biomass hydrolysis alongside the most abundant glucose. These sugars can then be supplied to an additional microbe for bio-based production of a value-added chemical. We have demonstrated the ability of the fungal hydrolysate to support growth of the model organism, *S. cerevisiae,* presenting a co-culture based consolidated bioprocessing strategy that utilizes crude, rather than pretreated, biomass. While additional work may be required to improve enzymatic production and hydrolysis, our regulation studies provide a path forward for optimizing production of biomass degrading enzymes, identifying conditions that improve enzymatic production as well as potential repressors of biomass degrading enzymes. The two-stage fermentation approach described here allows for the consolidation of biomass pretreatment and hydrolysis into a single step to supply a monosaccharide-rich hydrolysate that can be fed to a model organism for growth and production. The second growth step allows for the precise control of the production bioreactor such that conditions can be optimized for the desired product rather than for fungal growth.

4.4. Materials and Methods

4.4.1. Culture maintenance and growth measurement

Anaerobic media preparation and gut fungal culture procedures were used throughout this work. Anaerobic gut fungi were routinely grown in 10 mL cultures of Medium C [148] containing ground reed canary grass (4 mm particle size) in 15 mL Hungate tubes. The tube headspace was filled with 100% CO_z and cultures were grown at 39°C. Cultures were transferred to new media every 3-5 days to continue growth. For differential expression experiments, gut fungi were grown in 80 mL of medium C in 120 ml serum bottles and all subsequent cultures were started from the same source culture. Fungi were grown on a variety of carbon sources including glucose (anhydrous, Thermo Fisher Scientific, Canoga Park, CA), maltose (Sigma-Aldrich, St Louis, MO), cellobiose (Sigma-Aldrich), Avicel (PH-101, 50 µm particle size, Sigma-Aldrich), corn stover, reed canary grass, switchgrass, and alfalfa stems; biomass substrates were provided by the USDA-ARS Research Center (Madison, WI). Soluble substrates were added to a final medium concentration of 5 g/L while particulate substrates were added to a final concentration of 10 g/L.

To obtain growth information, the pressure of fermentation gases was measured during growth. Accumulation of pressure in the headspace of the closed Hungate tubes is correlated to fungal growth and inversely correlated to substrate loss [152]. Cultures that accumulated pressure significantly more than the blank control (10 mL Medium C culture containing no carbon source, but inoculated with fungi) were considered positive for growth. Effective net specific growth rates were determined from the pressure accumulation data of 3 x replicate cultures during the phase of exponential gas accumulation.

For growth and sugar release experiments, gut fungal cultures were grown on Avicel and reed canary grass (4 mm granulated particles) in 10 mL cultures containing anaerobic Medium C. Cellulose cultures contained 100 or 200 mg of cellulose, and biomass cultures contained either 100 mg or 500 mg of reed canary grass. Pressure measurements were taken three times per day to track growth of the fungi. Aliquots of 0.1 mL were removed from cultures for glucose determinations using a YSI 2900 substrate analyzer with YSI 2365 glucose detection membrane kits (YSI Inc., Yellow Springs, OH).

4.4.2. RNA isolation

RNA was isolated from growing fungal cultures during the exponential growth phase using the Qiagen RNeasy Mini Kit (Qiagen, Valencia, CA). The protocol for plants and fungi was followed, including a liquid nitrogen grinding step to disrupt cell walls and an on-column

DNase digest. The RNA quality was determined through measurement on an Agilent Tapestation 2200 (Agilent, Santa Clara, CA) to obtain RINe scores. The total RNA quantity was determined by using Qubit Fluorometric Quantitation (Qubit, New York, NY) using the high sensitivity RNA reagents.

4.4.3. RNA Library Preparation and Sequencing

Sequencing libraries were prepped using an Illumina TruSeq Stranded mRNA library prep kit (Illumina Inc., San Diego, CA) following the kit protocol. Two separate libraries were created for each fungus. For each sample from *Neocallimastix californiae* 600 ng of total RNA was used while for each sample from *Anaeromyces robustus* 400 ng of total RNA was used as input for the library preparation. Starting quantities of RNA were determined by the lowest concentration sample to ensure equal starting material for each sample at the start of library preparation. Once the library preparation was completed, samples from each fungus were pooled together into two separate cDNA libraries with a final concentration of 10 nM. Each library was then diluted to 2 nM, denatured, diluted to 20 pM, prior to diluting to the final loading concentration of 1.8 pM. Libraries were sequenced on a NextSeq 500 (Illumina, San Diego, CA) using High Output 150 Cycle reagent kits. Samples for *N californiae* and *A. robustus* were sequenced on separate flow cells.

4.4.4. Metabolic map reconstruction

Enzymes present in the metabolic maps of isolated anaerobic fungi were determined based on the annotation of the transcriptomes, specifically, by the presence of EC numbers. Metabolic maps present in the KEGG database were filled in based on EC numbers present within the transcriptome annotations. All enzymes identified as present based on this initial analysis were checked for exact EC number presence to avoid false assignments based on incomplete EC numbers (i.e., to ensure the functionality of bisphosphoglycerate phosphatase, EC:3.1.3.13 was not reported based on the presence of general phosphatase functionality designated by class EC:3.1.3.-). Gaps in metabolic maps were then checked by searching the entire annotation, including BLAST and InterPro, for key words.

4.4.5. Expression data analysis

Counts of transcripts were quantified by using the RSEM analysis[123] present within the Trinity[122] programming package. Transcriptomes previously obtained[130] were used as reference templates to obtain count data. Expected counts from this analysis were then fed into the DESeq2 package[126] in the R programming language to determine statistically significant changes in expression with a minimum of one $\log_2$ fold change in expression and p-value ≤ 0.01. Results from all substrates were compared to the base case of glucose to determine fold change in expression of all transcripts. Bar plots showing change in expression were made using the raw transcripts per million (TPM)[123] output from the RSEM analysis.

4.4.6. Analysis of sugars (HPLC)

Sulfuric acid (0.85 M) was added (1 in 10 volumes) to fungal culture supernatants, that were then vortexed and allowed to stand for 5 min at room temperature. Nine volumes of water were added and the sample again vortexed briefly, centrifuged for 5 minutes at 21000xg, and the supernatants were extracted with a syringe and filtered into HPLC vials using a 0.22µm filter. Samples were run on an Agilent 1260 Infinity HPLC (Agilent, Santa Clara, CA) using a Bio-Rad Aminex HPX-87P column (Part No. 1250098, Bio-Rad, Hercules, CA) with inline filter (Part No. 5067-1551,Agilent,Santa Clara, CA), Bio-rad Micro-Guard De-

Ashing column (Part No. 1250118, Bio-Rad), and Bio-Rad Micro-Guard CarboP column (Part No. 1250119, Bio-Rad) in the following orientiation: Inline filter>De-Ashing>CarboP>HPX-87P. Samples were run with a water mobile phase at a flow rate of 0.5 mL/min and column temperature of 80°C. Signals were detected using a refractive index detector.

HPLC standards were created for cellobiose, maltose, sucrose, glucose, fructose, galactose, xylose, mannose, and arabinose. Each sugar was dissolved in medium C to create 10 g/L (w/v) stock solutions. Serial dilutions from this stock were used to create 1%, 0.1% and 0.01% standards and the above protocol was followed to run each standard.

4.4.7. Yeast and Bacteria Culture

Following release of sugar by gut fungi in cultures grown on cellulose, the liquid medium was removed from the Hungate tube using a syringe needle and placed in a sterile growth tube which was then inoculated with *Saccharomyces cerevisiae* (BJ5464) or *Escherichia coli* (XL1-Blue). Growth of cultures was tracked using optical density measurements at 600 nm (OD_{600}). Cultures were inoculated at a target OD_{600} of 0.5 for yeast cultures and 0.1 for bacteria cultures and grown aerobically in shaker incubators set to 30°C and 225 rpm for yeast, and 37°C and 225 rpm for *E. coli*.

5. Understanding the gut fungal response to catabolite repression

Several figures in this chapter are from Solomon et al., *Science*. 2016[130]. Reprinted with permission from AAAS.

5.1. Introduction

Agricultural wastes, invasive plant species, and energy crops are renewable, non-food resources for fermentable sugars that can be used to produce biofuels and chemicals[12,193]. However, the recalcitrant nature of lignocellulosic biomass makes it difficult to degrade as the lignin inactivates many biomass degrading enzymes[194]. This leads to a requirement for biomass pretreatments to physically separate the recalcitrant lignin from the sugar rich cellulose and hemicellulose prior to enzymatic hydrolysis to release monosaccharides[195]. Typical enzyme cocktails for hydrolysis require enzymes from multiple organisms, such as the filamentous fungus *Trichoderma reesei*, which only secrete a subset of the enzymes required to break down pretreated cellulosic substrates[196]. Fortunately, there are microbes that routinely degrade complex lignocellulose in nature, like those found in the digestive tracts of large herbivores that have evolved a full suite of enzymes to hydrolyze lignocellulosic substrates without pretreatment[63]. Among the organisms in the rumen microbiome are the anaerobic gut fungi. These fungi are the most primitive known free-living fungi[197] and are considered the primary colonizers of plant material in the rumen[66].

Anaerobic gut fungi follow a similar life cycle to that of the pathogenic chytrids, reproducing asexually via motile zoospores that colonize biomass substrates[62]. They rely on a combination of invasive rhizoidal growth and powerful secreted enzymes to degrade plant biomass[127,180]. To date, the strict anaerobic lifestyle along with complex nutritional requirements have hindered isolation attempts and molecular characterization[11] and precluded their application in industrial processes. To gain better insight into how gut fungi degrade complex biomass, we have used next generation sequencing to both annotate their repertoire of enzymes and study the regulation of carbohydrate active enzymes (CAZymes). In other cellulolytic fungi, carbon catabolite repression is an important regulatory mechanism to regulate the of genes encoding both cellulases and hemicellulases[182] – largely to prevent the overproduction of these metabolically expensive enzymes under environmental conditions where they are not needed. Carbon catabolite repression results when a more readily used carbon source is available and interacts with transcription factors to inhibit the expression of other carbon sources. In cellulolytic fungi, this occurs when glucose is available and represses the expression of enzymes required to process more complex cellulose[182]. Carbon catabolite repression is also common in other microorganisms such as yeast[198] and bacteria [199]. This type of repression allows microbes to use certain carbon sources preferentially in order to ensure the fastest growth possible[199]. For industrial strains, regulation such as this is often a target for genetic engineering to remove the inhibition and maximize enzyme production capability[200].

Here we have used a simple carbohydrate, glucose, to trigger carbon catabolite repression in anaerobic gut fungi. RNA sequencing for various time points after the glucose pulse allowed us to examine how the global expression patterns of genes in gut fungi remodel in response to a carbon catabolite repressor. Clustering of the expression results allows us to group similarly regulated transcripts into regulons that can be used to predict the function of unannotated transcripts, identifying novel enzymes with potential to exploit for biomass degradation. We first performed these analyses for a single isolated strain, *Piromyces finnis*. The same analysis was then performed with two other strains, *Neocallimastix californiae* and

Anaeromyces robustus to determine how conserved the regulation patterns were across fungi within different genera.

5.2. Results and Discussion

5.2.1. Glucose carbon catabolite repression in Piromyces finnis

All three species of fungi exhibit similar growth rates on substrates ranging in complexity from simple sugars to complex biomass (Figure 5.1.A, Table 2.1). Using functional annotations of the transcriptomes of these anaerobic gut fungi, we assigned putative functions to many transcripts based on their sequence similarity to genes within online databases like NCBI[101] or protein domains within the InterPro database[102]. This allowed us to identify the variety of CAZymes[136] with broad functionality that allow the gut fungi to effectively degrade lignocellulose (Figure 5.1.B, Figure 2.5). It is also likely that they are capable of tight control of the expression of these enzymes such that they tailor their enzyme repertoire to the carbon source they are presented with.

Microbes are frugal organisms that typically repress alternative carbohydrate utilization pathways when glucose is available since it is easier to catabolize. In simpler microorganisms, this typically means transitioning to glucose metabolism from sugars that require extra enzymatic steps to feed into glycolysis by modulating expression of enzymatic pathways and transporters[198,199]. In more complex, cellulolytic microorganisms this means shutting down the expression of enzymes required to hydrolyze glucose from cellulose. Given the evidence of carbon catabolite repression of cellulases in other cellulolytic fungi such as *Trichoderma reesei, Aspergillus niger,* and *Neurospora crassa*[182], we decided to search for similar regulation mechanisms in anaerobic gut fungi and hypothesized that these regulation patterns would identify critical enzymes for biomass degradation that evade annotation by sequence similarity to known biomass degrading enzymes. Gut fungal cultures were grown on reed canary grass and pulsed with 5 mg of glucose during exponential growth (Figure 5.2.A). RNA sequencing was performed on cell samples at various time intervals after glucose pulse to quantify the change in expression of all transcripts immediately after the pulse and after all glucose was consumed.

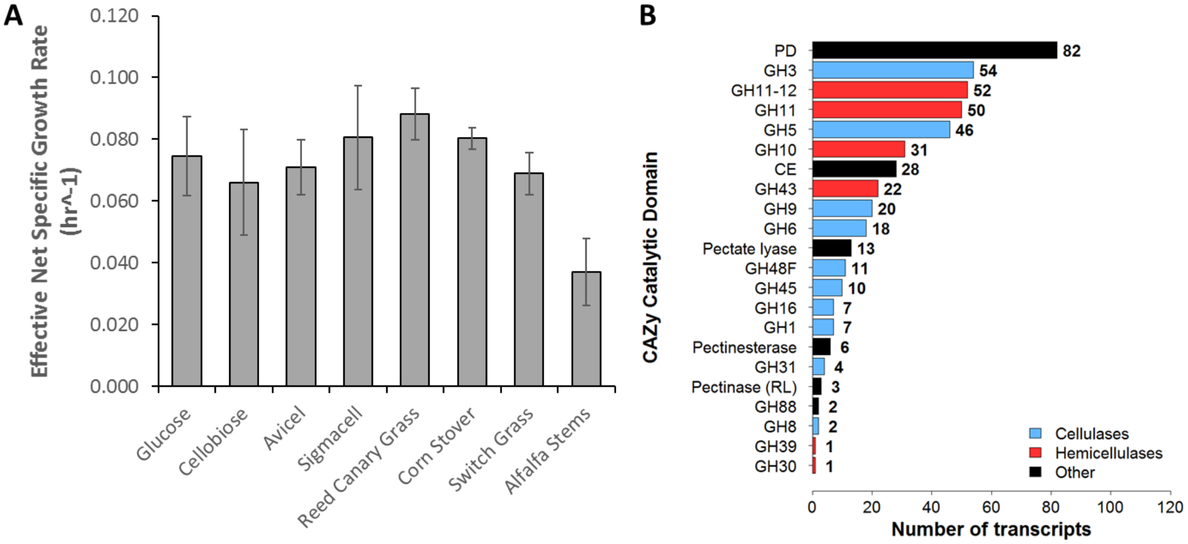


Figure 5.1. *P. finnis* demonstrates consistent growth on simple and complex substrates

A) Effective net specific growth rates determined from pressure accumulation measurements for *Piromyces finnis* reveal similar growth rates regardless of substrate complexity. **B)** *P. finnis* possesses a wide variety cellulases, hemicellases, and accessory enzymes for biomass degradation. GH – glycoside hydrolase; CE – carbohydrate esterase; PD – polysaccharide deacetylase. From Solomon et al., *Science*. 2016[130]. Reprinted with permission from AAAS.

In *Piromyces finnis*, glucose pulse resulted in more than a 2-fold change in 374 transcripts with p ≤ 0.01 (Figure 5.2.B). One third of these regulated transcripts contained CAZyme domains. As expected, these CAZymes were almost exclusively repressed in response to glucose addition. These changes in expression were also reflected in the activity of cellulose precipitated enzymes in the culture supernatant (Figure 5.2.C). After the pulsed glucose was consumed by the fungus, expression returned to the baseline levels measured at time zero along with cellulose precipitate enzyme activity on carboxymethyl cellulose (CMC).

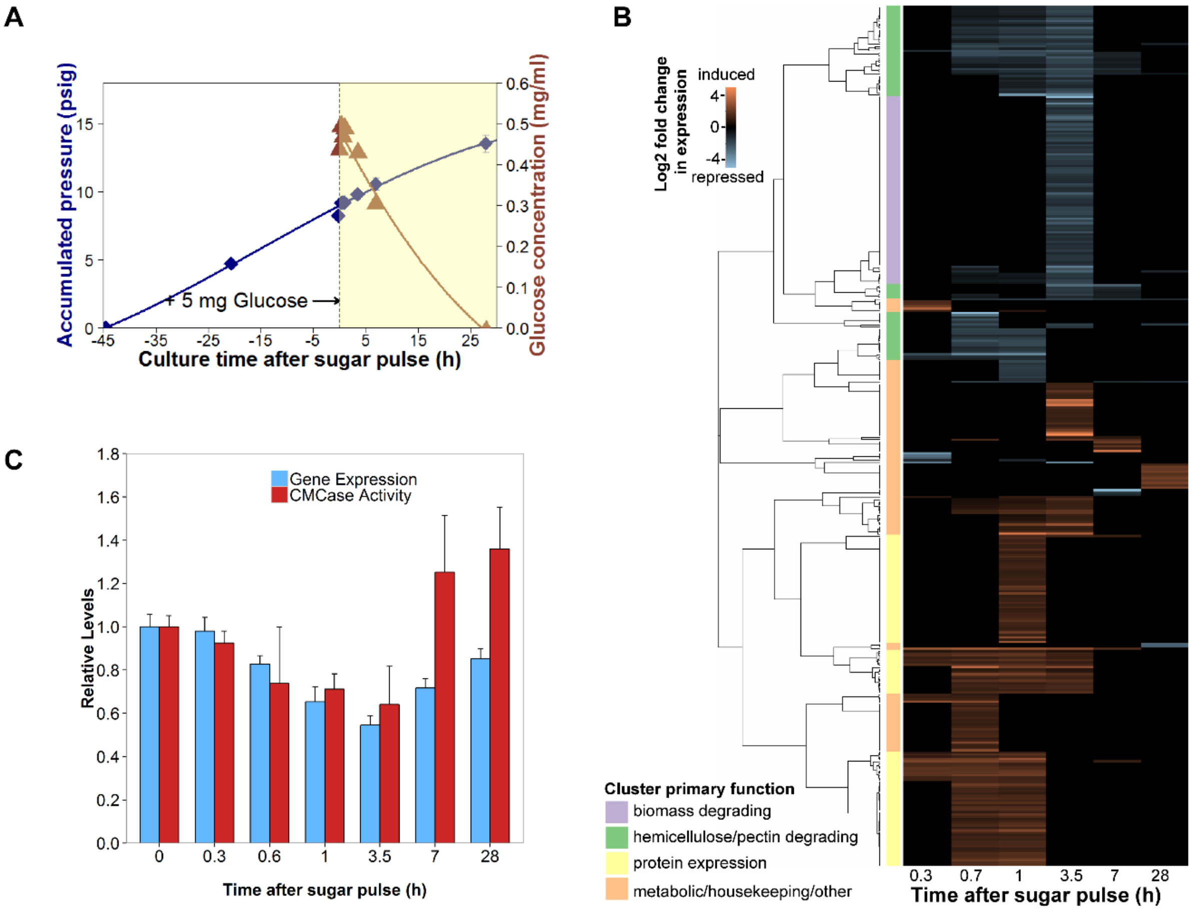


Figure 5.2. *P. finnis* regulates carbohydrate active enzymes in response to glucose pulse

A) Glucose pulsed into *P. finnis* cultures growing on reed canary grass during exponential growth. B) This pulse resulted in repression of CAZymes and enhanced expression of metabolic and housekeeping genes. C) CAZyme expression changes were consistent with enzyme activity of cellulose precipitated enzymes in the culture supernatant. From Solomon et al., *Science*. 2016[130]. Reprinted with permission from AAAS.

The transcriptional response occurred rapidly compared to the rate of growth of *P. finnis*. Significant changes in the expression of housekeeping genes and genes involved in protein expression were observed as soon as 20 minutes after the glucose pulse. Many biomass degrading enzymes were downregulated within 40 minutes after the pulse, with the fastest response specifically among hemicellulase encoding transcripts. Cellulases and other biomass degrading enzymes with a broader range of activities, including hemicellulases and accessory enzymes, were downregulated on a longer timescale of 3.5 hours. More responsive regulation of hemicellulases is also conserved among higher fungi[201-204] and is believed to have arisen due to the structure of lignocellulose itself. Hemicellulose and pectin surround cellulose, leading to the need for cellulases only after the hemicellulose and pectin have been digested.

Table 5.1. Transcripts clustered by expression level have conserved functions

Cluster	Conserved Function	Cluster size	Up or down regulated
1	Hemicellulose/Pectin Degrading	7	Down
2	Hemicellulose/Pectin Degrading	22	Down
3	Hemicellulose/Pectin Degrading	17	Down
4	Metabolic/housekeeping/other	6	Up
5	Hemicellulose/Pectin Degrading	6	Down
6	Biomass degrading	82	Down
7	Protein expression	50	Up
8	Metabolic/housekeeping/other	25	Up
9	Metabolic/housekeeping/other	3	Down
10	Metabolic/housekeeping/other	17	Up
11	Protein expression	47	Up
12	Metabolic/housekeeping/other	3	Up
13	Protein expression	19	Up
14	Metabolic/housekeeping/other	4	Down
15	None	1	Down
16	Metabolic/housekeeping/other	11	Up
17	Metabolic/housekeeping/other	5	Up
18	None	2	Up
19	Metabolic/housekeeping/other	24	Up
20	Metabolic/housekeeping/other	9	Down
21	Hemicellulose/Pectin Degrading	14	Down

From Solomon et al., *Science*. 2016[130]. Reprinted with permission from AAAS.

An in depth cluster analysis[205] of the regulatory patterns observed for these transcripts also revealed coordinated expression of biomass degrading enzymes. Hierarchical clustering revealed 21 distinct clusters, or “regulons”, containing glucose-responsive genes of related function (Table 5.1). These clusters likely respond to the same regulatory mechanism and allow for broad ranging response to a single stimulus to achieve a specific goal. In the case of this experiment that goal is to transition to the metabolism of a simpler carbon source by shutting down unnecessary production of cellulases. The fact that many of these regulons contained transcripts with conserved predicted function, suggests that they may be used to identify divergent proteins that perform the same, or similar, function, but are distinct from any other known protein sequences. Divergent sequences are likely to be present in the anaerobic gut fungal transcriptomes due to the poor characterization and lack of genome sequences for these organisms. While our transcriptome annotations identified putative functions for many transcripts, there are many more transcripts that did not have significant similarity to known proteins or protein domains to predict their putative function. These sequences along with the regulons of biomass degrading enzymes provided an opportunity to select candidate sequences that may be novel biomass degrading enzymes. Within regulons associated with almost exclusively biomass degrading function, we identified 17 such transcripts that may represent novel biomass degrading enzymes (Table 5.2).

While downregulated clusters were primarily comprised of biomass degrading enzymes, the functions within upregulated clusters were consistent with those involved in logarithmic growth on glucose and were likely upregulated as the fungal cells shifted to more rapid growth on glucose. Protein expression clusters contained transcripts for predicted chaperone proteins, rRNA processing proteins, elongation factors, and enzymes involved in amino acid and nucleotide synthesis. Metabolic/housekeeping clusters were less conserved in specific functions, but included a broad array of metabolic, protein expression, and housekeeping genes involved in processes like cell wall synthesis, transport, and central metabolism.

Table 5.2. Unannotated transcripts in biomass degrading regulons

Unannotated transcript ID	**Cluster**
comp12262_c0_seq1	5 – Biomass degrading
comp12026_c1_seq1	6 – Biomass degrading
comp12362_c0_seq1	6 – Biomass degrading
comp7503_c0_seq2	6 – Biomass degrading
comp11992_c0_seq2	6 – Biomass degrading
comp11882_c0_seq1	6 – Biomass degrading
comp11735_c0_seq1	6 – Biomass degrading
comp12028_c12_seq1	6 – Biomass degrading
comp7496_c0_seq1	6 – Biomass degrading
comp5143_c0_seq1	6 – Biomass degrading
comp10778_c1_seq1	6 – Biomass degrading
comp13233_c0_seq1	6 – Biomass degrading
comp6536_c0_seq1	6 – Biomass degrading
comp11012_c2_seq1	6 – Biomass degrading
comp7326_c0_seq1	6 – Biomass degrading
comp14924_c0_seq1	21 – Hemicellulose/Pectin Degrading
comp11723_c0_seq2	21 – Hemicellulose/Pectin Degrading

From Solomon et al., *Science*. 2016[130]. Reprinted with permission from AAAS.

Overall, the distinct clustering of transcripts based on their regulation patterns into groups with conserved functions suggests that this is a valuable tool to identify unique enzymes with great potential for industrial application. Anaerobic gut fungi, including *P. finnis*, possess an incredible ability to degrade biomass and it is likely that they possess enzymes that are unlike those already characterized.

5.2.2. Sequence analysis of co-regulated transcripts

Unannotated, co-regulated transcripts (Table 5.2) that were identified in *P. finnis* were further examined through additional sequence analysis. While alignment to known sequences in gene databases yielded no significant functional prediction, based on their regulation pattern these transcripts are expected to be involved in lignin degradation, cellulosome structure, or uncharacterized cellulase/hemicellulase/accessory function. Proteomic and genomic analysis of *P. finnis* has identified gene sequences for cellulosome structure scaffoldin proteins using Hidden Markov Models to identify conserved sequence motifs[82]. These non-catalytic proteins are expected to coordinate biomass degrading enzyme complex formation through interaction between cohesin domains on the scaffoldin protein and dockerin domains on biomass degrading enzymes. We identified two scaffoldin sequences within the co-regulated set of transcripts that had open reading frames (ORFs) 4,230 and 5,013 base pairs (bp) in length (1,410 and 1,671 amino acids). These large proteins are not likely to have novel catalytic function, but are important for efficient biomass degradation, explaining their presence in the biomass degrading regulons.

To identify the most promising candidates for novel biomass degrading function, we aligned the co-regulated sequences from *P. finnis* to the transcriptomes of *A. robustus* and *N. californiae*. We hypothesized that those sequences that were conserved among multiple fungi were most likely to have catalytic function. This alignment yielded no significant matches for 10 co-regulated transcripts, suggesting that these sequences are less likely to play a role in biomass degradation. Many of these transcripts also had short predicted open reading frames, such that they may not represent protein encoding genes. We also identified four non-scaffoldin sequences in both *A. robustus* and *N. californiae* that were significantly similar to co-regulated transcripts in *P. finnis* (Table 5.3). This identifies these transcripts as the best place to start biochemical studies to express and characterize unannotated proteins for potential biomass degrading activity.

Table 5.3. Sequence alignment of *P. finnis* co-regulated transcripts to *A. robustus* and *N. californiae*

P. finnis **Unannotated transcript ID**	**Scaffoldin?**	**Sequence hit *A. robustus***	**Sequence hit *N. californiae***
comp12262_c0_seq1	N	Locus2793v1rpkm17.21	-
comp12026_c1_seq1	N	-	-
comp12362_c0_seq1	N	Locus1323v1rpkm59.77	Locus936v1rpkm136.63
comp7503_c0_seq2	Y	Locus2632v1rpkm18.90	Locus2411v1rpkm34.85 Locus4280v1rpkm14.40 Locus12584v1rpkm1.78 Locus12584v2rpkm0.00_PRE
comp11992_c0_seq2	Y	-	-
comp11882_c0_seq1	N	Locus721v1rpkm140.98	-
comp11735_c0_seq1	N	-	-
comp12028_c12_seq1	N	-	-
comp7496_c0_seq1	N	-	-
comp5143_c0_seq1	N	-	-
comp10778_c1_seq1	N	-	-
comp13233_c0_seq1	N	-	-
comp6536_c0_seq1	N	Locus4155v1rpkm8.56	-
comp11012_c2_seq1	N	-	-
comp7326_c0_seq1	N	-	Locus6670v1rpkm6.81
comp14924_c0_seq1	N	-	Locus3185v1rpkm22.91 Locus1571v1rpkm64.27
comp11723_c0_seq2	N	-	-

*Open reading frame sequences for these transcripts are presented in Appendix D (Ch 7.4)

While there are many known enzymes that carry out the hydrolysis of cellulose and hemicellulose[136] both under anaerobic and aerobic conditions[18,206], there are many fewer known enzymatic mechanisms for the degradation of lignin. Those that are known were identified in aerobic fungi, such as white rot fungi, and rely on the presence of oxygen for activity[207]. Anaerobic gut fungi efficiently degrade lignocellulosic biomass and it is therefore likely that they possess unknown mechanisms for the manipulation and depolymerization of lignin that protects the cellulose and hemicellulose in plant cell walls.

5.2.3. Regulons may contain candidate lignin breakdown enzymes

The ability of gut fungi to hydrolyze the cellulose and hemicellulose found within plant cell walls has been well characterized[10,11,66,79,82,130]. However, relatively little work has been done to understand how the gut fungi affect lignin in plant biomass and identify the enzymes responsible for that activity. Previous work has identified approximately 20% reduction in lignin by characterizing the plant tissue composition before and after incubation with gut fungi compared to greater than 50% of cellulose and hemicellulose was digested[208]. Other work has suggested that this weight loss of lignin was primarily a result of the solubilization of lignin with fungi incubated samples showing increases in the concentration of *p*-hydroxyl, vanillyl, syringyl, and cinnamyl phenols[209]. However, this study did not examine the effect of isolated strains of gut fungi, but rather incubation of plant material with rumen fluid using various antibiotic treatments to target bacterial and fungal populations.

Our analysis of the growth of the isolated strain *Anaeromyces robustus* on switchgrass identified similar results. Gut fungi primarily decompose the cellulose and hemicellulose within plant material, resulting in an enrichment in lignin content after incubation with fungi from 18% to 22% (Figure 5.3). These results suggest that gut fungi may be ignoring the lignin and focusing on hydrolysis of cellulose and hemicellulose, leading to this increase in lignin content. However, analysis of total phenol concentration revealed that soluble phenols were released from plant material into the liquid media. A Prussian assay was used to determine phenol concentration in culture media autoclaved with switchgrass, in the same culture media after incubation of switchgrass with fungi, and in culture media containing fungi grown on soluble sugars. This analysis identified an increase of phenol concentration after fungal incubation from 0.08 M to 0.25 M.

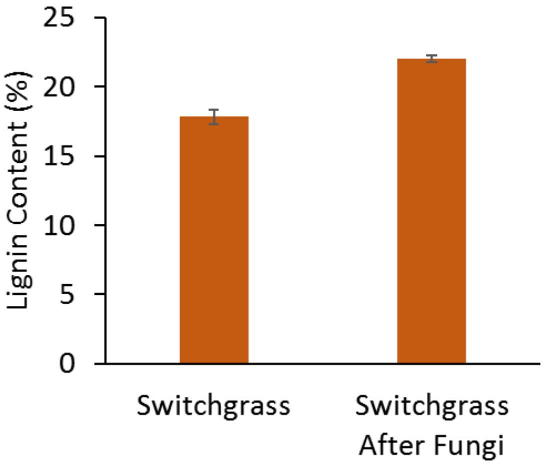


Figure 5.3. Acid hydrolysis reveals lignin content increased by fungal growth

Lignin content was determined as the acid insoluble fraction of plant biomass after acid hydrolysis treatment. These revealed an increase in lignin content from 18% in switchgrass not incubated with fungi to 22% after incubation with *Anaeromyces robustus*. This analysis was completed in the Foston lab at Washington University in St. Louis by Marcus Foston and James Meyer.

The occurrence of soluble phenolic compounds in the liquid media of fungal cultures grown on switchgrass suggests that there is some enzymatic mechanism in place capable of releasing small subunits of the lignin biopolymer. However, most of the known enzymatic mechanisms for depolymerization of lignin are found in aerobic organisms, like white rot fungi. These enzymes, like peroxidases and laccases use oxidative mechanisms to depolymerize lignin[207] and as such are not likely to be responsible for lignin degradation in anaerobic microorganisms. The regulation patterns identified in the carbon catabolite repression of transcripts in *P. finnis* (Figure 5.2, Table 5.1, Table 5.2) as well as potential patterns *N. californiae* (Figure 5.7) and *A. robustus* (Figure 5.10) present opportunities to identify novel lignin active enzymes responsible for this degradation.

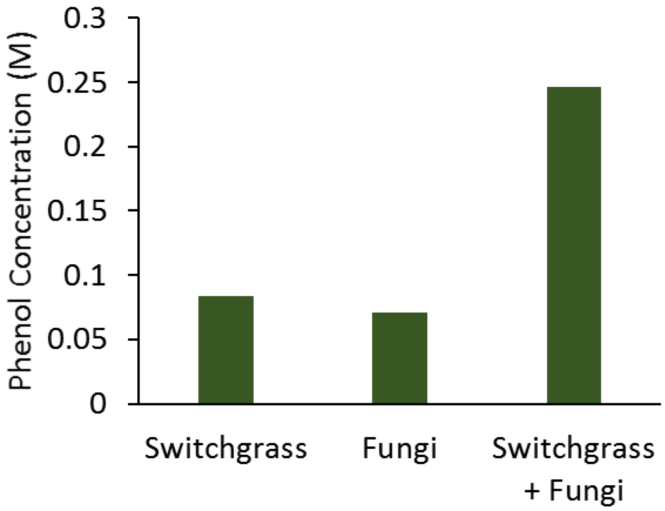


Figure 5.4. Phenol concentration in liquid media

A Prussian assay was used to identify broad phenol concentration in the liquid media of fungal cultures. Phenol concentration after incubation of switchgrass with gut fungi resulted in a drastic increase compared to switchgrass autoclaved in media without fungal incubation and fungi grown on soluble sugars. This analysis was completed by the Foston lab at Washington University in St. Louis by Marcus Foston and James Meyer.

5.2.4. Substrate based tuning of biomass degrading enzyme expression in P. finnis

To develop a full understanding of the regulatory role of key biomass degrading enzymes in *P. finnis* we also performed RNA sequencing on cultures grown on isolated carbon sources ranging in complexity similar to the experiments completed for *A. robustus* and *N. californiae* in Chapter 4. *P. finnis* was grown on glucose, cellobiose, microcrystalline cellulose (Avicel$^{®}$), crystalline cellulose filter paper, and reed canary grass. RNA was then isolated from each of these substrates and sequenced for differential expression analysis, comparing expression on each substrate to the expression levels in the base case of glucose growth. This study revealed a significant remodeling of transcriptome expression as a function of substrate availability. Approximately 10% of all transcripts (2,596) showed significant changes in expression on at least one of these substrates compared to glucose. Among these transcription changes, there was significant remodeling of the carbohydrate active enzyme expression.

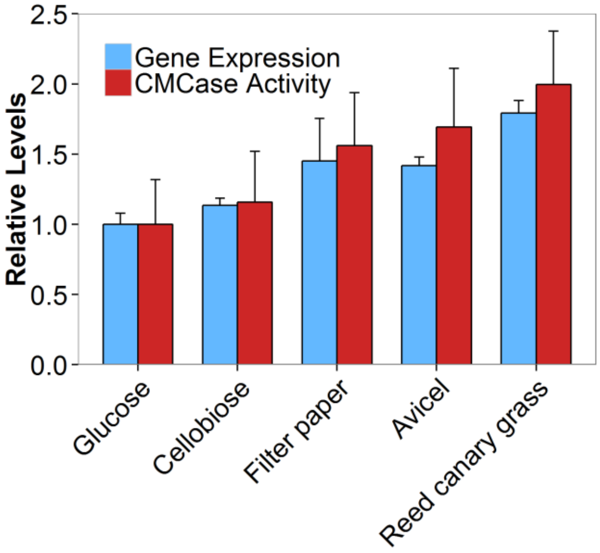


Figure 5.5. Expression levels of carbohydrate active enzymes on different substrates

The relative expression of carbohydrate active enzymes increased as a function of substrate complexity and the activity of cellulose precipitated enzymes in the supernatant also increased with the complexity of the substrate. From Solomon et al., *Science*. 2016[130]. Reprinted with permission from AAAS.

The expression levels as a percentage of overall transcript expression for carbohydrate active enzymes increased as the complexity of the carbon source increased. The highest expression levels were observed during growth on reed canary grass and the lowest during growth on glucose (Figure 5.5). This is not surprising since glucose was identified as a repressor of a broad range of carbohydrate active enzymes during the glucose pulse experiment. However, it is interesting that the expression is gradually modulating moving from glucose, to cellobiose, to cellulose, and finally to reed canary grass. This suggests that there may be multiple regulatory strategies for the expression of the full suite of carbohydrate active enzymes.

To look more in depth at the regulation of specific classes of CAZymes, gene set enrichment analysis (GSEA)[210,211] was completed. This allows the examination of gene sets containing only transcripts assigned to specific glycoside hydrolase (GH) classes across all growth conditions. The expression levels of all genes within the gene set are used to determine if the entire set is enriched under the conditions tested. This means that a gene set will not be identified as enriched if only a few members of a large set are upregulated, but only if the majority of members are regulated. For *P. finnis* this analysis was completed for gene sets containing transcripts annotated as cellulases (e.g. – GH5, GH6, GH9), hemicellulases (e.g. – GH10, GH11, GH11/12, GH43), and accessory enzymes (e.g. – carbohydrate esterases, pectin degrading). Additional gene sets examined include dockerin tagged transcripts expected to be part of the fungal cellulosome complex, putative antisense RNA, and glucose responsive transcript clusters identified in the glucose pulse differential expression analysis (Figure 5.6).

With increasing complexity of substrate, the number and functional diversity of CAZymes domains increased. This includes enrichment of GH5 and GH10 gene sets during growth on cellobiose, a dimer of glucose and product of enzymatic digestion of cellulose. Fungal cellulosome associated, or dockerin containing, transcripts were enriched on cellulosic filter paper, Avicel, and reed canary grass, presumably to allow for more synergistic degradation approaches for cellulose. Interestingly, substrates that did not contain hemicellulose induced expression of seemingly unnecessary hemicellulase gene sets such as GH10. This suggests that there may be a common regulatory network for at least a subset of the cellulases and hemicellulases produced by this gut fungal strain. However, there are likely additional regulatory strategies in place as GH11 and GH11/12 transcripts only showed enrichment during growth on reed canary grass.

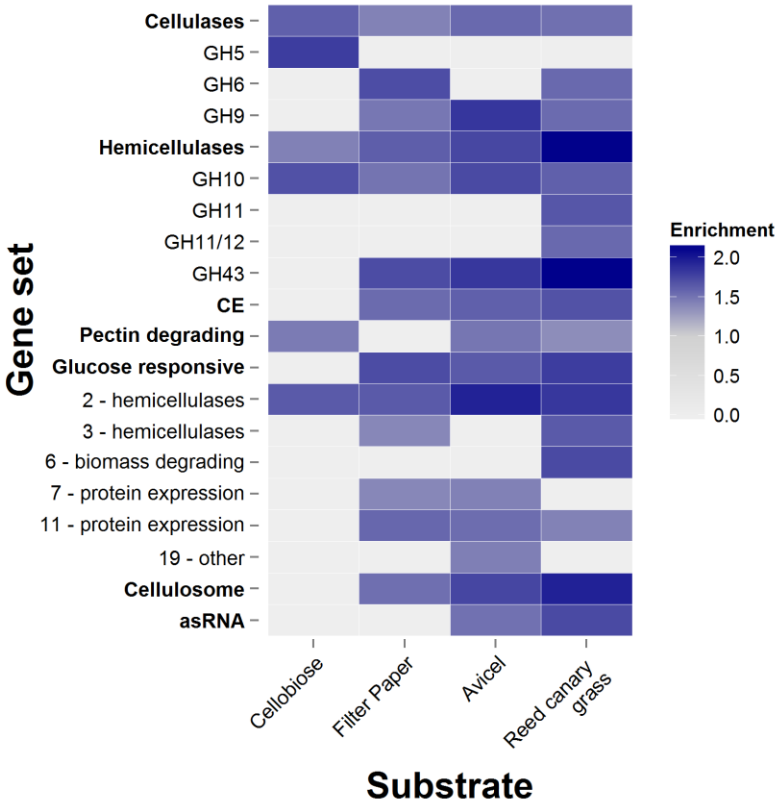


Figure 5.6. Gene set enrichment analysis (GSEA) for key CAZyme families in *P. finnis*

GSEA reveals enrichment of specific GH classes on more complex substrates when compared to the expression on glucose. Growth on reed canary grass is enriched for a broad range of GH families including cellulases, hemicellulases, and carbohydrate esterases. Cellobiose and cellulose growth conditions are primarily enriched in cellulases. From Solomon et al., *Science*. 2016[130]. Reprinted with permission from AAAS.

GSEA also revealed shifts between enzyme types for similar reactions, suggesting a highly specific, tailored response to different substrates. During growth on cellobiose, which requires the activity of β-glucosidases (GH5, GH9) to cleave it into glucose molecules, GH5 transcripts were enriched, but during growth on Avicel, filter paper, and reed canary grass, GH9 transcripts were enriched. This transition suggests possible synergies between all expressed enzymes and has implications for enzyme formulations for cellulose degradation.

Glucose responsive genes identified in the glucose perturbation experiment (Figure 5.2) showed enrichment under all conditions except for growth on cellobiose. This result is not surprising as the genes were repressed by glucose that should be absent or present in very low concentrations, during growth on complex substrates. It is possible that cellobiose is cleaved to glucose rapidly enough that a similar repression is observed under that growth conditions. Gene sets based on clusters, or regulons, (Table 5.1) also showed enrichment on more complex substrates. Protein expression clusters containing proteins such as chaperonins and rRNA processing proteins were enriched on insoluble substrates, indicating their role in mediating production and folding of lignocellulolytic enzymes necessary during growth on these conditions. One hemicellulase regulon was enriched under all non-glucose conditions, suggesting that they have a role in initial degradation of biomass for sensing and signaling of insoluble substrates to trigger expression of additional enzymes. When glucose is not available, these enzymes are expressed at a basal level and begin to degrade cellulosic material to provide soluble sensing molecules to trigger a specific catabolic response.

5.2.5. Remodeling of Neocallimastix californiae *transcriptome in response to glucose pulse*

The glucose pulse regulation experiment was replicated with *Anaeromyces robustus* and *Neocallimastix californiae* to determine how conserved the regulons containing biomass degrading genes were across anaerobic fungi. In addition to RNA sequencing of cultures pulsed with glucose, RNA-Seq was completed on a control set of cultures that were not pulsed with glucose to ensure that differences in expression observed were truly a function of the glucose pulse rather than the stage of growth. Analysis of the entire transcriptome of each fungus showed a large amount of transcriptional remodeling not only in response to the glucose pulse, but also in response to continued growth.

Table 5.4. *Neocallimastix californiae* glucose pulse regulation summary

	# Transcripts Regulated	
Time (hrs)	Glucose Pulse	No Glucose Pulse
0.6	269	527
1.1	515	600
2	970	1269
4	691	1880
6	498	2187
8	1223	2061
24	3908	3780
Total Unique Transcripts	4969	5908
Total Unique GH Transcripts	412	373
Up Regulated through 8 hrs	68	175
Down Regulated through 8 hrs	249	91

Addition of a 5-mg glucose pulse to cultures of *N. californiae* supported on reed canary grass was followed by a significant change in gene expression (Table 5.4, Figure 5.7). In total, after the pulse 4,969 transcripts showed significant regulation ($\log_2$-fold change > 1; $p < 0.01$). At early time points after the pulse, a few hundred transcripts were regulated, but a greater number of transcripts were regulated at later times, with the greatest number of regulated transcripts at 24 hours after glucose pulse. Surprisingly, more transcripts (5,908) were regulated in the control cultures that were not pulsed with glucose, not only in terms of total unique transcripts, but also in terms of total number of transcripts regulated at each time point except for 24 hours after pulse. While a greater number of regulated transcripts was expected in the pulsed case where a stimulus should be triggering transcriptional changes, this data highlights how dynamic gene expression is in these organisms. Even during normal growth, there are expression changes that may depend on a variety of cues and cellular processes.

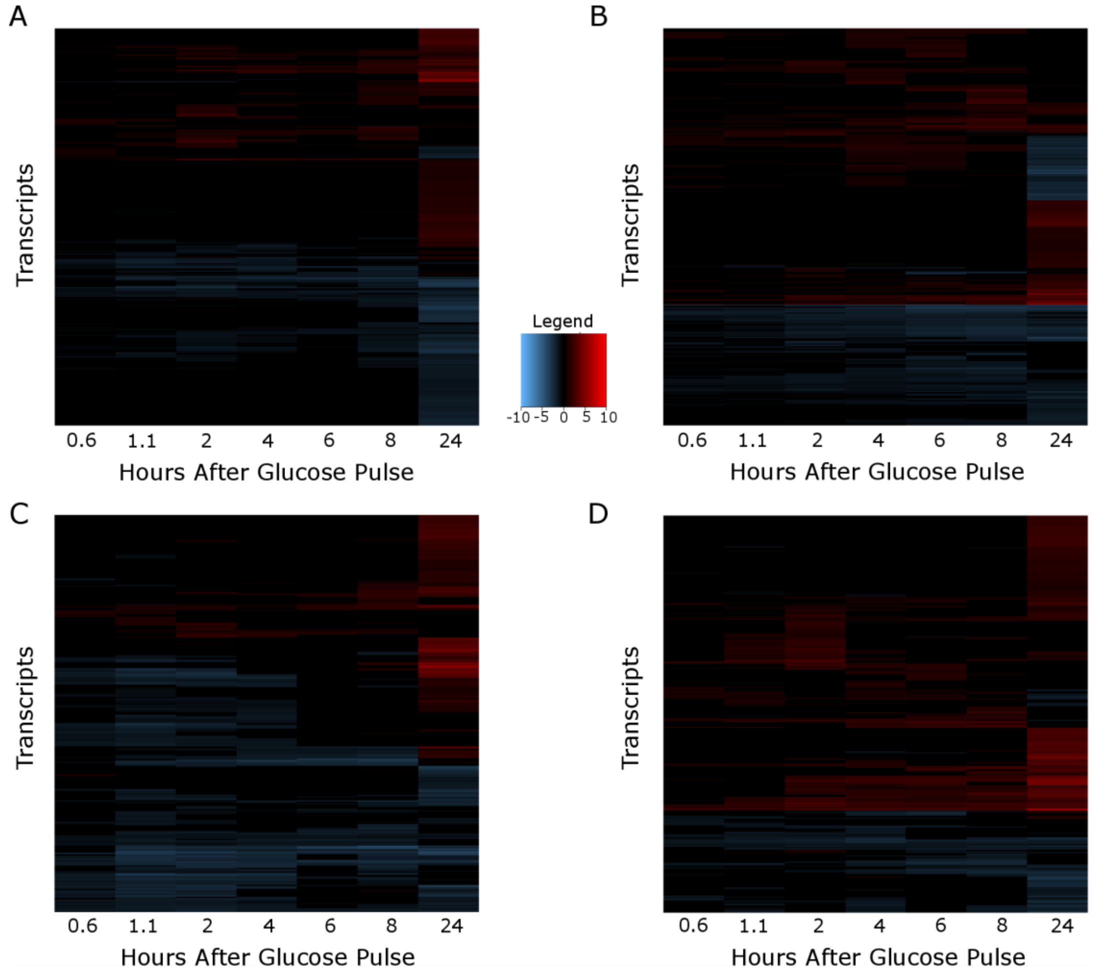


Figure 5.7. *Neocallimastix californiae* transcriptome response to glucose pulse

Many transcripts were regulated both in response to glucose pulse (A) and in the un-pulsed control cultures (B). More downregulated transcripts exist in the pulsed samples, but the difference is not immediately obvious compared to the un-pulsed control. When examining only the CAZymes, there is a significant difference in the pulsed (C) and un-pulsed (D) results, with much more downregulation in pulsed samples and upregulation in un-pulsed samples.

Examining overall transcriptional changes does not tell the entire story in these two sets of cultures. While more transcripts are regulated in the control samples, the regulation of carbohydrate active enzymes is drastically different in each of the two experimental cases.

The pulsed samples showed regulation of slightly more CAZymes, but the direction of their regulation is also important. Throughout the first eight hours after the pulse, the cells are expected to respond to the pulse as the added glucose is depleted. During this time, pulsed cultures showed down-regulation of 249 and up-regulation of only 68 CAZyme transcripts. However, un-pulsed control cultures showed down regulation of only 91 CAZyme transcripts and up-regulation of 175. Thus, in response to the pulse, *N. californiae* shows considerable remodeling of CAZyme expression compared to the un-pulsed controls. However, due to the large number of regulated transcripts, regulons enriched in specific functions and candidate genes for novel biomass degrading enzymes were not identified. The glucose pulse was nearly depleted after 8 hours (Figure 5.8) and entirely depleted after 24 hours. This was reflected in the expression changes as these time points marked a shift in up-regulation of many CAZymes (Figure 5.7.C).

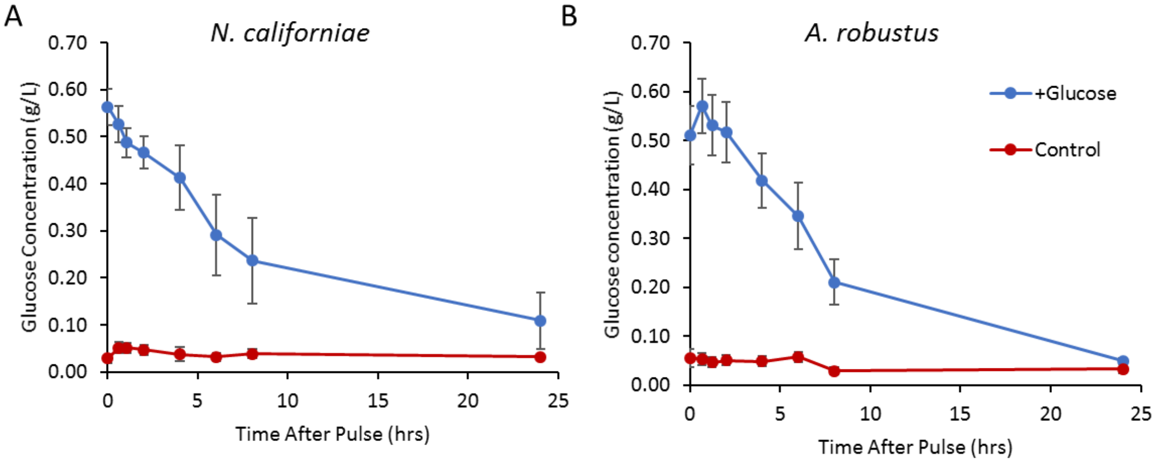


Figure 5.8. Glucose depletion by *N. californiae* and *A. robustus*

A) Glucose depletion by *N. californiae* after introduction of a 5-mg pulse of glucose to 10-mL batch cultures. B) Glucose depletion by *A. robustus* after introduction of a 5-mg pulse glucose to 10-mL batch cultures. For both fungi, glucose is mostly (but not entirely) depleted until between 8 and 24 hours.

Although there was downregulation of many CAZyme transcripts in response to the glucose pulse compared to the control (Table 5.4), the overall expression of CAZyme transcripts did not show significant changes in *N. californiae*. Figure 5.9 shows the expression of all cellulases, hemicellulases, and accessory enzymes as Transcripts Per Million (TPM). TPM is a normalized measurement of expression that equalizes the total expression in each sample, making it easier to compare samples with differing numbers of sequencing reads used to determine expression counts[123]. In this way, we can examine how the percent of total expression assigned to these CAZymes changes after the pulse. At all time points after the pulse, the total expression of these CAZyme classes remains the same. While more CAZyme transcripts were downregulated after the pulse, the total expression of all CAZymes did not change, suggesting that instead of shutting down biomass degrading activity, *N. californiae* shifted expression to alternative CAZymes.

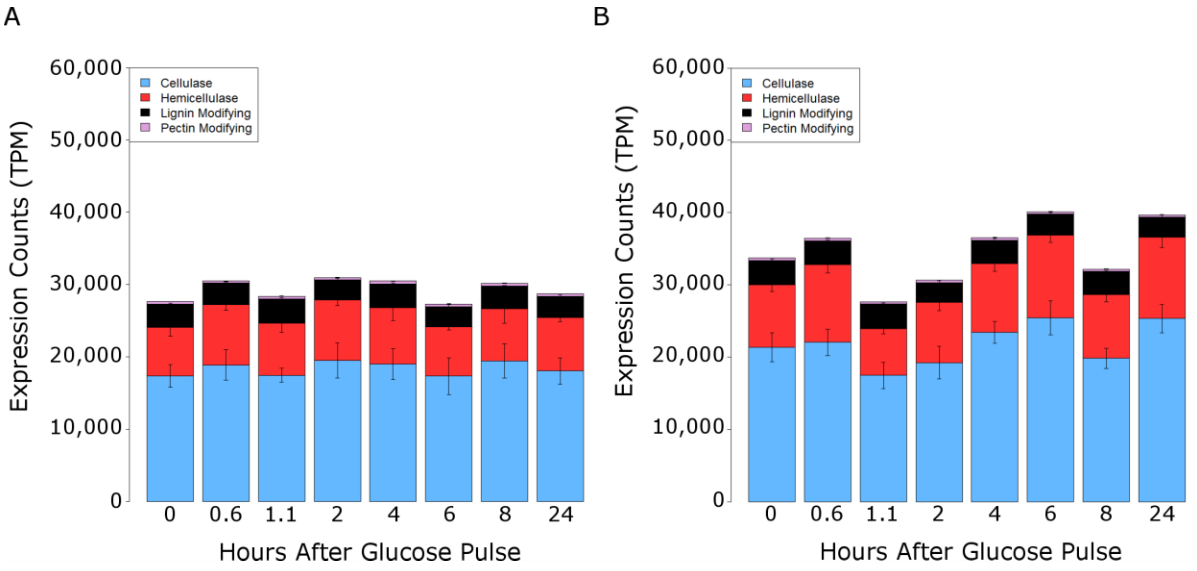


Figure 5.9. *Neocallimastix californiae* CAZyme expression after glucose pulse

A) Total expression levels of CAZymes after glucose pulse was added to growing culture of *N. californiae*. B) Total CAZyme expression in the control set of cultures that were not pulsed with glucose.

5.2.6. *Remodeling of* Anaeromyces robustus *transcriptome in response to glucose pulse*

Glucose carbon catabolite expression in *A. robustus* was also examined using a 5-mg glucose pulse to observe any changes in expression that occurred. Like the *N. californiae* experiment previously described, RNA was isolated from a set of cultures that were not pulsed as a control for the transcriptional response to changes in growth phases. Again, the fungal cultures showed regulation in many transcripts under both conditions.

Table 5.5. *Anaeromyces robustus* glucose pulse regulation summary

	# Transcripts Regulated	
Time (hrs)	Glucose Pulse	No Glucose Pulse
0.5	888	64
1.2	804	235
2	945	215
4	3741	86
6	3706	534
8	3827	486
24	1464	694
Total Unique Transcripts	6117	1533
Total Unique GH Transcripts	229	129
Up Regulated through 8 hrs	37	36
Down Regulated through 8 hrs	181	72

Anaeromyces robustus demonstrated regulation of many transcripts in response to the glucose pulse with nearly 1,000 transcripts showing significant regulation at each time point measured after the pulse. However, unlike *N. californiae* there was a significant difference between the pulsed cultures and the un-pulsed control with 6,117 transcripts regulated in the pulsed set and only 1,533 regulated in the control (Table 5.5). This suggests that there was significantly more regulation when the fungus needed to transition to growth on a different substrate, a result that was not surprising. When looking at the CAZymes only, there was regulation of 229 transcripts in the pulsed samples and 129 in the un-pulse control. In the pulsed case, only 37 of those CAZymes showed net up-regulation through the first eight hours after the pulse and 181 showed net down-regulation. The un-pulsed control showed up-regulation of 36 CAZyme transcripts and down-regulation of 72. This indicates a distinct increase in down-regulation of CAZymes in response to a glucose pulse, an observation that is clear from the heat maps in Figure 5.10.

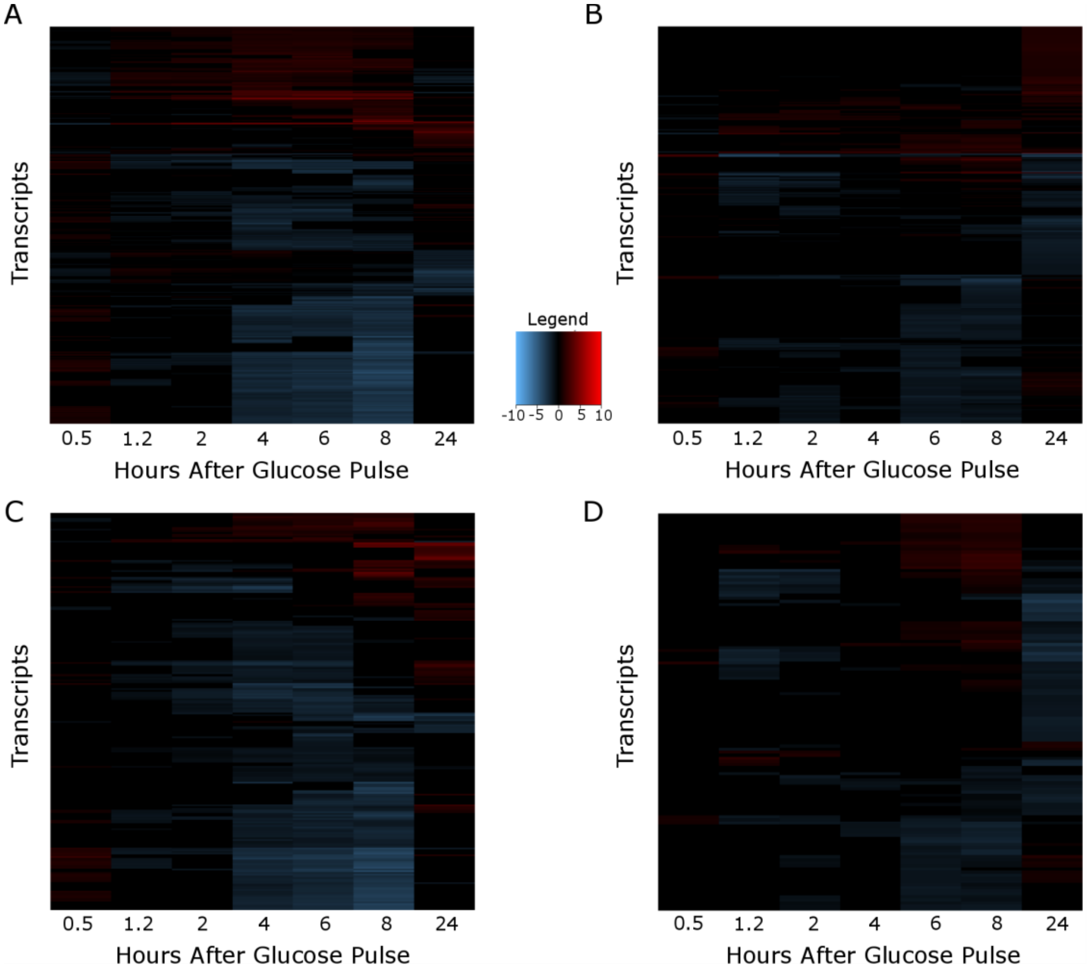


Figure 5.10. *Anaeromyces robustus* transcriptome response to glucose pulse

Anaeromyces robustus revealed regulation of many transcripts in response to a glucose pulse (A) as well as in the un-pulsed control (B), but regulated many more under the glucose pulse conditions. When examining CAZymes only, there was primarily down-regulation of CAZyme transcripts in response to the glucose pulse (C). In the un-pulsed control (D) there was significantly less down-regulation of CAZymes.

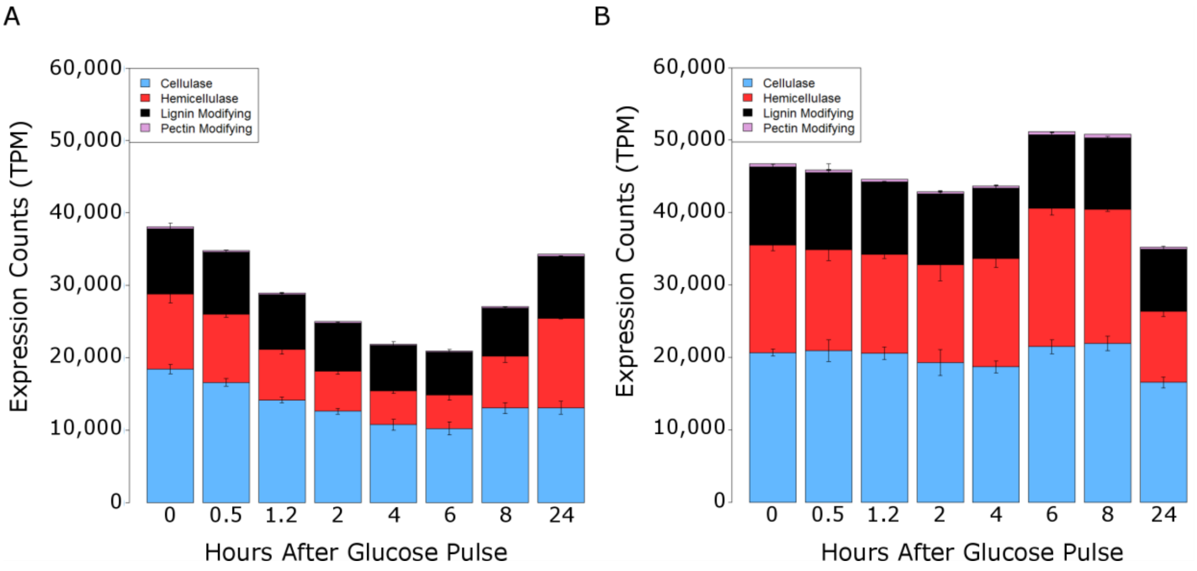


Figure 5.11. *Anaeromyces robustus* CAZyme expression after glucose pulse

A) Total expression levels of CAZymes after glucose pulse was added to growing culture of *A. robustus*. B) Total CAZyme expression in the control set of cultures that were not pulsed with glucose.

Examination of the normalized overall expression of CAZymes in TPM yielded a result similar to that of *Piromyces finnis*. After the glucose pulse was administered, the overall expression of cellulases, hemicellulases, and accessory enzymes each decreased. The expression of these enzyme classes reached a minimum expression level 6 hours after the pulse and the expression began to increase at the 8 hour and 24 hour time points, revealing a gradual return to the initial expression levels before the pulse. Measurement of the glucose concentration after the pulse (Figure 5.8.B) shows that most, but not all glucose is depleted after eight hours. This is likely the cause of the gradual increase seen in Figure 5.11; after eight hours, the glucose concentration has likely dropped enough to alleviate some of the repression, but not all. After 24 hours, the glucose level has returned to the basal level measured in the control cultures, and thus, the expression levels have nearly returned to the level before the pulse. The heatmap of CAZyme expression (Figure 5.10.C) also reveals a shift in the regulation at the 8 hour and 24 hour time points. At eight hours, fewer transcripts are down-regulated and a small subset showed increased expression compared to expression before the pulse. After 24 hours, almost all down-regulation is gone as the expression returns to the levels required for biomass degradation.

5.3. Conclusions

It is clear from these results that gut fungi tightly control gene expression in response to external stimuli and changes in their environment. Furthermore, carbon catabolite repression plays a significant role in transcriptional regulation, particularly in the case of carbohydrate active enzymes. Of the three isolates tested, only *Piromyces finnis* revealed distinct regulation clusters, or regulons, containing transcripts with conserved function. Several clusters containing almost exclusively hemicellulose and pectin degrading transcripts, or general biomass degrading transcripts were identified based on their regulatory patterns. Given this conservation of regulatory patterns among genes of similar function, we identified candidate genes for novel biomass degrading function. These transcripts were then aligned to the transcriptomes of *A. robustus* and *N. californiae* to find conserved gene sequences. These conserved sequences can now be expressed in heterologous hosts to examine their function.

Interestingly, the results from the same experiment carried out on *Neocallimastix californiae* and *Anaeromyces robustus,* did not provide the same opportunity to identify regulons containing transcripts of conserved predicted function for identification of novel enzymes. Likely due to the sheer number of regulated transcripts, cluster analysis did not provide insightful information for this purpose. In the case of *N. californiae* the regulation may not have been as clear due to the vast difference in the size of the genome compared to the smaller genomes of *A. robustus* and *P. finnis* (Table 2.4). However, the study still resulted in interesting findings. The regulation measured in *N. californiae* revealed that there was not a concerted, global effect on the total expression of CAZymes in response to a glucose pulse. This may indicate that the concentration of glucose added to the cultures was insufficient to trigger large scale regulation of biomass degrading enzymes, or that there are more complex regulatory mechanisms at work. While the overall CAZyme expression remained unchanged after the glucose pulse, there were a large number of enzymes that were downregulated in response to the pulse, but not in the control cultures. This suggests that rather than a net down-regulation of CAZymes, the glucose pulse triggered a reorganization of the CAZymes expressed.

Anaeromyces robustus, while not providing significant regulon information, did reveal the predicted response to the glucose pulse. Carbohydrate active enzymes within all classes of activity – including cellulases, hemicellulases, and accessory enzymes – were down regulated in response to the pulse. After the added glucose was depleted from the cultures, the expression returned to the level prior to the pulse. The results for *P. finnis* and *A. robustus* highlight the effect of carbon catabolite repression on the biomass degrading function of anaerobic gut fungi. In order to develop industrial processes using these enzymes it will be important to monitor the sugar concentration in cultures in order to optimize the production of enzymes and prevent reduction in biomass degrading activity during hydrolysis of complex substrates. Although *N. californiae* did not show the same level of global regulation of CAZymes in response to the glucose pulse, this suggests that the regulatory cues among distinct species are unique to one another. Similar to the substrate regulation results in Chapter

4, this means that it is valuable to study the regulation of each unique species in order to determine the proper process considerations for industrial application.

It is possible that differences in the sequencing analysis performed for these isolates compared to *P.finnis* is responsible for some of the disparity in the results. In the differential expression analysis for *P. finnis* sequencing was performed on an Illumina MiSeq and the Illumina TruSeq library preparation was performed following the provided protocol, including 15 PCR amplification cycles. However, in the sequencing for *A. robustus* and *N. californiae* the library preparation was modified to include fewer PCR cycles in an attempt to reduce the effect of amplification bias on the transcript quantification. While it is possible that amplification bias played a role in reducing the number of regulated transcripts in the experiment for *P. finnis*, it is also possible that the reduction in amplification cycles for the other two species allowed for more noise in the data, resulting in more transcripts identified as significantly regulated. Which of these two approaches yields the most significant information is still unclear. Furthermore, changes in the version of R programming packages used, specifically package "DESeq2" and its supporting statistical packages may have resulted in variations in the way the data is processed.

5.4. Materials and methods

5.4.1. Growth characterization of Piromyces

Growth of *P. finnis* on different substrates was measured through pressure accumulation of fermentation gases in the head space of sealed culture tubes[152]. A variety of substrates were tested including soluble sugars: glucose, cellobiose, xylose; cellulose: Avicel, SigmaCell, carboxymethyl cellulose (CMC); hemicellulose: xylan from Beechwood; and C3 and C4 grasses: reed canary grass, corn stover, alfalfa stems, and switchgrass. The exponential growth phase, the linear portion of the log-linear plot of the data, as used to calculated an effective net specific growth rate. These effective net specific growth rates were then used to compare growth of the fungus across various substrates.

5.4.2. RNA isolation

RNA was isolated from cultures during exponential growth (P~3-8 psig) using a Qiagen RNEasy Mini Kit (Qiagen, Valencia, CA) following the manufacturer's instructions for Plants and Fungi. Sample quality was assessed by RIN score with a BioAnalyzer (Agilent Technologies, Santa Clara, CA). For the *Piromyces finnis de novo* transcriptome assembly, RNA samples from cultures grown on glucose and reed canary grass were prepared.

5.4.3. Piromyces *transcriptome acquisition*

Pooled libraries were normalized and denatured using 0.2 N NaOH prior to sequencing. Flowcell cluster amplification and sequencing were performed according to the manufacturer's protocols using the HiSeq 2500. Each run was a 6bp paired-end with an eight-base index barcode read. Data was analyzed using the Broad Institute Picard Pipeline which includes de-multiplexing and data aggregation. In total, more than 10^8 reads were acquired. The reads were then assembled into a *de novo* transcriptome of more than 27,000 transcripts with an average sequence depth of 400x using Trinity (r2013-02-25)[122]. For subsequent differential expression experiments, cDNA libraries were sequenced using a MiSeq (Illumina, San Diego, CA). Transcripts were grouped into gene families as determined by their component and subcomponent (compXX_c##) grouping within the Trinity platform. Reads from all conditions tested were aligned to the *de novo* transcriptome and expression was estimated using RSEM analysis[123].

5.4.4. Glucose perturbation experiments for Piromyces

Piromyces finnis was grown in parallel 10 mL cultures to mid log phase on reed canary grass (~2 days) before they were pulsed with 5mg of glucose. Four cultures were set aside as an untreated control prior to sugar addition and harvested for RNA and transcriptome quantification. After pulse, samples were taken at various time intervals (20 minutes, 40 minutes, 1 h, 3.5 h, 7 h, and 28 h) until all the glucose was consumed. For each time point, 3-4 tubes were sacrificed and the RNA isolated for transcriptome quantification. Glucose levels were tracked by assaying the culture supernatant of each culture tube with a glucose hexokinase-based assay (Megazyme, Bray, Ireland). The remaining supernatant was reserved at -80°C until cDNA prep and analysis.

5.4.5. Glucose perturbation experiments for Anaeromyces *and* Neocallimastix

Anaeromyces robustus and *Neocallimastix californiae* were each grown in parallel 10 mL culture to mid-log phase on reed canary grass. After reaching mid-log growth, half of the cultures were pulsed with 5 mg of glucose and half were not pulsed. After the pulse, samples were taken at various time intervals; 0.5, 1.2, 2, 4, 6, 8, and 24 hours for *Anaeromyces* and 0.6, 1.1, 2, 4, 6, 8, and 24 for *Neocallimastix*. For each time point 3-4 cultures tubes were sacrificed for RNA isolation from both the pulsed and un-pulsed sets of cultures. Glucosse levels were also tracked by assaying culture supernatants using a YSI 2900 substrate analyzer with YSI 2365 glucose detection membrane kits (YSI Inc., Yellow Springs, OH). RNA isolated from each culture was sequenced on an Illumina HiSeq and the reads were aligned to the *de novo* assembled transcriptomes (described in Chapter 2) and expression was estimated using RSEM analysis[123].

5.4.6. Substrate RNA profiling for Piromyces

Triplicate 10 mL anaerobic cultures were grown to mid-exponential phase (P ~5psig) on glucose, cellobiose, Avicel, Whatman #1 filter paper, and reed canary grass. Glucose and cellobiose were included at a concentration of 5 g/L, reed canary grass and avicel were included at 10 g/L, and Whatman filter paper was included as a ~1 cm square. When the cultures reached mid-log growth, RNA was isolated.

5.4.7. Differential Expression Analysis and expression clustering

Differential expression was determined using estimated count data determined by the RSEM algorithm and the Bioconductor DESeq2 package in the R programming language with default parameters[126]. Results were filtered for statistical significance using an adjusted p-value ≤ 0.01 and a $|\log_2\text{-fold change}| \geq 1$. Expression data was then clustered using complete hierarchical clustering based on a Pearson correlation distance metric (1-r) of the $\log_2$-fold changes. Clusters were defined at $h = 0.5$ to form the 21 regulons. Conserved functionalities were assigned to the clusters based on the most frequently occurring functions as determined by protein domain, or BLAST hit if no protein domain information was available.

5.4.8. Gene Set Enrichment Analysis

Enrichment for up- or down-regulation of specified gene sets was computed using the GSEA Preranked tool in GSEA v2.0.14[210] against a ranked list of genes. Ranking was based on the $\log_2$-fold change compared to glucose as determined by DESeq2. Gene sets between 15 and 500 members were specified based on predicted protein domains or presence in a regulon from the glucose perturbation study. Statistical significance was estimated from 1000 permutations of the dataset gene names.

5.4.9. Lignin content analysis

Analysis of lignin content was performed by growing gut fungal cultures in 60-mL serum bottles containing 2.0 grams of switchgrass or 0.2 g cellobiose. Cultures containing switchgrass were grown on minimal media (M2 media)[212] containing no rumen fluid and cultures containing cellobiose were grown on complex Medium C. Cultures were grown for one week, transferred into a 50-mL Falcon tube, and frozen at -80°C before they were shipped to the Foston Lab at Washington University in St. Louis. Molecular analyses to determine lignin content and phenol concentration were performed by Marcus Foston and James Meyer at WUStL.

6. Conclusions

6.1. Perspectives

6.1.1. Potential of non-model microbes for lignocellulose bioprocessing

Development of a sustainable, bio-based economy requires more efficient methods for the breakdown of non-food, lignocellulosic biomass into sugars that can be fed to microorganisms for production of desired fuels and chemicals[12,34]. This challenge can be addressed by turning to nature and studying the natural ecosystems in which lignocellulosic biomass is degraded. In the past, this approach has led to the study of aerobic fungi for their variety of cellulose degrading enzymes, particularly from *Trichoderma reesei*[206], as well as lignin degrading enzymes, in the case of white rot fungi[213]. It also brought on the study of anaerobic bacteria known to form extracellular complexes of cellulases called cellulosomes for efficient cellulose hydrolysis[214]. Compared to these microorganisms, the anaerobic gut fungi within the Neocallimastigomycota division, are understudied resources for the decomposition of complex lignocellulosic biomass. While gut fungi have long been studied for their role in agriculture and animal health[69,71,208], it was not until recently that they have been studied in the context of industrial bioprocessing[10,11,66,130]. This has been in part due to the difficulty of isolation and culture of these organisms, and challenges associated with their molecular and genomic study.

Development of new techniques for isolation, culture, and molecular analysis will no doubt increase the study of unexplored organisms like the anaerobic gut fungi. The rapid development of a variety of new sequencing technologies has already enabled a more in depth study of microorganisms that are difficult to isolate from their native ecosystem. As the cost of sequencing is reduced, the ability to sequence microbial ecosystems at a depth that allows detailed analysis of the full metagenome becomes possible. This allows for the discovery of key organisms and microbial communities that play an important role in biomass breakdown. Improved methods for the targeted enrichment of consortia capable of efficient biomass degradation will also lead to the identification of previously "unculturable" microbes that may represent a small percentage of the community, but play an important role in community[215]. As is the case with the anaerobic gut fungi, organisms that make up a small proportion of the full microbial community can fill an important niche that might otherwise be overlooked[65,66].

In addition to challenges in the culture and isolation of these types of organisms, they also typically lack genetic tools for functional modification. This imposes a limitation on their industrial use as they cannot be engineered for production of value-added fuels and chemicals. However, newly identified genome editing tools, such as CRISPR/Cas9 and TALENs, present an opportunity to address this need[216]. An alternative approach also focuses on the use of consortia to compartmentalize different necessary functions to organisms uniquely suited for specific roles, potentially eliminating the need for the engineering of non-model organisms.

6.1.2. Microbial consortia as improved biomass degrading systems

A shift in focus to the study of microbial consortia, and how the organisms within them work together to achieve specific goals, is a research area with great potential to drive the efficiency of lignocellulose hydrolysis[58]. Focusing on the isolation of consortia or "minimal systems" of organisms rather than isolated microbes can lead to improved biomass degradation, as observed in the culture of anaerobic gut fungi with archaeal methanogens that increase the overall biomass degradation performed by the gut fungi[9,67,127]. It is likely that there are additional, beneficial relationships within the rumen microbiome, as well as other microbial ecosystems, that may be used to enhance the native function of individual microbes.

Synthetic consortia and microbiomes also present a valuable opportunity to combine the best traits of different microbes in a single system, rather than engineer one model microorganism with all traits necessary to achieve a specific goal. We have demonstrated some success with this approach as discussed in Chapter 4, using gut fungi to degrade biomass and a model yeast, *Saccharomyces cerevisiae* for production. Other work in this area has combined the complex biomass degradation ability of aerobic fungi with the ease of engineering of bacteria and yeasts[60,61]. Such synthetic systems have many advantages, but also present challenges, particularly in the formation of a stable culture system, ensuring that the members of the consortia rely on each other to survive. To ensure stability of the consortia, the microbial members must be designed to require the presence of the other members. This can be done through the careful selection of organisms to create a nutritional mutualism, where both species benefit from each other, or commensalism, where one organisms depends on the other, in the community. The potential of these microbial consortia is clear, but developing a systematic way to engineer them remains a challenge.

6.2. Future directions

Anaerobic gut fungi have demonstrated potential for use in industrial bioprocessing with their incredible ability to degrade complex biomass without any pretreatments. In this work, they have also proven an ability to supply fermentable sugars from biomass to engineered model microorganisms in a two-stage fermentation scheme. To further advance their use in industrial biotechnology, an important next step is to engineer stable co-cultures of anaerobic gut fungi and model microbes, such as *Saccharomyces cerevisiae*. While the two-stage fermentation consolidates a process that typically requires three steps for production of chemicals from biomass, stable co-cultures will enable a one-pot system for chemical production directly from crude biomass, further consolidating the overall process. In order to do so, organisms that have little overlap in sugar metabolism with the gut fungi present an opportunity to host multiple organisms that grow in parallel. Furthermore, engineering a system where the second microbe requires additional nutrients produced by the gut fungi will also promote overall culture stability. This can be done be genetically engineering auxotrophy for specific amino acids or other nutrients that are produced by the gut fungi. These approaches will lead to the development of designer consortia that are well suited for both biomass degradation and chemical production.

In addition to design of consortia, development of genetic tools for the engineering of anaerobic gut fungi will also elevate their application to industrial bio-based chemical production. Using a combination of transcriptomic and genomic sequencing, we have identified putative promoters for the controlled expression of gut fungal genes. If gut fungi can be engineered to produce heterologous genes, such as flavin-based fluorescent proteins, the function of these predicted promoters can be verified and a minimal promoter regions can be identified. Based on the data obtained through transcriptomic experiments, there are candidates for both induced and constitutive promoters that have the potential to be valuable resources for engineering production of heterologous proteins and pathways.

Using functional annotations of transcriptomic and genomic sequences, we characterized cellular metabolism including glycolysis, pentose phosphate, and other sugar catabolic pathways as well as amino acid synthesis pathways. Differential expression data can also be used to predict how the fluxes through these pathways may change under different growth conditions. To complement the data that has already been obtained, metabolomic characterization is necessary to complete the cellular metabolic pathways. Metabolomics will allow for the identification of the most prevalent metabolic end points, and confirm the presence or absence of pathways that were identified as incomplete based on transcriptomic and genomic data. The incorporation of metabolomics will lead to the development of a complete metabolic model that can be used to design metabolic engineering approaches. Furthermore, based on the complete model additional organisms can be chosen for designer consortia that take gut fungal metabolic outputs and produce value-added chemicals. With these next steps, anaerobic gut fungi will be brought closer to industrial application.

6.3. Overall conclusions

The breakdown of lignocellulosic biomass is a complex problem, but we have shown that gut fungi present a valuable resource for developing more efficient methods for consolidated bioprocessing. We isolated, characterized, and classified several unique strains of anaerobic gut fungi (*Neocallimastix californiae*, *Anaeromyces robustus*, *Piromyces finnis*, and *Caecomyces churrovis*) whose growth was not limited on complex biomass substrates compared to simple monosaccharides. These fungi possess a wide array of enzymes that allow for the efficient hydrolysis of crude plant material without the pretreatment that other enzymatic hydrolysis methods require. In fact, compared to other members of the fungal kingdom, whose members are largely known for their ability to degrade biomass, gut fungal transcriptomes contain many more genes encoding for biomass degrading function[130]. We characterized the biomass degrading enzyme repertoire through the use of transcriptome and genome sequencing along with bioinformatic analysis to predict the function of genes based on their sequence similarity to genes of known function.

By studying the global regulation of the transcriptome within different strains of anaerobic gut fungi we gained even greater insight into their biomass degrading function. We identified the control of expression of biomass degrading enzymes in response to growth on substrates of varying complexity. This work highlighted the necessary conditions for expression of the entire suite of biomass degrading enzymes in both *N. californiae* and *A. robustus*. In the case of *A. robustus*, all necessary enzymes are produced during growth on cellobiose, a simple disaccharide that is a breakout product of the enzymatic digestion of cellulose. This suggested that a gut fungus derived enzymatic cocktail for purely enzymatic digestion of crude biomass, can be grown on a soluble sugar source, simplifying the required purification processes to obtain the full enzymatic cocktail. Conversely, *N. californiae* only produced the full suite of cellulases, hemicellulases, and accessory enzymes during growth on crude biomass, indicating that *A. robustus*is the better candidate for enzyme production and purification strategies. Regulation triggered by introduction of a carbon catabolite repressor, glucose, also identified conserved regulation patterns among biomass degrading enzymes in *P. finnis* that may be used to identify novel genes for putative biomass degrading function. Carbon catabolite repression studies also highlighted the importance of sugar concentration as a key regulator of biomass degrading enzyme production and as a culture condition that must be considered in the optimization of biomass degrading activity.

Gut fungi have demonstrated that they are a valuable resource for enzymes that may be used to create better biomass degrading enzyme cocktails for use in lignocellulose hydrolysis, but they also demonstrated an ability to release excess sugars during growth in batch cultures. This discovery was leveraged to create a two-stage fermentation scheme whereby gut fungi are used to consolidate the pretreatment and hydrolysis steps of typical bio-based production processes and supply fermentable sugars to model microorganisms directly from crude biomass. In this scenario, rather than the cost of expensive chemicals or energy intensive operating conditions, there is an opportunity cost in the amount of sugar necessary to support growth of the gut fungi as they hydrolyze the biomass. There is still room to optimize this production scheme using careful selection of model microbe partners to take advantage of all sugars hydrolyzed from crude biomass.

We have shown the potential for gut fungi in industrial bio-based processes, whether that be through the use of gut fungi to consolidate pretreatment and hydrolysis, or gut fungal enzymes that can improve upon existing enzymatic hydrolysis cocktails. However, there are necessary improvements required to implement these unique organisms in production pipelines. We have developed simple and reliable methods for cryogenic storage, a crucial step towards developing and maintaining a production strain, but more detailed metabolic models and genetic tools are the remaining pieces that will make gut fungi a valuable industrial resource. Full genome scale metabolic models are necessary for pathway engineering to introduce new functionalities into the organism, and can also lead to the development of designer consortia that leverage the ability of fungi to degrade biomass. Genetic tools, on the other hand, will allow for development of production strains of gut fungi that are capable of both hydrolyzing crude biomass and producing a valuable chemical. While there are many challenges ahead for the incorporation of gut fungi into industrial bioprocesses, here we have made great strides to understanding where they may be best applied and how to develop them for industrial use.

7. Appendices

7.1. Appendix A: Substrate differential expression analysis in *Pycnoporus cinnabarinus*

Production of chemicals and fuels from biomass is critical for a renewable economy. While cellulose and hemicellulose have been valuable sugar feedstocks for microbial production[2,23], lignin has proven difficult to convert into valuable products. Lignin valorization aims to convert lignin into renewable chemical feedstocks or biofuels[217-219], but lignin's heterogeneous nature often results in complex product distributions upon depolymerization. Incorporating biological lignin depolymerization processes into valorization efforts is a promising opportunity in making lignin a commercially viable renewable chemical feedstock[220]. An in depth understanding of biological lignin breakdown could help streamline bio-reactors and provide inspiration for biomimetic approaches to lignin depolymerization.

White rot fungi are capable of degrading woody plant materials, breaking down lignin to leave soft, white, rotted wood behind. The common mechanisms by which white rot fungi accomplish this is through the use of lignin peroxidase, manganese peroxidase, versatile peroxidase, and laccase enzymes to oxidatively degrade lignin[207,221]. The fungus *Pycnoporus cinnabarinus* was once thought to only generate laccase enzymes to facilitate lignin breakdown[222,223]. However, the presence of lignin peroxidases has recently been detected in extracellular assays and the genome of *P. cinnabarinus* BRFM137, which contained the genes necessary to express lignin peroxidases as well as their support enzymes[224,225]. Based on these new insights, further exploration of *P. cinnabarinus* PB 94's role in lignin disassembly was undertaken through differential expression analysis[226,227].

Next generation sequencing (NGS) is a powerful tool for understanding biological processes related to lignin breakdown through transcriptomic analysis[228-230]. In particular, global analysis of gene expression under different physiological conditions can give detailed insights into the enzymatic profile of lignin-degrading fungi in response to lignin-rich substrates[231]. Transcriptome analysis has previously been employed to study the closely related species of white rot fungi, *Pycnoporus coccineus,* detailing the metabolic changes that occur between hard and soft wood lignin degradation[232]. We have used RNAseq to explore the changes in lignin-degrading capabilities of *P. cinnabarinus* PB 94 comparing expression during growth on sugar and biomass of varying lignin content. Differential expression analysis reveals genes that are significantly regulated in response to growth on lignin-rich substrates and presents a picture of the metabolic changes that occur in response to growth on lignin. Close attention was paid to the regulation of carbohydrate active enzymes (CAZymes)[136], particularly members of the auxiliary activity (AA) family containing enzymes responsible for lignin degradation[233].

7.1.1. Results and Discussion

***de novo* Assembly of *Pycnoporus cinnabarinus* PB 94 Transcriptome yields large exome**

Initially, the exome of *Pycnoporus cinnabarinus* BRFM137 was used as a reference for alignment and abundance estimation using RSEM analysis[224]. However, alignment to BRFM137's genome was poor and resulted in a significant number of genes with no aligned reads. Therefore, the transcriptome of PB 94 was assembled *de novo* using RNA isolated from cultures grown on five substrates: glucose (G), maltose (M), cellobiose (CB), Poplar (Po), and switchgrass (SG), as well as a control (C) culture with no substrate (Table 7.1). Cultures grown on soluble substrates yielded high quality RNA as determined by Agilent TapeStation RINe scores of 7.5 or greater while cultures grown on biomass, Po and SG, demonstrated isolation of lower quality RNA, yielding partially degraded RNA with RINe scores between 5.5 and 6.6. The RNA degradation observed on biomass substrates, which was not observed on soluble substrates, may be due to the generation of radical species by the lignin degrading-enzymes expressed under these conditions, resulting in lower stability of RNA. To accommodate the partially degraded RNA samples, ribosomal depletion, rather than poly-A enrichment was used to remove ribosomal RNA from Po and SG cultures[234]. Poly-A enrichment was used to isolate mRNA from samples with high RINe scores.

Table 7.1. Growth conditions for transcriptome analysis

Substrate	Harvested	Concentration	Stored*
Control	5 days	NA	Yes
Glucose	5 days	20 g/L	Yes
Maltose	5 days	20 g/L	Yes
Cellobiose	9 days	20 g/L	Yes
Poplar	12 days	2 g	No
Switchgrass	12 days	2 g	No

* Cultures stored in 1 mL of RNAlater at -80 °C after harvesting cells

The *de novo* assembly used approximately 30 million reads from the five different growth conditions for equal representation of each substrate. This resulted in a transcriptome comprised of 45,286 transcripts, including isoforms, of a predicted 27,990 genes (Table 7.2), much larger than the previous prediction of 10,442 genes in BRFM137[224]. While some of these additional transcripts may have been artifacts of the alignment, it also highlighted the importance of transcriptomes for gene prediction. The use of transcriptomes in gene prediction may result in the identification of additional genes not identified by alternate gene models, particularly for organisms which have not been well characterized by experimental data[235]. Therefore, the *de novo* transcriptome was used in subsequent analyses.

Table 7.2. Transcriptome assembly statistics

Transcriptome Assembly	
# Transcripts	45,286
# Predicted Genes	27,990
Transcriptome Size	57,410,976
# Reads	178,409,214
Read Length (bp)	75
Coverage	233.1
Conditions Used	
Substrate	**#Reads**
Control (no substrate)	30,074,212
Glucose	28,959,734
Maltose	30,000,002
Cellobiose	30,031,750
Poplar	32,452,406
Switchgrass	26,891,110

Extensive global regulation across substrates in *P cinnabarinus* PB 94

Differential expression analysis compared gene expression counts on substrates of differing lignin content to the control culture. Poplar and switchgrass were chosen for their large difference in lignin concentration by dry weight of ~25% and ~17%, respectively[236]. The heat map in Figure 7.1 reveals a total of 3575 transcripts were significantly regulated when the four substrates are compared to the control. Of these regulated transcripts, 2519 were not assigned any gene ontology (GO) terms by the BLAST2GO annotation pipeline. The highest represented GO terms in each of the three GO classes (biological process, molecular function, and cellular component) were involved in translation (76 transcripts) and transmembrane transport (50 transcripts) found within the biological process class. Within the molecular function ontology notable annotations included 88 transcripts with oxidoreductase activity, 53 heme binding, 9 substrate-specific transmembrane transporter activity, and 9 carbohydrate/cellulose binding. It was expected that many housekeeping genes, such as those involved in translation, would be regulated, but not such significant regulation in transmembrane transporter proteins. This suggests that a great deal of transcriptional level regulation is involved in the tailoring of the cell membrane to handle different metabolic inputs presented by each substrate as well as the export of proteins from the cell to act on biomass.

The largest changes in regulation were seen for the biomass cultures with 2556 regulated transcripts on poplar and 1933 regulated transcripts on switchgrass. Cellobiose, a β(1,4) glucose dimer and a breakout product of cellulose deconstruction, had relatively fewer regulated transcripts with 398. Maltose, α(1,4) glucose dimer, had the least observed regulation with only 89 regulated transcripts. The low level of regulation observed for maltose was likely a result of the presence of maltose in the agar of the maintenance plate used to inoculate all cultures, including the control culture, that may be masking the regulation of maltose metabolism. The distribution of upregulated transcripts (Figure 7.1) shows considerable overlap between poplar and switchgrass as well as the largest subset of unique transcripts demonstrating significant regulation. The larger transcript profile is likely due to growth on more challenging substrates, requiring a wider array of carbohydrate active enzymes (CAZy) as well as changes in metabolism necessary for the wider array of metabolites provided from biomass.

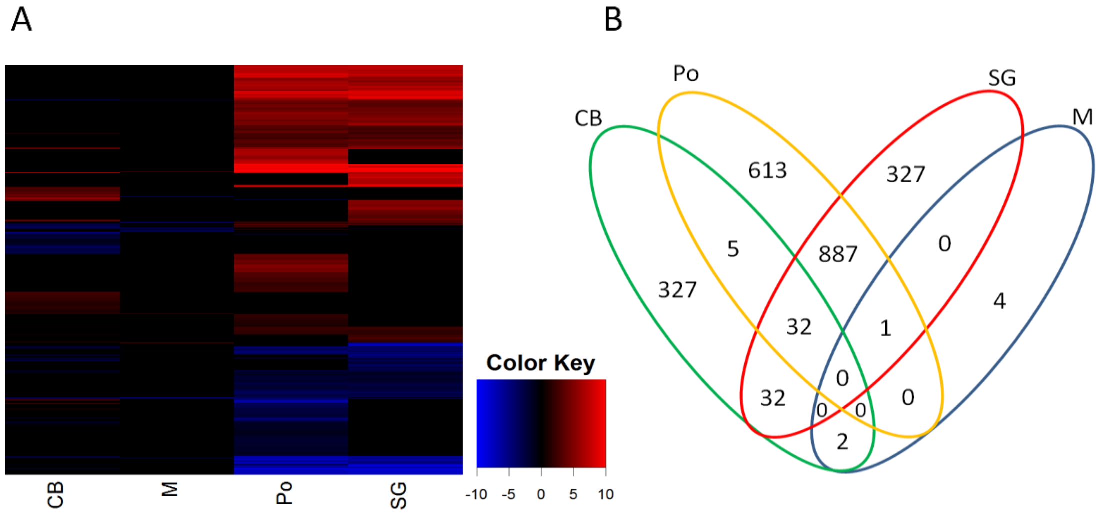


Figure 7.1. Transcriptional regulation of different substrates compared to control

A. Heat map of regulated transcripts in *P. cinnabarinus* PB 94 as compared to the control. Red indicates significant up regulation while blue show significant down regulation. Where CB is cellobiose, M is maltose, Po is Poplar wood shavings, and SG is milled switchgrass. B. Venn diagram of upregulated transcripts of *P. cinnabarinus* PB 94 grown on different carbon sources. CB is cellobiose, Po is Polar, SG is switchgrass, and M is maltose.

Expression of Auxiliary Activity Transcripts greatly increases on lignocellulose

Of the five distinct CAZy classes (Glycoside Hydrolases, Glycosyl Transferases, Polysaccharide Lyases, Carbon Esterases, and Auxiliary Activities), the Auxiliary Activities (AA) family that contains enzymes directly or indirectly involved in lignin decomposition, were of primary interest. The AA family can be further broken down into subclasses, two of which (AA1 and AA2) are comprised of enzymes known to degrade lignin. The AA1 subclass contains laccase, ferroxidase, and a laccase-like multi-copper oxidase. *P. cinnabarinus* is well known for expression of laccases, which were once thought to be the only lignin-degrading enzyme produced by *P. cinnabarinus*[222,223,237,238]. However, through recent extracellular assays and genomic analyses, *P. cinnabarinus* has been shown to also express members of the

AA2 subclass[224,225,239,240], including manganese peroxidase (MnP), lignin peroxidase (LiP), and versatile peroxidase (VP). Manganese peroxidase generates Mn(III), chelated with small organic acids, from the oxidation of Mn(II)[241,242]. The Mn(III) complex then diffuses through lignin, generating phenoxy radicals at terminal phenols[243]. Lignin peroxidase directly oxidizes the aryl rings contained in lignin or aryl rings in small molecules that then serve as redox mediators to oxidize lignin[244]. Versatile peroxidase has the activity of both MnP and LiP[245]. MnP, VP and LiP catalytic cycles are initiated with peroxides that are provided in part by another subclass of AAs, AA3_3. The alcohol oxidases (GO:0047639) that make up AA3_3 catalyze the reaction of a primary alcohol and O_2 to an aldehyde and H_2O_2; this hydrogen peroxide then feeds into the AA2 catalytic cycles.

P. cinnabarinus PB 94 demonstrated substrate based control of expression for several auxiliary activity gene families. There was a marked increase in transcript expression counts for the AA3 and AA2 families when *P. cinnabarinus* PB 94 was grown on the lignin-rich substrates of poplar and switchgrass (Figure 7.2,Table 7.3) demonstrating a tight control such that these enzymes are expressed in high quantities only when necessary. Laccase (AA1) expression did not increase with increasing lignin content and in general exhibited steady basal expression where any apparent increase is within the standard error (Figure 7.2), an observation consistent with previous studies of *P. cinnabarinus*[223,224], except in the case of cellobiose. Interestingly, a single laccase gene (TR10024|c0_g1_i1) was responsible for this increased expression on cellobiose in all three biological replicates. Cellobiose is not known to cause up-regulation of laccases so this may represent the first reported case of this effect. If this gene is regulated by cellobiose, then the native regulation likely relies on small amounts of cellobiose released from biomass to trigger expression. Growth on cellobiose alone provides a much higher concentration of the substrate than would be seen during the degradation of cellulose from biomass. Therefore, growth on cellobiose may result in an inflated response compared to growth on biomass substrates where cellobiose may also be present, but in lower concentrations.

Table 7.3. Log_2-fold change of lignin active enzymes compared to control

Transcript	CB	Po	SG	Gene description
TR3772|c0_g1_i1	2.8	3	3.8	Laccase
TR10024|c0_g1_i1	4.1			Laccase
TR1596|c0_g1_i3			7.7	Laccase
TR16670|c0_g2_i1	2.7		3.6	Manganese Peroxidase
TR9691|c0_g2_i1			7.5	Manganese Peroxidase
TR17164|c0_g1_i1			7.9	Manganese peroxidase 1 precursor
TR11892|c0_g1_i3		8.6		Manganese-dependent peroxidase
TR11892|c0_g1_i1		7.8	9.4	Manganese-dependent peroxidase
TR9691|c0_g1_i1			5.6	Manganese-dependent peroxidase
TR18856|c0_g1_i4		3.5	3.7	Possible laccase
TR18856|c0_g1_i1		3.4	3.7	Possible laccase
TR14782|c0_g1_i1		14.2		Prepropeptide lignin peroxidase
TR12424|c2_g2_i1		13.1	9.9	Prepropeptide lignin peroxidase
TR1930|c0_g1_i1		11.9	8.4	Prepropeptide lignin peroxidase
TR15946|c0_g1_i1		11.1		Prepropeptide lignin peroxidase
TR886|c0_g1_i1		9.0	7.1	Prepropeptide lignin peroxidase
TR11254|c0_g2_i1		7.8	6.5	Prepropeptide lignin peroxidase
TR18900|c1_g1_i1		6.7	6.8	Prepropeptide lignin peroxidase
TR16664|c0_g1_i		5.0		Prepropeptide lignin peroxidase
TR11254|c0_g2_i2			7.1	Prepropeptide lignin peroxidase

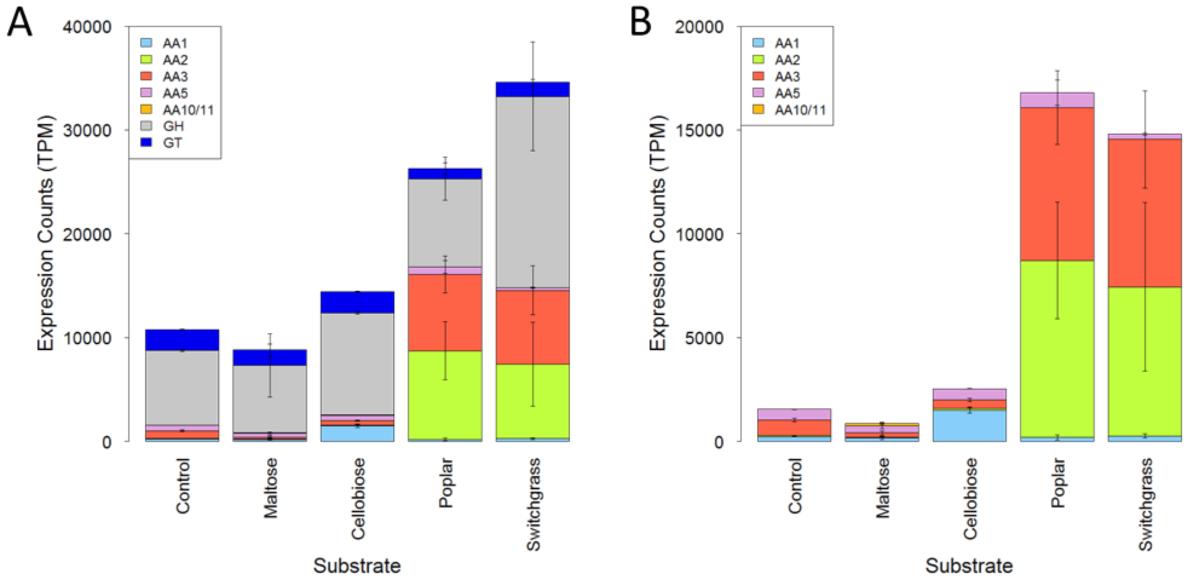


Figure 7.2. Total expression of CAZymes as a function of substrate

A. The CAZy profile for *P. cinnabarinus* PB 94 transcriptome showing an increase in the expression of AA2 and AA3 families when the fungi is grown on lignin-rich substrates. B. CAZy auxiliary activities enzyme profile highlighting the dramatic increase of AA2 and AA3 with lignin-rich substrates.

Within the AA1 and AA2 subfamilies three laccases, five manganese peroxidases, two possible laccases, and nine lignin peroxidases were upregulated compared to the control (Table 7.3). No versatile peroxidase was observed in the transcriptome, although they have been identified in other strains of *P. cinnabarinus*[224]. Interestingly, more MnP genes were upregulated on switchgrass and more lignin peroxidase genes are upregulated on poplar. This difference in expression between poplar and switchgrass may be due to the higher lignin content in poplar compared to switchgrass[246-248]. However, both LiP and MnP show increased expression in poplar and switchgrass as compared to cellobiose and maltose.

Table 7.4. Log$_2$-fold change in regulated alcohol oxidases compared to control

Transcript	Po	SG
TR8734|c0_g1_i1	9.8	9.2
TR13643|c4_g2_i1	8.2	7.1
TR8734|c0_g2_i1	7.9	7.1
TR13643|c4_g1_i1	7.5	6.5
TR17454|c0_g1_i2	7.5	6.7
TR9467|c2_g1_i1	7.0	6.7
TR13643|c1_g1_i1	6.8	5.8
TR17454|c0_g1_i4	6.4	
TR17454|c0_g1_i5	6.3	
TR3037|c1_g1_i3	4.9	4.5
TR3040|c0_g1_i1	4.9	5.1
TR13838|c0_g1_i3	4.5	
TR3745|c1_g1_i1	4.4	4.8
TR13838|c0_g1_i1	4.1	
TR19162|c0_g1_i1	3.6	
TR3745|c1_g1_i2	3.4	
TR3037|c0_g1_i1		7.6
TR3037|c0_g2_i2		7.4

A total of 72 diffeent alcohol oxidases were found in the *de novo* transcriptome of *P. cinnabarinus* PB 94. A subset of 18 alcohol oxidases was upregulated when *P. cinnabarinus* PB 94 was grown on lignin-rich substrates (Table 7.4). Much of the observed expression comes from three alcohol oxidase transcripts: TR8734|c0_g1_i1, TR13643|c4_g1_i1, and TR8734|c0_g2_i1, with expression levels of approximately 600, 2000, and 2000 TPM, respectively. These three alcohol oxidases, present in the AA3 auxiliary activity family, only demonstrated increased expression when grown on lignin-rich biomass compared to other substrates. As a whole, the AA3 family demonstrated a drastic increase in expression when grown on biomass, revealing that expression of these genes was dictated by necessity from growth on more complex substrates.

Gene set enrichment analysis revealed that two AA families were significantly enriched on biomass substrates (Figure 7.3) when expression was compared to the control culture. Both AA2 and AA3 families were enriched on switchgrass while only family AA2 was enriched on poplar. It was surprising that both groups were not enriched on poplar, which has higher lignin content and therefore a larger requirement for the activities of these enzymes. However, when the same data sets were compared to expression on maltose, both switchgrass and poplar demonstrated enrichment of both AA2 and AA3 transcripts. This discrepancy is likely due to the large size of the AA3 gene set. Larger gene sets are less likely to be termed enriched in GSEA if only a few members are responsible for the bulk of the expression change. Although hydrogen peroxide producers within this class would likely show enrichment under these conditions, the AA3 family is a diverse family and not all members are necessarily involved in lignin degradation.

Additional sources of hydrogen peroxide required for oxidase activity are glyoxal oxidases (AA5_1) and GMC oxidoreductases (AA3_2)[249,250]. Four glyoxal oxidases and 11 glyoxal oxidase precursors were present in the transcriptome, but only the precursors displayed significant regulation. Out of the four glyoxal oxidases, only TR15336|c0_g2_i1 displayed significant expression on the order of hundreds of TPM counts while the remaining three had much lower expression counts with only 1 to 10 TPM. There were 26 GMC oxidoreductase transcripts identified in the transcriptome that may also contribute to peroxide generation. Transcript counts were typically low for the GMC oxidoreductases, with counts less than 10 TPM except for TR11006|c1_g1_i1 and TR11006|c2_g1_i1, which had ~500-1000 TPM on maltose, cellobiose, and the control. Interestingly, these two transcripts were down regulated on poplar and switchgrass compared to the control despite their activity being more important with higher lignin content. Two GMC oxidoreductases, TR15260|c0_g1_i1 and TR15260|c0_g2_i1, had low basal expression levels with counts of approximately 30 TPM across all substrate types.

Differential expression and GSEA reveal a clear importance of AA2 and AA3 enzymes in enabling growth of *P. cinnabarinus* on biomass substrates. While the expression of laccases remained largely unchanged on biomass substrates, these enzymes may be responsible for beginning the breakdown of biomass and releasing molecules that trigger increased expression of additional lignin active enzymes. The enrichment of alcohol oxidases on lignin rich substrates suggests that the expression of alcohol oxidases involved in lignin break down is tuned to the substrate. Low expression levels of glyoxal oxidases and GMC oxidoreductases along with down regulation of some transcripts on biomass, suggests that they are not critical for biomass degradation, but may play a small role in supplying hydrogen peroxide for lignin deconstruction. These analyses also revealed significant expression changes of other CAZY families associated with cellulose and hemicellulose breakdown.

Glycoside Hydrolases demonstrate significant regulation across conditions

The glycoside hydrolase (GH) family displays the highest level of expression under all conditions (Figure 2A) with 535 transcripts (Table 7.5). The AAs make up the second largest family expressed and a considerable subset were upregulated in the presence of lignin-rich substrates. The third largest representative from the CAZy family is the glycosyl transferases (GTs). Fewer than 100 representatives of the carbohydrate esterase (CE) and polysaccharide lyase (PL) transcripts are observed. The high expression levels of the GHs and GTs is consistent with the literature of many fungal species, as they represent the primary means of carbohydrate metabolism, though the white rot fungi are well known for their auxiliary activities that degrade lignin[213,251].

Table 7.5. Carbohydrate active enzyme genes present in transcriptome and their regulation

CAZy designation	**# of Transcripts**	**Po regulation**		**SG regulation**		**CB regulation**	
		Up	**Down**	**Up**	**Down**	**Up**	**Down**
Glycoside Hydrolase (GH)	535	116	2	91	7	43	27
Auxiliary Activities (AA)	194	32	0	29	2	3	3
Glycosyl Transferase (GT)	157	6	3	2	1	4	0
Carbohydrate Esterase (CE)	75	20	1	17	0	1	1
Polysaccharide Lyase (PL)	5	2	0	1	0	0	0

Po: Poplar, SG: Switchgrass, CB: Cellobiose

The total GH expression levels remained consistent between substrate types (Figure 7.2) with the exception of SG, which had a statistically significant increase in GH expression. While the total expression of glycoside hydrolases was stable across substrates, the glycoside hydrolase subfamilies expressed were dependent upon the substrate available. Poplar and switchgrass demonstrated significant overlap in GH expression, while cellobiose resulted in upregulation of alternate GH families. In total 186 transcripts from the GH family were upregulated with the majority upregulated in the biomass cultures. Only 43 transcripts were upregulated on cellobiose, with some overlap with switchgrass and poplar, and no upregulation from maltose cultures. Determining what GH's are activated for maltose is hampered due to the overlapping carbon source present in the control culture.

Gene set enrichment analysis revealed several GH gene sets that are enriched under at least one of the conditions. Surprisingly, gene sets of cellulases (GH9, GH48F, GH1), hemicellulases (GH10, GH11, GH11/12), and other accessory enzymes (GH88, carbohydrate esterase, Polysaccharide deacetylase) were enriched on maltose, but not under any other conditions when compared to the no substrate control (Figure 7.3.A). This suggests that maltose may trigger the expression of a variety of CAZymes, but does not have a significant effect on the expression of Auxiliary Activity enzymes. The discrepancy between this observation and that from the comparison of normalized counts (Figure 7.2) that resulted in no increase in overall GH expression on maltose may be the result of small increases in the expression of many members of the gene set rather than large changes in only a few members of the set. Such a behavior may result in small change in overall expression, but the identification of an enriched gene set, identifying the importance of utilizing multiple methods to obtain the best picture of expression changes under varying conditions.

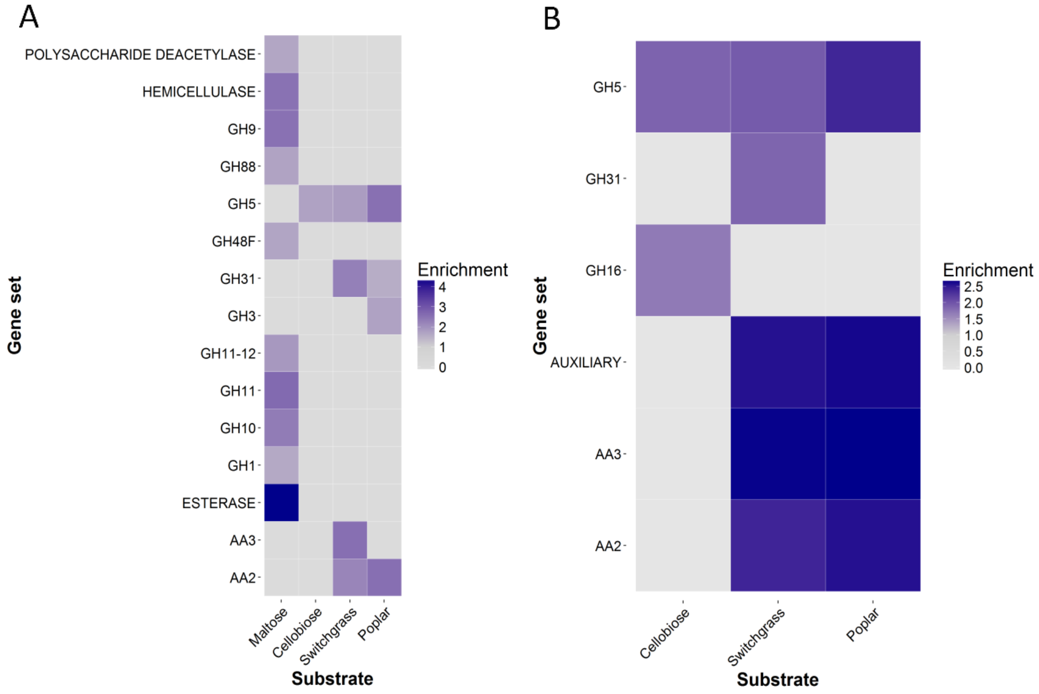


Figure 7.3. Gene Set Enrichment Analysis of CAZyme expression

Figure 3. Gene set enrichment results of different substrates as compared to the control culture (A) and the maltose culture (B). Switchgrass and Poplar show enrichment of the CAZy family AA2, which is consistent of the lignin-degrading heme peroxidases, MnP, LP, and VP.

Other conditions demonstrated enrichment of alternative GH families. Growth on cellobiose resulted in the enrichment of GH5 and GH16 cellulases when compared to expression on maltose (Figure 7.3.B). This result is not surprising as these enzymes are involved in the breakdown of cellulose, from which cellobiose is a degradation product. SG and Po demonstrated enrichment of GH5, GH31, and GH3 cellulases compared to control cultures (and GH5 and GH31 only compared to maltose cultures). Though cellulases are enriched on biomass, there was no enrichment hemicellulase GH families when compared to either the control culture or to the maltose cultures. These results suggest that for growth on biomass the regulation of AA families may be more important than the regulation of cellulases and hemicellulases and is in line with the behavior of white-rot fungi to break down lignin from woody materials and leave behind cellulose enriched detritus.

7.1.2. Conclusions

Study of *P. cinnabarinus* transcriptomics and regulation adds to our understanding of lignin breakdown in the environment. *P. cinnabarinus* PB 94 makes use of a full suite of lignin-degrading enzymes including laccases, MnP, and LP as well as alcohol oxidases that supply hydrogen peroxide to fuel the catalytic cycle of the lignin-degrading peroxidases. When grown on a variety of substrates, we observed that the expression of these enzymes is tailored to the substrate that is present. While the expression of glycoside hydrolases and glycosyltransferases was relatively consistent across all substrates, only displaying upregulation on switchgrass, the auxiliary activity groups AA2 and AA3 demonstrated drastic increases in overall expression during growth on lignin-containing substrates. AA1 laccase expression did not increase on lignin rich substrates. Instead, expression of a single laccase transcript displayed a dramatic increase during growth on cellobiose, suggesting that this molecule may be a regulatory trigger for expression. In many cases, enrichment of entire groups of enzymes was dictated by relatively small subsets of transcripts. This points to these few transcripts as some of the key players in the activities. Alignment of RNAseq reads from PB 94 to the *P. cinnabarinus* BRFM137 exome was poor, demanding a *de novo* assembly for PB 94. This assembly resulted in many more predicted genes compared to BRFM137. While some of this difference may have resulted from the assembly method, it highlights the benefit of transcriptomics for gene identification to complement other gene modeling methods, particularly for understudied organisms. Comparing the two strains of *P. cinnabarinus,* the transcriptome of strain PB 94 contains more MnP and LP genes compared to the previously studied strain BRFM137[224], and exhibits upregulation of their expression under lignin-rich conditions. Furthermore, *P. cinnabarinus* PB 94 uses numerous strategies to deal with lignin during carbohydrate metabolism. The insight provided here into how lignin degradation is controlled by *P. cinnabarinus* PB 94 can be used to better develop biotechnological processes for lignin depolymerization. Holistic knowledge of the entire processes is important in optimizing any biologically based technology, not only to quickly address any problems that may arise, but also to ensure maximum efficiency. This work represents an important step in technological advancement for the use of lignin as a chemical feedstock.

7.1.3. Materials and Methods

Growth and Isolation of P. cinnabarinus PB 94

P. cinnabarinus PB 94 was obtained from the ATCC and maintained on Remel malt extract agar plates (33.6 g/L). Experimental cultures were grown on five different carbon sources: glucose, maltose, cellobiose, switchgrass, and Poplar wood chips, as well as a control culture consisting solely of the semi-minimal media (sMM) and the agar plug from the *P. cinnabarinus* PB 94 maintenance plate. The sMM consisted of diammonium tartrate (1.84 g/L), disodium tartrate (2.3 g/L), KH_2PO_4 (1.33 g/L), $CaCl_2$•$2H_2O$ (0.1 g/L), $MgSO_4$•$7H_2O$ (0.5 g/L), $FeSO_4$•$7H_2O$ (0.07 g/L), $ZnSO_4$•$7H_2O$ (0.046 g/L), $MnSO_4$•H_2O (0.035 g/L), $CuSO_4$•$5H_2O$ (0.007 g/L), and yeast extract (1 g/L). Liquid cultures consisted of 45 mL sMM (autoclaved) combined with 5 mL of maltose or cellobiose dissolved in sMM (final concentration 20 g/L). Maltose and cellobiose solutions added to the sMM were filtered through a 0.22 µM sterile filter. Biomass cultures contained 2 g biomass (Switchgrass or Poplar) and 20 mL of sMM. Each culture was inoculated with a 5-mm cube of agar overgrown with *P. cinnabarinus* PB 94. The maltose and control cultures were grown for five days and the cellobiose culture was grown for nine days. Liquid cultures were centrifuged at 3220 RPM for ten minutes at 4 °C to isolate the fungal cells from the sMM, 1 mL of RNAlater™ was added to the cells and the samples were stored at -80 °C. Biomass cultures were grown for twelve days and the RNA was immediately isolated at the time of harvest.

RNA Isolation

Total RNA was isolated using a Qiagen™ RNeasy mini kit with Qiagen™ Qiashedder spin columns. Fungal cells were homogenized via grinding with a mortar and pestle under liquid nitrogen. RNA isolation then proceeded following the plant and fungi protocol with on column DNA digest. The RNA concentration was determined using a Qubit fluorometric assay and the RNA integrity number (RINe) was determined using an Agilent 2200 tape station. RNA was isolated from three biological replicates for each growth condition.

cDNA Library Preparation and Sequencing

cDNA libraries for cultures grown on soluble substrates were prepared using the TruSeq™ Stranded mRNA Library Prep Kit (Illumina, San Diego, CA) following the standard protocol. mRNA from cultures grown on solid biomass was selected using the Ribo Zero Gold™ rRNA removal kit for yeast due to low RINe scores (~5). The yeast kit was selected after comparing ribosomal RNA sequences for *P. cinnabarinus* to those used in the Ribo-zero kit using the RNA MatchMaker tool from Epicentre (www.epibio.com/rnamatchmaker). After cDNA library preparation, each sample was normalized and pooled together for a final concentration of 4 pM and sequenced on an Illumina NextSeq 500 the v2 mid-output 150 cycle kit, with paired end, 75 base reads.

The transcriptome of *P. cinnabrainus* was assembled *de novo* using the Trinity algorithm[122] and reads from one sample under each growth condition. Transcript IDs are formatted as ‘TR##|c#_g#_i#’ where ‘TR##|c#_g#’ indicates a gene and ‘i#’ after the gene indicates isoforms, if any are identified.

Blast2GO was used to perform BLASTx and InterPro scan annotation of the *de novo* transcriptome to provide insight into the predicted function of transcripts.

Differential Expression Analysis

Transcripts were quantified for differential expression using the Trinity utility function, align_and_estimate_abundance.pl, to perform RSEM analysis[123]. Differential expression analysis was then performed using the DESeq2 package for the R programming platform[126]. The threshold for a gene to be considered regulated was a Log2 fold change value of at least 1 and a p-value of 0.01. Gene set enrichment analysis (GSEA)[210,211] was performed using gene sets for CAZy predicted function determined based on the protein domain annotations assigned by the InterPro scan. To compare expression between transcripts, counts measured in transcripts per million (TPM) were used as determined by the RSEM analysis.

7.2. Appendix B: Bioinformatic identification of membrane proteins

The work described here is featured in the Open Access article: Seppälä, S., Solomon, K. V., Gilmore, S. P., Henske, J. K. & O'Malley, M. A. Mapping the membrane proteome of anaerobic gut fungi identifies a wealth of carbohydrate binding proteins and transporters. Microbial Cell Factories 15, 212, (2016)[179]

7.2.1. Introduction

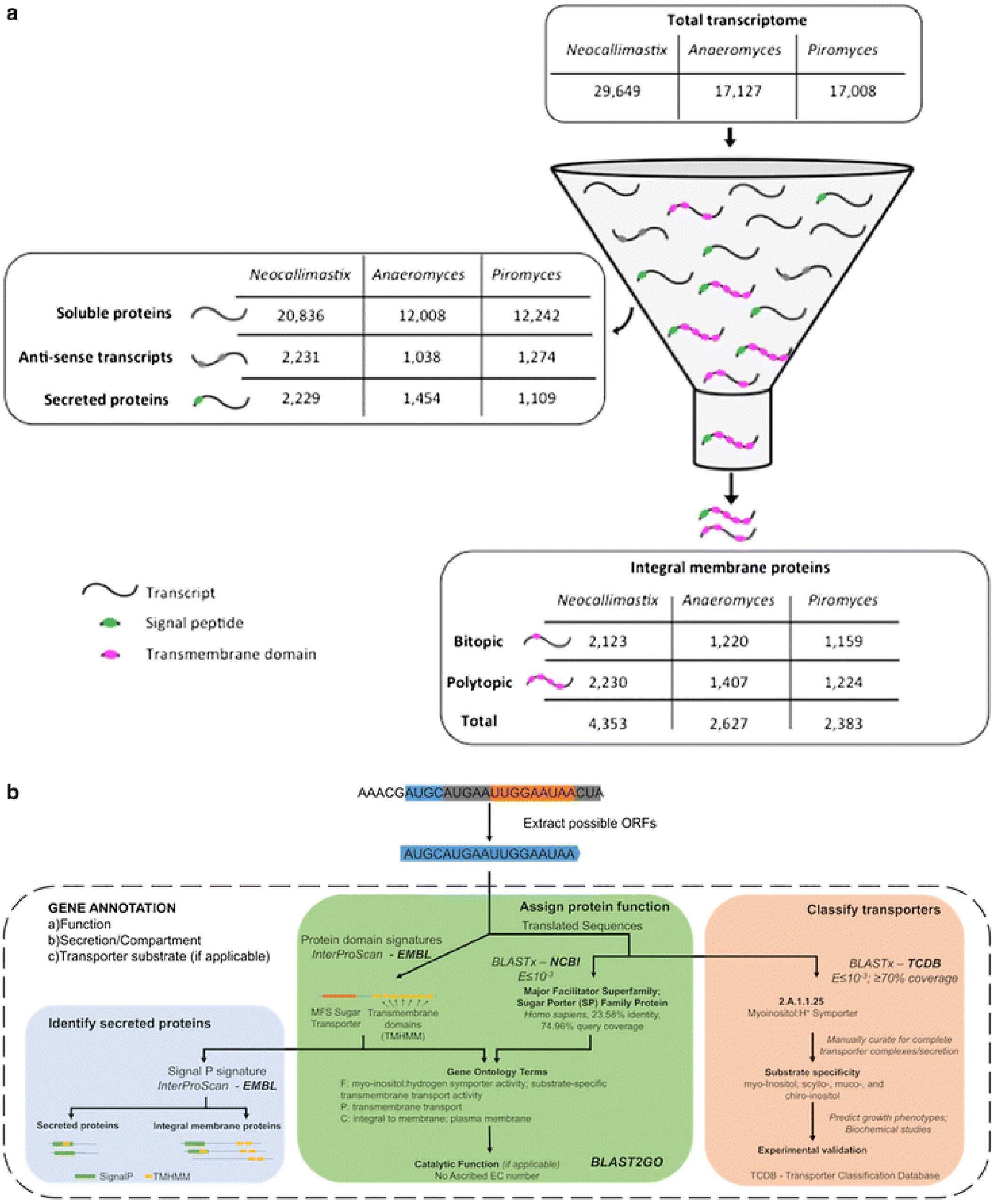
Cellular membranes are an important barrier separating the cellular contents from the surrounding environment, selectively transporting molecules in and out of the cell and sensing environmental cues to trigger changes in cellular function[252]. Membrane function, particularly in regard to transport proteins, is an important consideration for industrial biotechnology to improve strain performance and stability[253]. Membrane embedded transport proteins dictate the uptake and secretion of molecules. Engineering cellular uptake can improve substrate utilization and consequently flux towards product formation[254-256]. Similarly, employing the correct secretion systems can also increase flux as well as prevent product related toxicity and facilitate product purification by secreting the product outside of the cell[257,258]. To support these efforts, there is a need to identify novel transporter proteins from the magnitude of sequencing data available.

Bioinformatic tools are valuable resources for predicting the function of unknown gene sequences. By comparing the nucleotide and predicted amino acid sequences of assembled transcripts to sequences within databases comprised of known proteins, putative function of the protein encoded in that sequence can be identified[101-103]. This computational approach can also be tailored to search specifically for membrane proteins. The transcriptomic data collected from three isolated strains of anaerobic gut fungi, *Neocallimastix californiae*, *Anaeromyces robustus*, and *Piromyces finnis,* was mined for integral membrane proteins to characterize these organisms' ability to survive in the competitive microbial environment of the herbivore gut. While these gut fungi secrete a wealth of biomass degrading enzymes[82,83], we hypothesized that they also possess membrane-embedded transporter and receptor machinery to support their lignocellulolytic metabolism.

7.2.2. Results and Discussion

A bioinformatic analysis pipeline was designed to identify putative membrane protein sequences along with soluble and secreted proteins and predict putative membrane protein function in the transcriptomes of *A. robustus*, *N. californiae,* and *P. finnis* (Figure 7.4). Combining the use of multiple sequence analysis tools including EMBL InterProScan[102], NCBI BLAST[101,103], Gene Ontology (GO)[154], and the transporter classification database (TCDB)[259], we separated predicted transmembrane sequences from soluble proteins and assigned putative function to membrane proteins. We identified that approximately 15% of the transcriptomes of each fungus represent transmembrane sequences. Nearly half of these proteins are bitopic, containing only one transmembrane segment, that may be cleaved and released into the extracellular environment[260].

Gene Ontology was used for broad classification of function by binning predicted activities into Transport, Sensing and Signaling, Catalytic, Other, and Unknown functional groups. Approximately one third of all transmembrane proteins in each strain were assigned to transport, sensing/signaling, or catalysis (Figure 7.5). While many sequences were assigned multiple GO-terms, here transcripts were counted only once and thus this classification is not exhaustive. Interestingly, approximately half of membrane proteins have no GO-annotation, a result likely caused by low natural abundance of these proteins and difficulty in annotating these transcriptomes. Only about 30% of each transcriptome can be annotated by the NCBI database[79,83].


a
Transcript
Signal peptide
Transmembrane domain
b
AAACGAUGCAUGAAUUGGAAUAACUA
Extract possible ORFs
AUGCAUGAAUUGGAAUAA
GENE ANNOTATION
a)Function
b)Secretion/Compartment
c)Transporter substrate (if applicable)
Identify secreted proteins
Signal P signature
InterProScan - EMBL
Secreted proteins
Integral membrane proteins
SignalP
TMHMM
Assign protein function
Translated Sequences
Protein domain signatures
InterProScan - EMBL
MFS Sugar Transporter
Transmembrane domains (TMHMM)
BLASTx – NCBI
E≤10⁻³
Major Facilitator Superfamily; Sugar Porter (SP) Family Protein
Homo sapiens, 23.58% identity, 74.96% query coverage
Gene Ontology Terms
F: myo-inositol:hydrogen symporter activity; substrate-specific transmembrane transport activity
P: transmembrane transport
C: integral to membrane; plasma membrane
Catalytic Function (if applicable)
No Ascribed EC number
BLAST2GO
Classify transporters
BLASTx – TCDB
E≤10⁻³; ≥70% coverage
2.A.1.1.25
Myoinositol:H⁺ Symporter
Manually curate for complete transporter complexes/secretion
Substrate specificity
myo-Inositol; scyllo-, muco-, and chiro-inositol
Predict growth phenotypes; Biochemical studies
Experimental validation
TCDB - Transporter Classification Database

Total transcriptome

Neocallimastix	Anaeromyces	Piromyces
29,649	17,127	17,008

	Neocallimastix	Anaeromyces	Piromyces
Soluble proteins	20,836	12,008	12,242
Anti-sense transcripts	2,231	1,038	1,274
Secreted proteins	2,229	1,454	1,109

Integral membrane proteins

	Neocallimastix	Anaeromyces	Piromyces
Bitopic	2,123	1,220	1,159
Polytopic	2,230	1,407	1,224
Total	4,353	2,627	2,383

Figure 7.4. Membrane protein bioinformatic analysis pipeline

Membrane proteins were separated from soluble proteins based on the presences of predicted transmembrane domains (TMHMM) (A). Predicted function was assigned using a combine sequence analysis approach using EMBL InterProScan, NCBI BLAST, the transporter classification database (TCDB), and Gene Ontology (B).

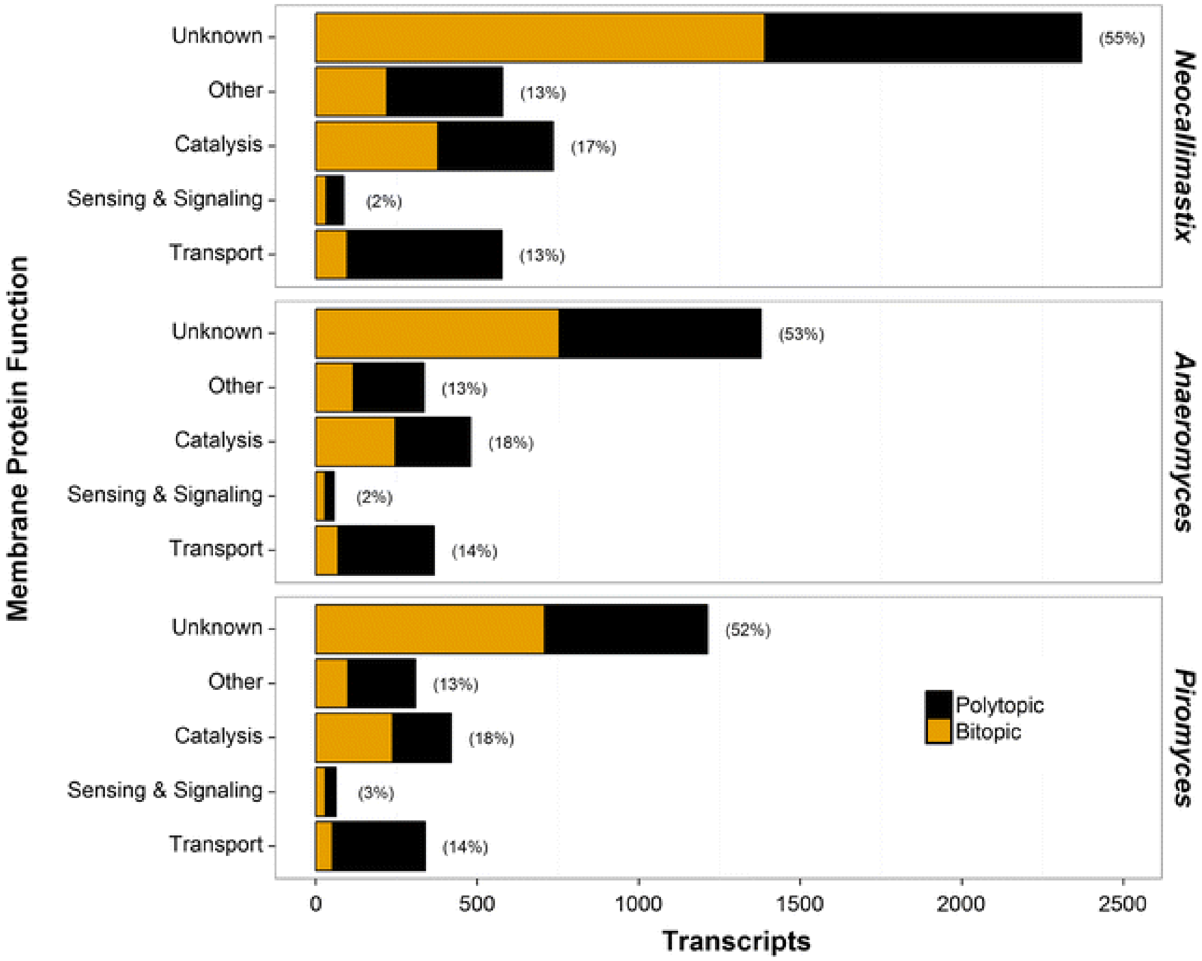


Figure 7.5. Gene Ontology summary for gut fungal membrane proteins

Putative integral membrane protein function was classified by gene ontology (GO) and binned into the major categories of Transport, Sensing and Signaling, Catalysis, Other, and Unknown. While many sequences had no predicted function by this method, approximately one third were assigned to transport, sensing/signaling, or catalytic function.

To further examine, the transport mechanisms present in the gut fungi, we aligned the assembled transcripts from each fungus to the TCDB using BLASTx alignment[103,259]. For this alignment, all transcripts were used, not just predicted transmembrane transcripts, as many transporters contain multiple sub units, including soluble subunits peripherally associated with the membrane. Strict 70% coverage, 70% identity, and alignment E-value less than 10^{-3}

criteria was used to increase confidence in the predictions. Using this criteria, we identified a total of 983 solute transporters, 282 protein biogenesis/general secretory pathway transporters, 223 nuclear import/export transporters, 103 peroxisomal import machinery proteins, 57 plastid import machinery transporters, and 220 transporters of other functions (Table 7.6). Among the solute transporters are predicted functions in transport of a broad range of molecules that can be beneficial for engineering of production organisms. These solutes include sugars, organic ions and metabolites, drugs and lipids, and ions and trace metals. Sugar transporters include predicted transport of mannose, fructose, xylose, sucrose, cellobiose, and myoinositol. While experimental characterization is necessary to confirm these functions, these transporters represent an opportunity to improve carbohydrate uptake in production microbes. Other metabolite, drug and lipid, and trace metal transporters can be used to improve cell health and prevent toxicity of fuel and chemical products.

Table 7.6. Putative functions of fugal transporters predicted by TCDB alignment

	Neocallimastix	***Anaeromyces***	***Piromyces***	**Total**
Solute Transporter	435	312	236	983
Protein biogenesis/ secretory pathway	138	73	71	282
Nuclear import/export	90	64	69	223
Peroxisomal import machinery	38	29	36	103
Plastid import machinery	29	13	15	57
Other	96	63	61	220

In addition to transporters, we identified putative sensing proteins, specifically G protein-coupled receptors (GPCRs). GPCRs represent the largest class of receptors in eukaryotes[261] and contain a highly conserved seven transmembrane domains[262]. In addition to receptor function prediction from InterPro and BLAST annotations, the presence of a full seven transmembrane domains was also used to select putative GPCRs from the transcriptome sequences. Using this approach, we identified 53 putative GPCRs in *N. californiae*, 25 in *A. robustus*, and 34 in *P. finnis* (Table 7.7). Of the five (Glutamate, rhodopsin, adhesion, frizzled, and secretin) or six (A-F) classes of GPCRs[263,264], we identified primarily proteins in the Class C/Glutamate family of receptors.

Table 7.7. G protein-coupled receptors identified in gut fungi

	Neocallimastix	***Anaeromyces***	***Piromyces***
Rhodopsin/Dicty-CAR	2	1	2
Class C (Glutamate)	51	24	32
Total	53	25	34

The Glutamate, or class C GPCRs were believe to be absent from the fungal kingdom until recently[265] and include glutamate receptors, calcium sensing receptors, sweet taste receptors, and gamma aminobutyric acid receptors type B ($GABA_B$)[266]. These receptors typically have long ligand binding domains (>400 amino acids) called the Atrial Natriuretic Factor (ANF) receptor domain, which is related to the prokaryotic amino acid binding domain proteins in the structural SBP Type I superfamily[267,268]. The gut fungal GPCRs we identified have a non-canonical architecture containing putative carbohydrate binding domains. The GPCR sequences identify in the gut fungal transcriptomes are predicted to have the large extracellular domains, but rather than the ANF domain, approximately 30% of the GPCRs display pectin lyase fold/virulence factor protein domains (IPR011050; IPR012334). Pectin is a major component of plant cell walls and pectin and pectate lyases are virulence factors that are secreted by plant pathogens[269] and these receptors may be involved in biomass sensing. Nearly half of the gut fungal GPCRs identified contain an amino-terminal SBP Type II domain

(SCOP superfamily SSF53850) that are similar to prokaryotic substrate binding proteins associated with sugar uptake systems. The diversity of the amino-terminal domains in the gut fungal GPCRs corroborate the prediction that these proteins are involved in carbohydrate and biomass metabolism.

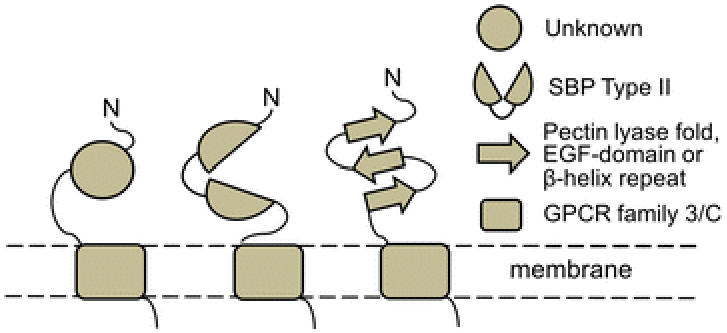


Figure 7.6. Gut Fungal GPCRs contain non-canonical extracellular domains

Extracellular domains identified in gut fungal GPCRs contain atypical domains including SBP Type II domains associatd with sugar uptake and pectin lyase fold domains.

7.2.3. Conclusions

Integral membrane proteins represent an important component of living cells and it is becoming increasingly clear that membrane transporters and receptors are essential for the engineering and stability of microbial production strains. We used a relatively simple bioinformatic strategy to identify predicted transmembrane proteins. This method identified a large number of transporter proteins associated with carbohydrate uptake and toxin export that can be used to improve production in engineered microbial strains by increasing flux toward product and alleviating toxic effects. Examination of GPCRs identified in the transcriptomes highlight potential application in the sensing of biomass and associated cellular responses to support biomass degradation and convey a competitive edge to the slow growing gut fungi in the microbial community of the rumen. Overall, this analysis identified a diversity of transporters and sensing proteins with potential for improving microbial engineering for lignocellulose-based production.

7.2.4. Materials and Methods

Identification of Integral Membrane and Other Secreted Proteins

We identified secreted proteins within the transcriptomes by parsing the annotation files provided by BLAST2GO for InterPro domain hits. Transmembrane domains were predicted by Phobius[270] and TMHMM[271]. Signal peptides were predicted by Phobius and SignalP[260].

Filtering and Classifying the Transcriptome

Membrane protein candidates were classified into one of four primary roles on the basis of their associated GO Terms in the precedence order: ‘Transport’, ‘Sensing & Signaling’, ‘Catalysis’, ‘Other’, and ‘Unknown’. Each GO annotation was parsed and searched for functional keywords as follows: Transport encompasses all membrane proteins with a stated “transport”, “symport”, or “V-type ATPase” role such as ABC transporters, P-type ATPase ion pumps, solute symporters, antiporters, and uniporters; Sensing & Signaling includes proteins annotated with a “receptor”, “signal”, or “sensor” function; Catalysis proteins all have roles that terminate in ‘-ase’; Unknown includes proteins that cannot be assigned a GO term while Other counts the remaining unassigned proteins. To better represent the total protein count encoded in the transcriptome, proteins with multiple functions are only assigned to the role of highest precedence. For example, ABC transporters with both transport and catalytic ATPase functions are binned only once under Transport.

Transporter Analysis

The translated amino acid sequence for each transcript was aligned to the Transporter Classification system Database (TCDB)[259] using a local installation of NCBI BLAST's blastp. TCDB database was downloaded January 15, 2015. To increase the confidence in our predictions, we filtered the results to include only hits that covered at least 70% of the amino acid sequences of both the query and the subject. After filtering by coverage, the hit with smallest E-value was selected, with a maximum cutoff of 10-3.

Identification of putative GPCRs

Transcripts with putative GPCR function were identified by searching the functional annotations provided by NCBI BLAST and InterPro databases for keywords 'GPCR' and 'G-protein coupled receptor'. From this subset, only sequences that contained between 7 and 9 transmembrane domains as identified by transmembrane Hidden Markov Models (TMHMM). This ensured that transcripts identified were full length GPCRs with 7 transmembrane domains and allowed for the presence of hydrophobic signal sequences that may also be identified as transmembrane domains. Predicted N-terminal domains were identified by the InterPro based annotations present in the extracellular N-terminal region. These were identified by selecting all domains from the GPCRs that were present before the first of the seven transmembrane sequences typical of GPCRs, restricting the search to only the N-terminal extracellular region.

7.3. Appendix C: Epigenetics and chromatin isolation in anaerobic gut fungi

7.3.1. Introduction

In addition to regulatory DNA sequences, the control of gene expression is dictated by a number of epigenetic factors. Epigenetics describes changes in gene expression that do not involve any changes to the DNA sequence itself include DNA methylation, histone modification, and chromatin structure[104]. While the impact of these factors is not yet fully understood it is becoming increasingly clear that they play an important role in gene expression. There are several sequencing methods that aim to probe the epigenetic features of the genome. These methods typically use enzymatic digestion of intact chromatin (DNA molecules that maintain their interaction with associate histone and regulatory proteins) to target isolation of specific regions of DNA based on their accessibility in the chromatin structure. These methods include DNase-Seq, MNase-Seq, and FAIRE-Seq[107]. Techniques such as these provide an opportunity to study accessible regions of DNA and highly expressed gene sequences in anaerobic gut fungi to provide insight into regulatory regions responsible for control of gene expression. However, to implement these techniques, it is first necessary to purify intact chromatin.

7.3.2. Results and Discussion

To ensure isolation of DNA with its associated proteins, as well as isolation of high molecular weight DNA fragments, liquid nitrogen grinding was used as a lysis methods to reduce shearing of chromatin. This did result primarily in isolation of high molecular weight DNA (>10kb), suggesting that the isolated DNA may be extracted from the cells in a manner that retains protein interactions. However, smearing in the lanes of the DNA gel indicates that while a much of the DNA is present as high molecular weight fragments, some of it is degraded (Figure 7.7). This degraded DNA may be a result the lytic lifecycle releasing DNA into the media where it is degraded, or through damage to the DNA during the isolation process.

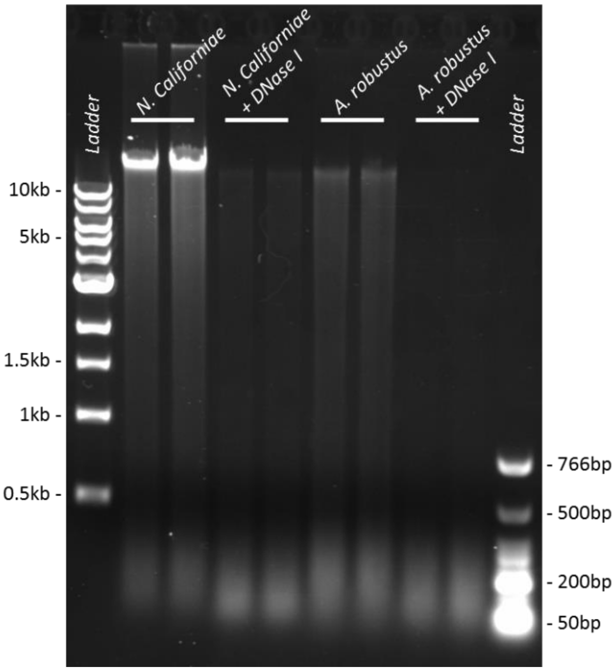


Figure 7.7. DNase-I digest of chromatin in *N. californiae*and *A. robustus*

DNA isolated by liquid nitrogen grinding was digested with DNase I prior to phenol-chloroform extraction to target digestion of exposed regions of DNA (DNA not interacting with histone proteins). Primarily high molecular weight DNA was present in samples not digested with DNase with some degraded DNA present as identified in smearing as well as low molecular weight fragments. Almost all high molecular weight DNA was digested in samples treated with DNase I (+ DNase I). Presence of degrading DNA in undigested samples does result in issues with isolating digested DNA fragments.

To assess DNase I digest of DNA, cells were ground in liquid nitrogen, and immediately resuspended in a nuclease digestion buffer[272] for DNase I incubation, and the reaction was stopped using a stop reaction buffer containing SDS to denature proteins[272]. Non-digested control samples were resuspended in a 50:50 mixture of nuclease digestion buffer and stop reaction buffer with no DNase I added. DNase I digestion resulted in the nearly complete degradation of all high molecular weight DNA (Figure 7.7). While this suggests that DNase I effectively digests exposed regions of DNA in the chromatin structure, it is still unclear if intact chromatin has been isolated.

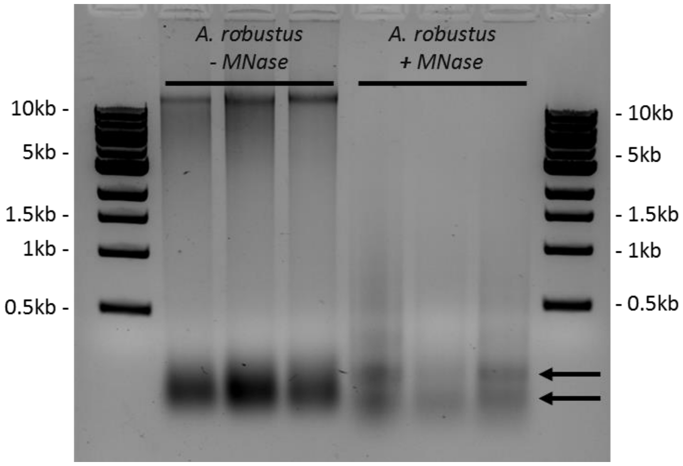


Figure 7.8. MNase digest of *A. robustus* DNA

DNA from *A. robustus* was isolated via liquid nitrogen grinded and subsequently digested by micrococcal nuclease (MNase). This resulted in the presence of two low molecular weight bands that are likely nucleosome regions of DNA not digested due to their interaction with histone proteins. Undigested DNA was primarily present as high molecular weight fragments although a significant amount of low molecular weight degraded DNA was also present.

Digestion with Micrococcal nuclease (MNase) was used to assess the success in chromatin isolation. While DNase I will cuts exposed regions of DNA in the chromatin structure into approximately 75 base pair fragments leading to sequencing of the exposed region[273], MNase more extensively digests exposed DNA leaving only DNA present in nucleosomes (DNA interacting with histones)[107]. This results in the presence of single nucleosomes(~150 bp in length) as well as nucleosome repeats of conserved lengths[272]. Thus, if banding is present in the digested DNA, intact chromatin was in fact isolated. When run on a DNA gel, MNase digested DNA did reveal two faint bands that may represent a single and double nucleosome region (Figure 7.8). However, the lack of additional bands may indicate that the full extent of native DNA-protein interactions is not maintained during the isolation procedure. To ensure isolation of exposed regions of DNA are isolated for sequencing, as is the goal with DNase-seq, alternative methods may be used, such as Formaldehyde Assisted Isolation of Regulatory Elements Sequencing (FAIRE-Seq). This method uses incubation of cell cultures with formaldehyde to crosslink DNA-interacting proteins to the genome. Then the DNA is fragmented via sonication and the fragmented DNA is separated from protein-interacting DNA by phenol:chloroform extraction[274]. This procedure has not yet been tested but present an opportunity to improve the isolation of exposed DNA regions.

7.3.3. Conclusions

We have begun to identify methods for the study of regulatory elements within the genomes of anaerobic gut fungi. However, a few challenges still remain before successful sequencing efforts can be completed. First, the presence of degraded DNA in non-digested DNA samples present a challenge to ensure that this DNA is not sequenced alongside the target, digested regions of DNA. Second, improved extraction methods to ensure that protein-DNA interactions are maintained through extraction and nuclease digestion. Alternative methods such as FAIRE-Seq, that uses formaldehyde to cross link the proteins to the DNA it interacts with to ensure that these interactions are not lost in the process. In addition to targeted isolation of DNA regions exposed in the chromatin structure, the real-time nature of Pacific Biosciences SMRT sequencing identifies methylated nucleotides in the genome sequence.

Analysis of adenosine methylation in a variety of fungi has linked the presence of this feature to in regions of DNA to highly expressed genes, presenting another opportunity to identify gene loci and regulatory elements in the genomes of anaerobic gut fungi[275]. Either through improving methods already implemented or introducing new methods for isolation and/or analysis of regulatory elements from gut fungal genomes valuable information about the control of gene expression in anaerobic gut fungi can be gathered.

7.3.4. Materials and Methods

DNA Isolation by liquid nitrogen grinding

To isolate DNA, cell cultures were grown for 5-7 days on soluble carbon sources such as glucose and cellobiose. Cultures were then spun down to collect cells which were pat dry on paper towels before grinding with a mortar and pestle under liquid nitrogen. Cells were then resuspended in either a nuclease reaction buffer, or mixture of nuclease reaction buffer and stop reaction buffer. After resuspended ground cells, phenol:chloroform:isoamyl alcohol (25:24:1) was added at a 1:1 ratio. The mixture was vortexed for 10-20 seconds, centrifuged for five minutes at 16,000xG at room temperature, and the aqueous (top) layer was collected. This process was then repeated for a second extraction. After two extractions, samples were treated with 10 μg/mL of RNase A was added for RNA digestion. Ethanol precipitation was then completed by adding 1/10 volume of 0.3M sodium acetate (pH 5.2) and 2-3 volumes of ethanol and incubating overnight at -20°C. Precipitated DNA was then centrifuged at 18,000xG for 30 minutes at 4°C. The pellet was rinsed with 70% ethanol and centrifuged for two minutes at 18,000xG and 4°C twice. Finally the pellet was dried for 10-15 minutes at room temperature and resuspended in nuclease free water or Tris-EDTA (TE) buffer.

DNase/MNase Digest

DNase digest was performed by resuspending ground cells in nuclease digestion buffer containing 250 mM sucrose, 60 mM KCl, 15 mM NaCl, 0.05 mM $CaCl_2$, 3 mM $MgCl_2$, and 15 mM Tris-HCl buffered to pH 7.5. Samples were incubated for five minutes with 500 Units of DNase I per milligram of starting material. For MNase digest, samples were incubated for 5 minutes at 37°C with 1000 gels units. After incubation a Stop Reaction buffer containing 40 mM EDTA and 2% SDS was added. Then the DNA extraction was performed as described above.

7.4. Appendix D: Sequences of co-regulated transcripts

7.4.1. Coregulated sequences in Piromyces finnis *from Table 5.2*

>comp12262_c0_seq1

Atgctcaagattgggaatgtaaaactcaaaagtctaatgaatgttacgctactcttaacgaatgttggtctcaaccatattctactgaacttgct
gaaaaatgtaatgctattaatg

>comp12026_c1_seq1

Ttggtgtttaatcaacaagaacatttgctaaaaattttatttttaaattttaaa

>comp12362_c0_seq1

atggccaagggaaagtatacttctaagttcactagcgttactgctgaacttttcaatgactattatgaagataccaacactgttattagtaagaa
gatggctaataatgccgcttctcaaaaaactaacaatgtaagtaccaaacaaaataaggctgttttatcttctaaaccacagaagaagcagaa
cttaaagaagcaaaacaattctgctaacaaaaaggttattacccaatctttaaagcaaactgctagtttccttaacaacaagcaaaaatacca
acaagaattcccaactcttggacaatcatacaagtcacaaactactcaaaaaccacaaccacaaatgaaacaacaacaacaacaaattaaa
aagcaacaaccacaaaagcttttcaacaaaactactgctccaagaacttacagatctccagctacttacaccacaaagcaaaccgaagaact
tcttcgccaattctattcactttgtcaaagttacaatggtatgaattactttggtaaagccgttttccaaaactgttcttggtctaagaaccaaaat
ggtcaatggttaaattcagcttctgcccttgctttaaagaatgctccaattgttcgtaaaatgcaacgtcaagcctctgttaagaagcaaccaat
gaaacaactcactcgccaagcttctgttaagaagcaacaaccatccatgaaaatgggtaaacaacaagctgaaaaattgttccttaatgaatt
cagcgaaatctcacaattattcaatggaatgagttacttcggtaaatcatcataccaagaaaattcatgggctaagaaccgtcaaggtcaatgg
gtttctaaggcttcttctatttctttaaagaatgccccaattgttcgtaaagttcaaactattcgttcacgtcaagcttccgctaagaagcaaccag
tccaacaaaaggctccaatccgtcaagctccagttaagaagcaaccagttcaacaaaaggctccagtccgtcaagctccagttaagaaacaa
ccagttcaacaaagaaaaatgaaaaagcaacaacgttctatgaagccagttgttaacagtgctatcactatggaacaaaagcaaatgaacca
aaaaatgattgttgctcaacaaaatgcttccatgaaattacaacaacaaattcttaagcaattcaatgaaaagcaacaacaattagaaaaga
agaagaagcttgaacaacaacaaaaattaaagcaacaacaacaattaaaaatccaaaagaagaatgctgaaaagaaagctcaaactgtta
agcaaccaactaaccttaagaagcaacaattagaaaaggaaaagaagagagttcaaaaagctcaaaagactcttaacaaaggaaacaagt
taaacaacaagaccaaaagaaatgttagacaaatgtcaagaagtcaagctaagaagctccgtaagaagcaacaattagttaacactattaa
caagcgtattcaagaattagttgctgaaaaacaacaaaagcaactcttcagtcaaatggtcaagcaaaaggctgaacaaattttagaagaaa
agcacgaacttgctattcaaaagagcattgccctctctaagaaggaagctcaaggtaacaagcttcgtgaaaaggaatcattcgataagcaac
aaaaacaaattcaaaagaagcttactaagcaaaagagtcaaatcaaacaaaaggtcaacaagaagaattctcaattagaagcccaaaaga
agaagcaattagctgaattaagaaacaatttaactccagaacaattcaacaaaattcaaagcattatgcaacaaaagggtaacaacactaag
gaacaaaacgaaagacaattacaatcagaacaagccaagaagaactggcaagaacaagttagaaagatgaaggctagacaacaagatgt
tcaacaaaagcaagaaaagcaatctttaatgaagaaacgtttagaaagtattaagagaaaattaactccaaaacaacaagaaaacttaatg
aacaagcttaagataaaggaacaacaaaagcaaaagacttctcaacttaagaaccgtcaaccaagagctaataacaacaacaagaaatttg
ttaagagagccccaatcaatcaagccccagttaaacaagctccagttaagaagcaaaacactaagcaaatgaacaagaagactgaagtttg
gacttttgtttcttacaacaagaattcaaaggttcaaaagccagtccaaaagaatgccgctccaaagccagttcaacaaaagtctaagaatgtc
aaacaattcaagactattttaactagaaagttcaatgcctcagaaactaaaatgtacaacactgaattcaccgaattatgtaatgttttcgaatc
aaccaaatatgttacttactctaactacaagacctggagtatgaacaagtctggtagatatgtttccaacgcttctgttattgcagctagaaatg
ccccaagaattactagaggattatctaagcaaatgttctttggtaaaaagatgaacatgaagactcaaaatgttaagccacaacaagttagaa
acaagaagaacaacagatctagatgc

```
>comp7503_c0_seq2

atgaaaagaagaaatataattaatttgctttctactttatgtgcattattagctactaaaggagtatccgcagatattccaaaatgtacagaaag
tagagacaataatggaattacatataatattgataatcaaggctataactattgtatatataataaaaaattatgtagtttctcagatgttcaaa
gttcaagtacaacagcaccagcagaattagatattaaagagtctggtttatatttttaaaaaaaggaaatgattatcaagaaataagtggtag
tgatgataccgaatcagtttccgcagagttaatttctgttgttgctgaaagttccaaacaaaaagaagttactgttcaaactatatcaaagggtc
attatataaattataataaagaaattatttactgtactgatggaaaaagctgtgtagttaagaatccaacaggacaaaatgcattattctttata
caaaatccaatttcaggaaagggtttaattaaatgggaaaatgaggctgaaaattatgatataacagatggatattatttaaatggttctggaa
catctccattaattctttgtaaaaacaaggaatgtaaagaagttagtgcaactgcatacaatgtatatcttgatgcatctgacaatagttcccaa
aaccttattacatgtagtgccgatactactgatccatctaaagttgtttgtacaagtgaaagaggagaagaaggtgctagttatattaacagtag
tgaaattgacaaagcaacaaaaccattaattcaatgcattagaagtaaatgtaaaactgttgaagttactgaatctagtgtatattatgaagat
aagaaggatcaatctaaaattattcaatgtacttcagctccaaaatgtacaaagatttctggaacagttggtgatatttatgttggtaaaagagg
tgatggtgaaactgatgccattattaagtgtgttaatgctggtacagaaagcgtcgttgtcaaatgtacaatggatacaaatccagctaaagat
ggttattatttaaacactggctctgattcatccaataatcaagtcattgcttgtgatgatggatgtaaatctcttaaagtcaatccaggatactac
aagaatgccaattctagtgaaagtgatggaagtgatgaatatattgaatgtaacaatgaatgtaagatcgcaaaagctactagtattaaacaa
tgtccaacagatttaagtgctgtttctatttctgaagcttgtgttagcaagtcatctacaaatgaatacaccttaaaattattatataagaatacca
atgttactgaatatacttataacacatctgatttatacttccacaccagcattagttctttcccaagtatttctagtggaaatggtgttaccactctt
ttcagactttcaaagtatggtattgaacgttatattgccagtggtgttatttctgtcaatccaagttctaatcaattagttactgatgtcaatagtaa
tgtaattggaactgatgttaacttatatgattgtagtagttcaactaagatttgtaataagcgttcttcttgtcaagctaactcttacatgtatgatg
ctgaaaataaaaaggcaatttactgtgataaagatgaaaaattaactgatgttagtagtaccagtggatattacattgattcagcaactgtcat
cagtaatagaaccccatacattatttcatgtgaaggtagtacttgtactcatcttttaccaactgtcgcttcttactttgtaaattcaggaaatgat
aatgatactaaagctttaatttactgtaatggaagtacatgtatcactactactgcttcaaccggtaattatattggaaaccaacaagctggtatt
attacttgtacctcacaaaccaactgtgtttacaaggatgcttcatctactggtaacgattccaattacattaattctggatcaaataaagcatcc
tttgctttaattggttgtactaagaagggatgtgttccaaaggctgccaatactggttactatttctctgataatgtttcttctcttattaactgtgaa
agtaacaatatttgtaatttaatcaatccaactgttaattactattactatgctgatacttctgatactggtaagaactatattattaactgttctaa
gatctctgcttctattgtttgtgctaaggaagttgctgatattggaagttacattactagtcaatcaaaccgtttaattacttgttctgctaatggag
gttgtaagcaagaaattgccaaaccaggttactatcaatcagctgttaagattaccattaacaccccaagagacctctcaagtgttggttctga
aagtgaattagttagtgacattacttctagagattctactactacttataatattattgaatgttctaatactaactgtgaactcttaactgccgaa
gaattatctaacattccaatttgtgaatataatactgacaagtgttacattactcttgcatatgccttaggaaaatctactgttaacactatttctg
ctggtggtatctgtactaatgctgaccgttcaactttctactttgccactgatactattgttgtcgctccaaatgttattgatggtagtacctcaactt
atgtttacactactactactactaactgtattgttgttagcaagaaatatgctgacttatactataccgttggttcagatatttaccgtttaaatgat
ggttctgtcagtcgtttctacgattctggtaactactttgttaatgttgaaaagaatactttaattaatggaaacaacgctgataattacaataatg
aaaatgtaaaactttaccaatgtaacggaactgcatgtagaatcttagataatccagaaaacaatacctactatgctgatgtcaacaaaagaa
ttcttaagttcaatgttaatagtgattcatactcatttgcatatgaaaaagatattatttgtatcttctccaataacaaatgtactccaaatgctga
cttaaatggaagagaattctgtattacttacaagggtgaaatcgctttagctgctcatgatattaagaaccgtgaaactggtgaatgttacaaag
cttcaagtattagcaattatatatatggatacaaccaatacttatacaaaatgaatctttactccgcaactattattgatgaaaatggttacaata
ttgttagtctttcaactaataacactattagtactaaggattacaagaacagacttctttctggtaattctatcaagatttacggatgtcattcttca
acttgtaaggtttatgaaccagaagaaggtgtctattactatgatggtgctgccaagactattattaagaaggataccaatggttgggtttctcc
atctacttcaggttatgctttagtttctgttaatccaggagaaaagtacatttaccaattcaagactgaacttgatgctgttactttaatatcaaag
gctactactggttattactataccgttgataatgaaatgtatgattgtaatgatagtgacaaggcctgtgttttaattaccgaaactgattactact
ttactaacactgatgaaatttactactgtgtttacgattctgaaaatttagaaaagactgaatgtactaagcaatcttgctacattggtcaaaact
attacattagtggaaactactacagatgtgaagccggttcataccttactccaatcaaatctagatactgtaaatatgatgaaaatgttattgta
aacttcccaaccatcttaaaggaagaattcccaaatagcattaagcaagcaattgaaaacattgaaaagaataataattcaactgctgtcgct
gctagatcaaacaagaagtacttatctgttgttccagctattttcactaactgtacttacaatgttgaagaaaccgaagcttcatatgatttcgttt
gtcttaacaactttgttgctgttaatgaagaagatgattctcttgaaatttgttctattgaaaaccttggttacgttgaatgtgttgacgatgaatct
aatccagaaaaatgtaatccaagttcagccttttcaagagttgtatttaacttctttactatagcagttactatttttgcttcattatatgtaatgctt
ttc
```

```
>comp11992_c0_seq2

atgcaacaaaaaaaaataatatggaattttattttaatttttactctttatatattaaaagtaaaatcagatgcttcacctttactggaatgtacta
cttgtgtaaatggtggttgtaataataagaaattctgctttaatggtagtactattaatgcagtaaactcagcaggtaatacagggtctgttttatt
tagtggtagaactccaggaaattattttttaaaaatggtgaaattgtaacaagtacaattgaaggtattgatgatggttattcttgtgatgcttc
atctggttgctcaaagataacagtagaatctagtatagagaagacatatattaattctaaagcaactagtttattatgtgcatatatattagcaa
ctggggggaatcaacagcctttgaatgtaaaaatggaacagcaaataagtcatattttgacaatacttccaacaaagtatttagttgttcaag
cagtagatgttcattaatatcagctattgccggttattatgttgattctggtgagtatagtactgaaggaaaaactattattaattgcaatgaaaa
tccttgtaagatagaaaaaccagatggaaatgtaattgttgaattttatcttaattccggatcggataaatcctcaaatccgattatatattataa
taaagaaggtggttataaaactataactggtgatacaactgtagcatacttagattatggtacaaaagatgatagtgttgaagatgcggttattt
ataataatgtaattatttgttcttcgacgacaaagtgttcttctgttgcttataaatcaggcatatttttaagtcctgcaaatagtgcaaatgttaat
gatagtaccaatataagtcaacttattgaatgtaattcaagtggttgtgcggaattagatgatactgaaattatggaatatattggaagcaattc
tgaaaattcattttatattgatgaaatatctaaaaacttaatcagttgtatggtagataacatagataacaatagcaaggtattaaaatgtagaa
tatctaataaagaaatctcaaataaatattatttagattattcaactttttctttatcagaaaactgtgatacagaaagccatataatgactattg
aagccgccgaaaagacgtctttttgtggtataaatattatttcttgtgattcattatcaaaatgtaaatcctctaatatttcagaagatagtaattt
tattgatggtgatactggtaataatttaattgtttgtactacttttggcagtgacttcttttgtgcagtattaggtgtagaagcattaggtttaagta
actattatattaatagtggtaattccggaatttatccattattatattgtaatggaaataaaaaatgtgttgaaaagaaagcaaatacaaacgg
gtattatataactgatactagcgaaaaaataaaggaatctccattagaaattgacaatagtggttatttaattcactgtaatagtgaaacaaaa
tgtgagaaattacttgacgttgccaatgatggctattatgttaatgttggtagtgtggatacaactaaacctttaatctattataatagtgaatcat
ctgaatttgaagagaaagaaacagtagcaaatacttattatttagattcttcttctttagcttcaggaacctactcaaatttaatttattgttcttct
accaagaattgtacttctattattcctaatgatggttattatattaatgctcctggtgaagatgaactaagtttaattattgtatgcgacaaaactg
gttgtagaactggtgaaaaaacagaggaaccaattcaaaattgtattgttgataatagtatgacattatacgttggaaaatactgtataggaag
agaaagtaatgatattgaaaccaaggatcttaatttcgttattaatgattttgttattgataatgaacctattgattcttcaaacaaaaacattac
atttgtttctaatggcactaagtatcattttgtcactgtacttgctaataacttcccgggtatatcaactacagttacaacacttttccaagtcaaat
ctaattctatttctagagttgttgatgatgccgtatatattattaattcaagaaatgaaaaagttgaatccataagtggatctgtttccatcggtaa
ttcctattcaatttatacttgttcaagtactactaaattatgtatacaagaaactagttgcccatcaggaacctacttctttgatgaagataatggt
aaaggttatttatgtagtgaaaaatcaataatgcctattacagatgaaggttattatgttgatggtggttatgtggtaaataaatctcttactcca
gctgtcttaaagtgtaatgaatctggtaattgtcaaagatttattccaactaatacctatttcattaatgctggtattgacaatgataaaaaagctt
taattcattgttctaatgatcaatgtatgactgaagaagcagccattggttattaccgtgctgaatttggggaatctggaatcattgtatgtacttc
aaacactaattgtaaaatttcttctcttcaatacaactattacattaacagtggagcagataatagcgtaaagccaatcattgcttgtaataaaa
atatctattgtaatactaaaaaggctgtgtctggttactatcttgttcaagaaaatagtaatttattaataaattgtaagagtggtatttcatgtga
agctgaagatgcttctgttggttattactacaattcagctaataatgacaacaattcaagtgttgaaaccgtcattaaatgtgttacttcttccttc
cttaattctgttgtttgtaccactgaaaagaagaatgttggattttatgtatctggagcagaaaacaatattttaattaattgtattggaggtaaat
gtaagagtattgttgttgataatggtattttccgttccgctgccactattaaaaccacagtaaagaatagttcacgtgacaaatacgaagaaga
agaagaagacatgaacttaattgaacacgctggaagaagtgatgaagaaattattgaacttgacagacaaagtaatgtaatgctaagaatga
ctgaaaagaaactatattcaagagctaatagtggtgatgatgaaaatatatcaactcttatttcttgtaatggtggtgtttgtaaagaattaactg
ctgaagaattaatgtcaattccaatttgttcttacaacaatgaattatgttacttggacaattcaaattatatcacttcttctaataagaataatctt
gtttcaagtgtaaatgccggcgaattttgtacagataaatctcgttctactatttactttgctttagataccattgtagaatataaagatgttatttc
tggtgtactttcttcttcaagtacttctagcaaaaattgtattaaggcttcttcccaatacgcctccaacttattcactattggtaataatatttatca
agttaatgatggttttattaaagaagtttatgatagtggttattactttatcaatgtgaagaaaaacattttagtatatggaaatgaaattaaaga
atataatgataataatgttcgtttatacaaatgttatgatagaggatgtcgtattatggaaaaaccatctagcaatactttctatactgatgtcac
taaacgtatcattaaatataccgttgaagataacaaatactcctttgttaataagaaagaaaatacctgtacctttgaaaataatacatgtaccc
ctaaatacgatatcggagaaaatgatttctgtatgacagctgaaggtaatattgttgtagcaggtgaaaagattaaatcaagagaaactggta
gatgttatatgagtaattccatttctgaaaatgtattagcattctcgtataactctgtcctttaccttttgaatagtaatgctgctaatcaagtagtt
accagtggttattactttgcagaaaataataaatacaatagtgcagaatacaagacatttaataccacctcttctggcattactctttatggatgc
attaatcaaaattgtaaaatttatcaacctcaacctgacatctactactttgatatgttgactaatttatttaattcaaaagaagaatgatgaatgg
atttcaccaataaaggttggtcatcttttagtttctattaatcctgaggaagtttatatttacagctataccatgtctgatagtaaggaacttctttt
aactaaaaccaacaaaaatggttattactacaccattgatagaaaaatgtataattgcgatactagcatgaaagcatgcaaagaaattgatga
```

tactgcatatattttaaccaacagtaatgaattgtattactgtttagtagatagtgaaggagaagaaactgaatgtacaaagaaaatctgtact
actggacaaatctactatattaagaatgactattacaaatgtactactggatcattctttgaattaatcagatcaagaaattgtgattatgatga
aaccgttgttattaatttcccagttatttatgctgattcattcccaattagtgtttacaattcaatttccaatattgcaaagaataatcattatgttcc
aactcaaaaaactagtcgtcaatctattgaatcttaccaaggtgttttcactaactgtacctatgatgtatatgatgaagacacaacttacgacc
aaatttgtatgcagaattacgtaaaattaaatagagataaggagccagatatttgttcagtaaaacatcttggttatacttattgttcagttgaag
atggtgataataaagataagtgtagtccaagtggagtcaatacacaaaaatctctttctattttaaaacttttaacactcattttatccactataa
ttatctttgttgtttat

>comp11882_c0_seq1

Ttgctccagaaactccagctgctccagaaactccagctccaaaggctttaccagaagctccagttgctccagaagctccagctgctccagttgc
tccaactactaagactgttgttatcaagactaccaagactttaccagtcattaagactaccaagactttaccaactatcgttgaaaccaactaaa
tcaaaagaatattattataaattagaagattataagtttttaatataaatattaattatataataaataaagct

>comp11735_c0_seq1

Atggacaacaccttaactaaacaattaaaaaaatgtttattattatatatttatataattttaataatactattattatattatatgatattgatttt
c

>comp12028_c12_seq1

Atgaaattaaggttaacttcaacaaccttactttcactaaggatggtatcttcactaccgttaacaaggaaacttgtggtgtttccaacgataaa
taaatttcataatactaaaaatatccatttatca

>comp7496_c0_seq1

Ctggaagaaaaaaaaataaaaattaaattctctagtgctagaatcactattttt

>comp5143_c0_seq1

Ttgggattgcaactactgcttggaatcttagggtctacagggctaaaagttactttagaatgtccaaataacaatattaatcaaattaaatgcgc
ctcatacgccacaatagaaaatttgaaatattcaagattgcaacaatcctggctggaaaattttaagaagttt

>comp10778_c1_seq1

ttgaaagattccattcaaaaagcagaactagaaattatgaaacttcagaactggagtaataataggcctcattcatctattaataatggagtag
gatccacttctgtcattaaagaagaagaagaaaaagaagaaaaagaggcggaagaagaaaaaaaagaaaatagaggaagcgaaggaga
aaaagaagaaaaagaaaaacacggcaataaatcattgagttcaccgaaaacaaattggcagggttcaagagatagaaccgatagaactag
aagatcaaagattaataattggaatcctaattcaagcgctgcaatttttaccaacattcaactctcttatttagatatgttggaaaacaaagcaa
agaataagattatcaagcttaataaaagctcaaccgcaacttcaaaagatactacggttcaaaccaaatcgaccaacaccgaaaatcctata
acagaagaagaaagcaatagtaaaacagaaaattcaaagttaaaaacctattctgttccaccccaaagacatcgtcgtagtagttctcttacc
caaacttttataaaggataatgaattcatcgccagaagaagggatagtttttctgctggaagtaaaattcatcctatggctcttccactttcttcc
ccatcctcaaagtataattccatttctacagacctcaccggtatatcggatggttcttctagcaatattttaagaaaaggtagcttaacaggaatg
agtaataattctagtttctattcacctcgttcttcttcccttttcatggacgatattaacattttacaaaagaatagtcaaaagagaatatctggtgt
atttagcccagatatgaaatccaattccatgttaagcgcctcggctggctcagaaataccggtggtagatcaaaaatcattaaataatgacttta
atttattccaatatggcatatcctctaatgatagtgatggtcctattcagcgtcattcaattcaatccaatggcagtcataattccttagatagtag
tggtgaagctatcggttatagtacttcgaaatcaacacctgatgttggtcaagtcttgacggtattacaaaaaaatggtttagaaggattgatt

tacctattcctccaataaaggaagatgattattccaccccatctactggtgttaaaaataccatcactcctccaagacgttcttcaagctaccata
gagtgtgtaatagtactggtagtattaacatgatggaaccatcattatctttatctcagcctttggccaccatcaccacttctaccatttcctcttcg
tctaaccatttaaatcctaataccgctgtaaccgaaggtcgttcaagaagtggaagcaaaactttcctttattccctctatcgttcaaatagtgca
agaagtgcaagaagtacgagaaagggctttaatacttatttggatacgaacgtaaaattatccaataatggtcatgtgtcatcaacaagtatg
acaccgcttgataccaaagatagtgacaatcctcgaatgagtaataaggataatgacatgaacaaaaatgaagctttttatacttcggatata
aattatcatccttatatgatggatttgaaatcctccaatgagctttaccttgcttcccttcaacaacaaattcaacaacaacagcaacagcagca
acaacgtcaattttcaatgatgccaaattcaccacctttaaataattccaattattcttctgttttggtttcaccaaccgttccaggaaacacatca
ttctccaatagtcatatcaatattcattctcaacctcaatcaccagttatttctcatattcaaccaaatacatttactggtggtatgaatgataata
attatgattcttcattttcttcttcatccaaatcattacttctcaagtattctgattcccaaaatcaattagtgaccgttgatttaaacgacgaatcc
ataattcctacttccttttaaaaaattatccaaagaataattcaacaaccgtcaacccagtgattagaaattctttaaaaattttaaatatattg
gaacaaacctatctt

>comp13233_c0_seq1

Atgaaattctcaactttattcactactctttctactgttgctagtgttgctttagcttcctactgcggtactcaatgtgatccaaacaaaattccaga
cacttcactttctggtccaattcaattagttgctgttaatgataataaagattctagtattcattatcaagttgctggtaccgttgtcatcgaaaatg
attgtgtcttcactgtaaagggtttcaaacttactccaaaaagtgatggtgcaaaatggtatggtgctagcgatccaaactcaaatgaaggtatt
cttctttctgaacaagaagttggtgttacttccactgctactgatttaagttataatattaaagataccagtttattctgtcatgcctctttaattaa
ggatgttggtaatggtggtattcttcgtttaatggatagaaattcacaacttcttgcttacgctaagatttctgctggtgctgcttctagtccagca
aaaccatctggtgatgctcaaaaaactactactaagaaggacgcttcagaaaccaaaccagcttcaactcaaactagtgaagctaccaagcct
ggtaatgctactgaaccagctactgaagctgccaatccatctgctggtacttctacaactactgctgctccagctgcttcttcaaccaataattca
aagccaatcactaatgtttcacaaactactagtggttctctttccaattacaaggttccatccgttgctctctatgctgcacttttagttcttgctttt
cttaaattt

>comp6536_c0_seq1

atgggcagaaaaggaagtatgaatttacaagttataactgatagctttgtaaaagatcaaagtcgaagaagaaattgtcagatattgagtgcc
agagttccaacaacacctattaaatctcaacaaagaaaattatcatcaccagaaattgaattaacttcatcaccaaataaaaaaccagaagtt
aaagctttatctattcagactgcaccaaaatctgctatcccaaaaggctctgttccaagatcaggttttccaatgtctgctgttccaaaatctgca
ttcccaagtttaagaatatcaactttagcaacttcaagtaataaaacttcaggaccaaaaacaccagtaagtgcatcaggtaaaaaaagtcaa
ggattaactttaaaaacaccaacttcaaccattccagaaatcgatgcaccagataccccctgttaatataatttatgcaaatgctttctttgataat
aaaactcgtgaagatgaagaaacagaagctataagcaagaaagtattaaaacaacaaagaaaagaaaacattcgtaatatgcgtcgtgaa
cgcatgaaacaaccaaaagatattagagttgcattaattcttggtcgtttatatgatgcttcaaattatgaactaggacctttatcttactggtctg
atccagagattgaaaatattgaagaactaaagtctaaggaaattaaacaaactgaagaaacagaaaataaaccgaagaagatgattcaatt
gaatccagatggagatatagaaattgaagatgaaaaggaaaaggaccagtcaaataaaattaatcatgataaaactaataatgtaaatgctc
acccagcttgggtgccagaagatgctggttgtagtaatgttgcatgtcaaaaaacagctactagaattaatggtaaactagttaaaaaaaata
gaaaaccatatgtagctattgcttctttctatgctaaattatatgatccaattatgatgaagaaaagagaagaagaagaagaactacaattaat
tgaaaaacaaagacagaattcaattattactgaattaaaaccaatgggaaaaagttttaattcaagaaatcaaaatggtaataattctaaata
taataataatagtttttataacaactataattatagaaaccaaggagatcaaaacaattatcaaagcagacaaaactttaacaaccaaactttt
agtaataatacctttagaagtaaaaaatacgaaaataacgatgaatccaataaaataaactttaaagataatagaaatcaacaacaaccaa
atactagaagtttcaacagtaccaaaaaaggttcaatgcaaactaataactatatgaataataattataatccaacatatagttatggttatgta
tatcaatacccagtatatcaatactacgatccaaatttagtaccaaattatgaatataccaactatggaacagactataatcaaaattatagta
acaattattataataaatcttataaatataaacctagaatgaattataataataataccaatcaatatacatttaataataaaaataatgatgct
aattcttttaaaaaatatgaaaaaaaagttaatattaaatct

>comp11012_c2_seq1

Atggtggcgtttggggtggtgtttctttcaagaataaaaacaatgctaagttgggatctggtatattatatttcaaggctagaactaatgatactg
atgctcttcttcaagtat

>comp7326_c0_seq1

Atggtggcgtttggggtggtgtttctttcaagaataaaaacaatgctaagttgggatctggtatattatatttcaaggctagaactaatgatactg
atgctcttcttcaagtat

>comp14924_c0_seq1

Atgaaattctacaacgctttattattattagctgccactcttaccttaactttaggtaacggtttagatgtttctgatattgaagacaacattagtg
gtgttggacttgaagatggttttggagaaagttctgaaccagaaattgataccaacatggctccagtagaaactccagaagctccagaaactc
cagttattgttccaccatctttcccatctgttgataacaatgtcccaactactgattctattccaccaccatctatggactctgttaactctaaccca
gttactgattctgttaacactaatccaagtgttgattcagtcccaccatactctgacaacactggtgttactccaccaagtgctgaaggtattagc
ggtcaaaatgttgacaacggtgaagcttcagatgattatggaagcactgatcaaattgatcaaattgataatgctgatgtcaacccacaagatg
ttgtagatgatggtgaagcttctgatgattacggtaatgctgatggtattgaccaagttgatggtattgacagtgctaatgccaatgatgttactt
ctgatgatgaagatgaaggtctttctactagtggaaaggttgctagtggtcttgctggtgccgctgctctttcttctgctggtgtcttctattacatc
aagaaatctaagcgcgctggtttacaaagtgttcgtactcaaattactatggtt

>comp11723_c0_seq2

Atggttactactactcaatcaaactactctattatgtctactgctaccaaggttaacttggctatgaatggaagtgataaaaagtctgaaagatc
taaaaagagaagcaactttttaaacgtttattaccacaacttaatgaaaatgaaggaggttctattcaacaattacatgctattcaaatgatta
tc

7.4.2. Anaeromyces robustus *sequences from Table 5.3*

>Locus2793v1rpkm17.21

Cctttgaaaaaaaaaaaaaaagaaaacagttaatttttatattttaattccaaatcttggaacatatttcattaatagaactacattttccatt
atcagatccttgagcataacattggttaagagcagaataacattcattggatttttgagttttacattcccaatcttgagcaccacagctatggta
ataatcagcagacatgttaccattgttattattgttataattgttgttgttgttattgttgttattgttgttactgttattgttgttataatttggaa
tagtctttgtagcagtgttaccattattgttgttattgtaatttggaatagttttagtagtattgttaccataattattattgttattgttgttgttgt
tgttgtaatttggaatagtctttgtagcagtgttaccattattattattgttgttgttgttgttgtaatttggaatagttttagtagtattgttaccata
attattattgttattgttattgttgttgttgtaatttggaatagtcttagtaacattgttaccgttattgtaattgttattgttgttgttgttgtaatttgg
aatagtcttagtaacattgttaccgttattgtaattgttgttattgttgttgttgtaacttggaatagtcttagtaacattgttaccgttattgtaa
ttgttattgttgttgctattgttgttgttgtaacttggaatagtcttagtacttccattgttattgttattgttattgtaattgttattgttattgtaattgt
tattgttgttgttgttgtaacttggaatagtcttagtacttccattgttattgttattgtaattgttattgttattactattataactggttggaatactc
ttactattactgctactaggaatagttttagttacaccgtaattactattgctatttccataacttggaattgttttggtttgagcaaaacagtaaac
aatata

>Locus1323v1rpkm59.77

gtaaagaatgctccagttgttgtacgtagtagtaaaatacaacaaagttcacgtaaagtttcctcttcaagaaatcaaccagtaagacaaatga
gtagacaacaatatatagtaaattttaatagtgaattcagtgatctttgccaatgttacgatggcatgaattactttggaaaatcttctttccaaa
agaacacctggactataaatagatcaggtaaatgggtttcttctagcacttctttttcaatgaaaaatgccccagttgtgatacgtagtagcaag
caacaacaaataactcgtcaagcttcagtaaagaaacaaacaacacaacaatcaagaaaattaaaaaaacaacaacgtagcatgccagttg
ttaatagtgctattacaaaagaacaacaacaaatgaatcaaaaaagtattgttgctcagcaaaatgcctctttaaaattacaaagcaaagttct
taaacaatttaaagaaaaacaagaacaattggaaaagaaaaagaagcttgaagaacaacaaaaagaaatgaaacaagaacaaatgaga
attcaacaaaagaaaaatgctgaaagaaaattaaaagaacaacaagctgctaaacaatctattaaaaagaatgttaaacaagcccaaagta
ttaaaaaacagcaattacaaaaaaataaaattagaaatcaaaaaataaatgctccagtaagaaaaatgactaaaggtcaagaaaagaaaa
tgcgtaagagacaacatcttattaatcttgttaataaaagaattcaagaattagttgctgaaaaacaacaaaaacaactctttagtcaaatggt
aaaacaaaaagctcaacaaatattagaagaaaagcatgaagctagtattcaaaagagtatgcaaaattacaaaaaaactcaaagtaataaa
ctacttgaaaaagaaacttatgaaaagcaacaaaaatcagttcaaaagaagcttgctaaacaaaagatgcaaattagaaacaaaattaaca
acaaatcttcaagtttagaaacccaaaaatcaaaacaattaaaagaattaaagagtaaattaactccagaacaatttagtaaaattcaaagc
attttacaacaaaataataacaacagtaaagaacaaaatgccagacaattacaatctgaacaagacaaaaagaattggcaagaacaagtta
ggaaaatgaaagaaagacaacaagaaattcaacaaaaacaacgtagacaatctttaattaagaaacgttttgaaagcattaagagaaaatt
aactccaaaacagcaaattaacttaatgaacaaccttaagttaaaacaacaacaaaaggaaaataagagtaacaactcaaaagttaacaac
aagaagcaacaagtagttcaaaaaccaactaagaaacaaacaataatgaaacaacaaaaagtcaatatgaacaaaaatcaagtttggaca
tttgtaaacaacaataagactcaacaaaagcctattcaaaagtctaaaggaactaaacaatggaaaactgtttcttctagaccattaaatcctt
ctgaaaccatgatgttcaataaagaatttgctgaattatgtaatgcttttgaatctactaattatgttacatactctaattataaaacctggagcat
gaataaatctggtaaatatgtatccaatgcctctattatttctgctagaaatgctccaagaaatactaagggtggattatctaagcaattaaattt
tagtggtaaaagtagtgtttctatggcacaaaaaaacaaaccacaacaagttagaaatatcaaaaacaataaatccagatat

>Locus2632v1rpkm18.90

atgagaacaataaattacttaaaattatttgcagtgtttggtttttcattgtttaataatgtaaatgcagcaattccaaaatgtaatcaaactgaa
actaaggatgcaagcggtaatactacatataaagtaacgactactgacgtaaactcaaaatactgtttatataatgacatgctttatacatatg
atagtcctactaaaccattagaaaaattagatttagacaattttgaacctggtttaaatttcataaaagatacatcatttgaaaaaattgatgatt
caaaaagtaatgcaaatttatttaaattaaatgtagatggtaataaaaaaactgtaagggaagaagaaattaccaaaggaaattatgttaatg
cagattcagaggttatttattgtacatcaggtagtagtgaaacaaaatcatgtcaacttgttgatacaagtagtttggcatctccatcatttttcat
tcaagaatctgtatcaggtaaaggtttaattatgtatagtgaaaatggtatagaacaatacactgaaattaaggatggttattatcttaatggaa
atttcaatatgttggaaggttctaaacaattacttaagtgttcacaaaagacttgtacagaagtagttgcaaatgatggtgatgtttatgataata
ttttggaagaagatgaagttattcaatgtttattggatgctactgctaatgttgttaaatgtaaaactacaaaacctgaatctccatcattctttat
taacaagagtgtattagatttatcagaaaaaccattaatttcttgtactaattctgtatgtaaatctgaagctccatccgatccatatttatacttt
gaaaatccattaaattatactaagattatctactgttcttcaacaaaatcaaagtgtagttatttagaaggtatggaagaaggtgatgcctttgtt
agcaattatgaagacttaaatggtattgtactctgttcaattgatcaatctgattcaacattaaaatgtcaacatactactggtagtccaaatgga
tattatttaaattctggtggtgataatggtactaatcaagttattaattgtgctgataataattgtgtaactaagagcgttaacccaggttattaca
ttaatgctaatccttctgaagacaaagaattaattgaatgtaaattagaccaatgtagctttattaaaaaggatgatgctgcatgtcctaccactt
ttactgttggtgcctgttataaagattctacattagtattcaataaacttgaaaatgaagaattagttagtactaaggatgatttatatgtttatgc
aactttgagaaaattcccaagtattacaactgaaacttctactcttttccgtataactccatacagtgttgaacgttttattgatagtggtgttgttg
tcattacttcttcaaatactttagctactgacattagtgaaaatagtagtgatattttattattcgaatgcagtactaatactaagctttgcgtaagt
gtatcaagttgtactaacaatacatacatgtatgatactgcaaatcataaagctttgtactgtaaaaatggtaaattacaaataaaatctgaaa
atggttattatgttgatggtagtagtgttgttaattcaaaaactccataccttattagctgtaaagatgatgtatgtactcatattttaccgactgtt
tcatcatacttcgttaatgctggtgaagatagcagtacaaacgctttaatttattgtaataataattcttgtaacactgtatctgctagtaatggat
attatgttgccaatcaacaaagtggtattattaattgctcatcatctagtagttgtgattacaaggatgtttctggtattggaaataatgccaattt
tgttaataatggtaataataaaacaacttacgctttaatttattgtaataagaagagctgtgttccaaagaaggctaaaaatggttactacttcc
ctgataatgctagtagtcttatatattgtgaaagttctaataactgctctgtaattattccaacagttaattactattattatgctgatagttctgat
aataagaattatattattaattgtaataaagtatctacttcaattgtttgttctaaggaacttgctgatactggtagttacttaactaatcaatcta
atgttttaattacctgtagtaagaatggttcatgtaaacaagttgttgctaagccaggttactaccaatctgctgttaaaattactattaattcttc
aagagatgtttctgatgctagtgctgaaagtgaattagtcagtggaatttctggaagagattcaactactacctattccattattgaatgtactca

aacaacttgtgaatatttaactgcagaagaattaagtactatcccagtttgtgaatacaatggtgataaatgttacattactttatcttacgcttta
agtaaatctgctgttaattctattgccgctggtaatctttgtactaatgccgatcgttccgttttctattttgctactgataccattgttgttgctccat
ccgttattgctggtcaaacctctacttatgtttacactaccactactacaaattgtattattgtatctaacaaatatagtaacttatactacactgtt
ggctcagatatttatcacttagatgatggtgttattagtcatttctatgataatggttactacttcattgatattgaaaagaatactttagttaatag
taatagtattgataattacaatagtgaaaatattaaactttataaatgtaatggtattgcttgttctattatagatgaaccagaagttgctacttac
tatgctgatgttaacaagagaatcataaaatacaatgttaacaatgacgcttaccaatttgcttatgaaaaggatatagtttgtatttttgcaaac
aataaatgtactccaaatgctgatttaaatgccagagaattctgtattacttacaaaggtgaacttgttttagctgctagtgatattaagaaccgt
gaaactggtgactgttacaaggctagcagtattaataactatatttacggatacaatcaatatttatacaaaatggatgttaactctgcttcagtt
atagaaaataatggttattatcttattagtctttctactaataacactattagtgcaaaggactataaaaacagacttatcaatgctaatttaatt
aagatttatggttgtcattcatcaacttgtaaggtatatgaaccagaagatggtgtttactactatgacagtaaggcaaagacaatgttaaaga
atactgatggaatttggtacactccaaaaacttcaggttatgccttagtttctgttaatccagaagaaaaatatatttacaaatttaagagtgaa
cttgatgatattactttattatctaaggctgcaactggttattactacactattgacaatgaaatgtatgaatgtaatgaaaatgataatgtttgtg
aacaaattactgaaagtgattactactttactaatactggtgaaatctactactgtgtttatgattctgaaaacttagaaaagactgaatgtacta
agcaatcatgctatgttggccaacattatttcattaaagatggttattataagtgtgaagctggatcatactttactgctgttaaatctaaaaact
gtaaatatgatgaaaacgttattattaacttcccaactattttaaaggaagaattcccaacaaatattaaacaagctattgaaaatgttgaaaa
gaataataattctactgctgttgcagctagaactaacaaaaaatacctttctgttattccagctattttcacaaactgtacatacaatgttgaaga
aactgaagcttcctatgaccttgtttgtattaataactatgttgctgtcaacgaagaagatgacactattgaaatttgttccattgaaaatcttggt
tatgttgaatgtgttgatgatgaaacaaaccaagaaaaatgtaatccaagttctgcctatgctagaattactttcaacttctttactgtagcttta
agtgttattgcagctttttattttattttt

>Locus721v1rpkm140.98

atgcgtttctcaactattttaactattgctcttacattatctttaaaggcttactctttaccagtagctgaagatactgaaactgtaggaattgaacc
attggctggtccagtgattgatgctccagttccaccagttcttccagttccaccagttgctccagaagccccagaagctccagtcccaccagtttc
tccagaagccccagttgctccagaagctccaaaggccccagaagctccagttgctccagaagctccaaaagctccagaagctccaaaggccc
cagaagctccagttgctccagaagctccaaaggccccagaagccccagtttctccagaagctccagttgctccagaagctccaaaggccccag
aagccccagttgctccagaagctccagttgctccagaagctccaaaggccccagaagccccagttgctccagaagctccaaaggctccagaa
gccccagttgctccagaagctccaaaggccccagaagccccagttgctccagaagctccaaaggctccagaagccccagttgctccagaagct
ccagttgctccagctgctaagattttaccagctaaaatagttgctagatcaattgaaactccagctccaaaggctcttccagctaagtctattcca
attgttcctggtaatattactttaccagaagctccagctgccaagactttaccagttaaggttgctccagaagctccagaagctccagttgctcca
gaagctccaaaggctccagaagctccaaaggccccagaagctccagttgctccagaagctccagttgctccagaagctccaaaggccccaga
agctccaaaggctccagaagccccagttgctccagaagctccaaaggctccagaagctccaaaggctctagaagccccagttgctccagaag
ctccaaaggccccagaagccccagttgctccagaagctccaaaggccccagaagccccagttgctccagaagctccaaaggccccagaagcc
ccagttgctccagaagctccagttgctccagaagctccaaaggccccagaagccccaaaggccccagaagccccaaaggccccagaagctcc
agttgctccagttccaccagctactaagactttaccagctaaattagttgctagatctgaagttgaaactgaagctccagttgctccagttccacc
agctgctccagttccaccagctgctccaaaggctccagttgctccagaagctccaaaggctccagaagctccagttgctccagttccaccagct
gctccaaaggctccagctgctccagaagctccagttgctccagaagccccagttgctccagaagccccagttgctccagttgctccagttgctcc
agaagctccagttgctccagaagccccagttgctccagttccaccagctgctccaaaggctccagctgctccagaagctccagttgctccagaa
gccccagttgctccagaagccccagttgctccagttgctccagttgctccagaagctccagttgctccagaagccccagttgctccagttgctcc
agaagctccagttgctccagttgctccagaagccccagaagctccaattgttccaggaaatgtagttttaccagaagttgaagttaat

>Locus4155v1rpkm8.56

atgggaagaaaaggtagtatgaacttacaagttatcactgataactttgttaaagatcaaagcaggagaagaaatgttcaaccagcaagtgct
agacttccaacaacaccaacagtaaaagctccaacaagaaaaatgtcatctcctgatattgaattaacttcttctccaaataaaaaagatgaa
ggaacatcatcattaaccattgaaacgattccaaaatcagctataccaaaaggatctggaccaagatcaggacttccattatcagctgctccaa
aatcagctgctccaagtttaagaatttctacattagttacagctggtaataaagctagtggtccgaaaactccaaaaactccaaaaactcctaa
aactccattaagtgcatctgttaaaaagaatccaaacttaactttacaaccaccaacatctagcattccagatattgatgcaccggatacccca

gtcaatgttatttatgcaaatgctttctttgataataaaactcgtgaagatgaagaaacagaagctattaataaaaaagttttaaagcaacaaa
gaaaagaaaatattcgtaatatgcgtcgtgaacgtatgaagcaaccaaaagatattagagttgcattaattcttggccgtttatttgatgcttctc
actatgaattaggtccattatcttattggtctgatccagaaattgaaaataatgaagaattaaaacctaaaattgttgaagaagataatacacc
aaaaccaaagaaaatgattcaaattaatgaagatgaaattgaaaaagatattaaagaaagaaatactgtaatagaaaatcatgataaaact
aataatgttaatgctcatccagcttgggtaccagaagaagctggtagtagtaatatttcatgtagaaaatcagctaccttagttgatggcaaact
tgttaaaaaaggtagaaaaccatatattgctgttgcttcaatttttgccaaattatatgacccagtattaatgaaaaagaaggaagatgaagaa
aaactcagaatgatagaaagacaaaagcaaaattcaattattactgaattaaaaccaatgggaaaaagtattaattctagaaatcaaaacaa
ttcaaaccaaaactatcaacaacaaaataatggtaattataataatagaagccaaaatggccataattatcattatcaaaatagacaatacca
aaattacaataataaaaactttaattcttacaataataataataaatttagaaatgaaaataatagaaatgaaactttccaagttaattttaga
gataatagatcacaaccacagtataatactagaagttatttatttaacaacagaagtactaatcaacaacaaaataataattatatgaataata
attataataaccagaatattaattatggttatgtctaccaatatccagtttatcaatactacgacccaaatatggttcaaagttataatagtaatg
ttgattatgctaattatgattcaagttatgaacaaagttatagtaactactattataattcaaatggcagtaataattatcagaataatagattta
agccaagaatgaataatcaaaatagtaatcaaaataatagaaataataataatgaatcttttaattatagaaaatatgaaaagaaaattaat
attaaatct

7.4.3. Neocallimastix californiae *sequences from Table 5.3*

>Locus936v1rpkm136.63

Atgaatttgtttggaaagcagctttaccgaagtaagtcataccattataactttggcataatgagtagaattcacgaagaagttcttcagtttgtt
tcgcagtataactggctggagatctgtaagtttttgatacaactgattgatttacaagcttttgagactgctttttcaattgctgttgtggtttttgtt
ttatttgttgctgtggtttttgttttatttgttgttgtggcttttgaacagcagatttatatgattgtccaagagttgggaattcttgttgatataatttt
tggttgt

>Locus2411v1rpkm34.85

atgaggacattaaattgtttgaaactaatttcattattaggagctacatttttaagtataaaaaatgtgaaagctacaattccaaagtgtgttaa
aggagatgataatgcagatacagtaccatcaggttataattattgcgatttagatggtctaataaaattttttgataatgatcaattgattaatga
atccaatgtagtatgtacaagtgaagggttacattttattcagaatagctctaatttgagtgcacttgaagaaaatccaaaagctgcaatattac
ttaacttaacgattagtgaagaaggtgaatgtagttataatgaattacaaattaaagaaggtttttatgatgttggtgaagacaatttgatatatt
gtgataaaaatcaaaaatgtaatacatacagttcaacaccagttaccgatccagtgtattatataagttatgatggaaatttaataaagcaaga
agagagtgcctttagtaattcagataaaaaagatggatattatgttaatggtaataagggtaaacaattaataaaatgtgaatcatccacttgt
acagaagtagcagctcaagatggagatgcatatgttgatgttgaaactgaacaaattattatttgtagtaacggagataacggagtgaaatgt
gaatataaagacgacatagatggttatattatcaatagtagtgcaattacttcaagtgttaatcctataattaattgtgaaagtggtagttgcaa
agaaagtcctgttccagatccatattcttattatgaatatgctcttgataatacaaaagtaattgcttgttcatctgttaaaaatacatgtaaatta
gaatcaggagaagtaggtgattattttgttgctatccaaggagacgaaaagaataaattaataaaatgctcaaatgaaaatgataaagtagta
tgcaaagttggaactgcggaaaatggttactacttaaattcaggtggtaattcatcagtaaaccaaactatttactgtgatagtggtagttgtga
tactattcatgttaatccaggttactacataaacagcggttcaattgatgatgaacaaaaagatggtcttattcaatgcgatttaaatatatgtga
tacaaaggatattagtgttatagattgtagtaaattaagttctatttcttacgctactgtatgttataaggattcagcattcaacttctacaaaagt
gatgatttaattaatccaaccaatttcacaaccgctggtgaagttttcatttttgattctttaaagagattcccaagtattggtagtgaagttacca
ctctttaccacttaactgaatatggtattgaacgctacatcggaagtggtgttattggtgttaaatcaagtacaaaccaaaaagtgagtgactta
gatggtgaacttggctcagagattatattatatgattgtagtactactaccaaacaatgcacaagaagaacttcctgtgtttcaaatacatacat
gtatgatattgaaaataaagcagctttatattgtaataatggtaaattagtatctgaaacaggaaaaggatattatgttgatagtgttaccatgg
ttggatcaaaaactccatatattattaattgtgatgaaaacgaatgtactcatgaagccccaactgttcaatcctactatattaatagtagtgaat
atgacggtaattctaagaagttaatctattgtaataattcgaactgttatactgttgctgctacttctggatactatatttctaatcaacaaaatgg
tattatttcttgtacttcttctacagcatgtacttacagagacgctgctactgcaggaaataatgttaattatgttaacgctggaaagaataagag

cttaaatgcattaatttactgtaatggaaagacctgtgttccaaagactgctaaagttggttattacttctctaatcgcgttaccactcttatttatt
gtgaaagtactaactcttgtaatgaaattaatccaactgaaaactactataactacgctgatactattgacggaaagaactacattattaagtgt
tcaaaggtttcaacttcaatcatttgttcaaaggaagttccagacactggtagctacttaactagtcaaactaatattttaattaattgtacaaag
aatggaagctgtaaacaaataaacgctaagccaggtttctaccaatcagccgtaaagattaccattaactcaaagagagatgttgataatgaa
aaagaattagttaaagatatttctggaagagattctactactacatacaatattattgaatgtactactactaattgtgaattattatcagcagaa
gaattagcttctattccaatttgtgaatacaacagtgacaagtgttatattactttaaattatgctatgagtaagagcgctgtcacttctatttctgc
tggtaatatttgtactaactctgatcgttcaaagatctacttcgctactgatactattgttgttgctccatctgttattgctggtcaaactgctaccta
tgtttacaccaccactaccaccaattgtttaattgctgactctaaatatgatgattactattacactattggttctgatatctaccgtattaacgatg
gttcaattagtcatttctatgatactggttactatttcattgatattgaaaagaatgctttagtcagcagtaataacattgacaattacaacaagg
aaaatgtcaaactttacaagtgtgatggtcttaactgtgtcattatagatgaaccagataacgctacctacttctctgatgttaataagagaattg
ttaaatataatattaatagtaattcatatgtctttgcctatgaaaaggatattatttgtatctttgctaacaacaaatgtactccaaatgctgattta
aataacagagaattctgtattacttacaagggtgaaattgctttagctgctgctgatattaagaaccgtgaaactggtgactgttacaaagctag
cagtataaatagcaaaatttacggttacaaccaatacttatataatatggatattggatctgctactgtagttgacaagaatggttattacattgt
tagtctttcaagtaacagtactgttgttacaaaggattacaagaacagaatggtaaacactaattcatttaaggtttacggttgttacaatacca
actgtaaggtttacactccagaaagtggtttatactactatgatgacaattcaaagactcttttaaagaatgaagataacacttgggttgctcca
tccaattccggatatgctttagtttctattaatccaaatgaaaagtatgtctataagttcaagattgaaaatgatgttgttaccttaatatcaaagg
ccggtactggatactattacactattgacaatgaaatgtatgaatgtagtgaaaatgacaattcttgtaagttaattgaagacagtgattactac
ttcactaacgccggtgaaatttactactgtgtctacgattctgaaaacttggaaaagactgaatgtactaagcaatcttgctatgctggtcaaaa
ttactacattggtgataactactacagatgtgaagctggctcataccttactccaattaagtcaagaaattgtaaatacgatgaaaacgttattg
ttaacttcccagttattttaaaggaagaattcccaactacaatcaaacaagcaattgaaagtgttgaaagaaacaacaactccactgctgtagc
tgtaagatctaacaagaagtacttaactgttattccagccattttcactaattgtacctacaacgttgaagaaactgaagccgcttacgacttcgt
ctgtcttaacaactacgttactgtcaacgaagaagaagatactgttgaaatttgctccattgaaaaccttggatatgttgaatgtattgaagatg
acgcaaacccagaaaagtgtaaaccaagttcagcattcactagagttgtattaaatgtctttagtgtaatattcactgcaattgtttcattataca
ttgttctttat

>Locus4280v1rpkm14.40

atgtggtcattaaattttaaaaaattaattccattattaggagcatcactgttatgtattaacagtgtaaaagcagaaattcctgtatgtacagga
gaaggagcgaatattaatataagtaatgaagattatgaatattgtatttataataaaaaaatatgcggatatacagcaagtaatccacctaca
gaaataaaagatttaaaaaccggatttcaatttataaaagtaaaaaaatatgagatattaaaaaaagagaaaactgatgtcgaagaagtac
atttaattgaagtagcagaagatggtacaatcacggagataacaacaattaatgaaggatattatattgaagccgataacattctagtatattg
tgataataatggaaaatgtagagtggaaacaccggagactccaggaccatcatattatattgattataagggagaattaataaaaaatgaac
aaggtagtgatcttgaaacaattgtaaaatcaaatggctactatattaatggaaacaaaaatactaaagcttcaaaacaattaattaaatgtg
atcctacttgtactgaagtagctgctcaagatggagatgcttatattgatgttactgaagaaggtcgcgttattacctgtcaagaaacagataaa
gaaagtaaggtagtaaaatgtgaatataaaactcctgatggtggttattatatcaataaaagtaaaattgattcagctgataagcctttaattg
attgtgaaagtgaaggaaaatgtaaaattgatggggttactattccagagccatactcatattatgaaaatgctttagatccaagtaaaattatt
tcatgttcttccactaaaaattcatgtaaattagaacaaggtaacgcaaacgaatactatgttcaaattaagggggagggtaagaataatgaa
ttaatgaaatgtagtgttaaaaaagatgaagttgaatgtacaaatgtaccaaatccacaagaaggatactacttaaacgctggtggtgattcat
cttcaaatcaagttatcttctgtgatgataataagaaatgtactaccaaacatgtttctccaggttatttcattaataacggtaaagatgaggatg
aagaaaccccagatgatcttattcaatgtgattttaatatctgtaaaactgttgtttccagtattaagtgcgacggtattaaaccaacatctgcta
cggtttgttttgatggaacaatgttccaattctacaaaagtgatgacttaagtaatcctcttaatgacaccactaacggtgacttatacatttatg
atacattaaagaagttcccaactattactaacagtgaaactattactctctaccgcttaagtggaaatggtgttgaacgttacattggaagtggt
gttgttggtgtgaattcaatttctaatcaaaaggctgctgatcttgattcctctgatgttataatttatgattgtagtagtaccactaaactttgcta
caagcgtacctcttgtatctccaatacttatgcctatgatattgaaaataaagccgctttattctgtaatggtggaaagttagaagctgttactgct
aaaggttattatcttgatagcgctgctatggtcggatccaaaaatccatacattataaaatgtgatgacagtgaatgtgtacacgaagctccaa
cagtttcatcatactatataaatgctaataccagtagttccaataaattaatttactgccataattctaattgttataccattgctgcctcttctgga
tactatgtatttaatcaacaaaatggtatcattagttgtacttcctctacttcctgtaccgaaaaggatgctaccactattggtggtaatgctcact
ttgttaacgctggagtagacaaaagaaccaactctttaattttctgtaatgaaaagacctgtgttccaaaggctgcaagaattggttactatttct

ccagcaatgtctctaaacttatttactgtgaaagtggtaatacctgtgctgaaattaatccaactgaaaactactactactctgctgatactgcag
aaagtaagaattatatcattaaatgttcaaaggtttctgcctctattatttgttctaaggaacttgctgacactggttcatacttaaccagtaaaac
taatgtattaatttcttgtaccaagaacggaagctgtaaacaaattgctgccaagccaggttactaccaatctgctgtaaagattactattaact
caaagagagatgtttctaatgttgacgaaaatgaaaccgtcagtgacatcgccggaagagattctactactacttacaatattattgaatgtac
caccagtaactgtgaattattatctgctgatgaattaagcgccattcctgtttgtgaatacaacagtgataagtgttatattacaaacaaatacgc
tatgggtaagagtgctgtcacttctattactgctggtaatctctgtacaaacgctgatcgttcaaagttctactttgccactgatactattgtcgttg
ccccatctgttattgctggacaaactgcaacatacatttacaccaccaccactactaactgtattattgctgactctaaatataaaaactattact
acactgtcggttccgatatctaccgtattgacgatggaacaattagccgctacgttgaatccggttaccatttccttaatgttgataagaacacct
tggttagcgaaaatactattgaaaattacaataatgaaagtgttaagctctacaaatgtaatggagtttcttgtaagattatggatgaaccaaa
agataccacttacttcgctgatgttaataagagaattataaagtataacgtcaataatgatgcctacaatttcgcctatgaaaaagatattattt
gtattttcgcaaacaacaaatgtactccaaatgctgatttaaatagcagagaattctgtattacttacaagggtgaaattgctttagctgctgctg
atattaagaatcgtgaaactggtgactgttacaaagctggttctataaataacaatatttatggattcagtcaatacttatacagaatggacgtt
agctctgctactcttgttgataagaatggttaccatattgtcagtctttcatctaacaacactgttgctactaaggattacaagaacagagttatta
atactaactctattaagatttacggttgttacaataccaactgtaaggtctatgacccagaagatggtgtttactactacgatgaagaaggtaag
gccttattaaagaatgaaaatgacgtttggactgttccagaagtttctggttacgccttagtttcaattaatccaaatgaaaagttcgtctacaag
tttaagaaagatatggatgaaattactttattatccaaggcctccactggttactactacactattgacaatgaaatgtatgaatgtagtgaaatt
gataatacctgtgaaaaaattgatgaaagcgattactacttcactaacactggtgaaatatactactgtgtttacgattctgaaaacttagaaaa
gactgaatgtaccaagcaatcttgctatgccggtcaaaattactacattggtggaaactactatagatgtgaagctggatcctacctcagtccaa
ttaagtcaagaaattgtaaatatgacgaaaatgttattattaacttcccaactattttatatgaagaattcccaggtcatattaagcaagctatga
gtaatgttgtaaagaataataattctactgctgttgctgttagatctaacaagaaatacatatctgttgttccagctatttacactaattgtacata
caatgttgaagaaactgaagctacctatgaatttgtctgtcttaacaactttgtctctgttaacaaagaagatgatacaattgaaatttgttctatt
gaaaaccttggctatgttgaatgtgttgatgatgattctaacccagaaaagtgtaacccaagtggtgcattcagcagaattgtacttaatgtctt
cagtgtaatcttcactgcccttgtttcattatatgttgttctttat

>Locus12584v1rpkm1.78

Atggtagaggtgttgagagtatccgtagatgtagttattgatgtttccggcacgataacattcaccagtttcacggttcttaatgtcagtggtggc
aagaacaatttcacccttgtaggtgatacagaattcttggttcttcaaatcagcatttggtgtacatttgttgttggcaaagatacaaatgatatc
cttttcataggcaaaggagaagacatcattattcacgttgtacttcaagattctcttgttgacatcagcatagtaagtgttggcatctggtttatcg
atgatgctacagctacttccgttacagcggtagagtttgacattttcatcattgtaggcatcaatttcattgccactaacaagttcgtttttggcga
cattaatgaagtagtaaccagtttcgtagaattgaaggatactaccttgatcaagagtgtaaatgttggatccgacagtgaagtacatatcact
gtaggaatcgttgacttcaagacagttagagttggtggtagtgtagacataggtggaggtgacaccagaaatgacatttggcttaacaacgac
ggtatcggtagcgaaatagaagacggaacgatcactgttagtacaaatgttaccagcagcaatagaagtagttgcagatttagtcatggcata
ttcaagagtaatataacacttgttattgttaaattcacacattggaatggcagc

>Locus12584v2rpkm0.00_PRE

Gtagaggtgttgagagtatccgtagatgtagttattgatgtttccggcacgataacattcaccagtttcacggttcttaatgtcagtggtggcaa
gaacaatttcacccttgtaggtgatacagaattcttggttcttcaaatcagcatttggtgtacatttgttgttggcaaagatacaaatgatatcctt
ttcataggcaaaggagaagacatcattattcacgttgtacttcaagattctcttgttgacatcagcatagtaagtgttggcatctggtttatcgat
gatgctacagctacttccgttacagcggtagagtttgacattttcatcattgtaggcatcaatttcattgccactaacaagttcgtttttggcgaca
ttaatgaagtagtaaccagtttcgtagaattgaaggatactaccttgatcaagagtgtaaatgttggatccgacagtgaagtacatatcactgta
ggaatcgttgacttcaagacagttagagttggtggtagtgtagacataggtggaggtgacaccagaaatgacatttggcttaacaacgacggt
atcggtagcgaaatagaagacggaacgatcactgtt

>Locus6670v1rpkm6.81

atgattccaaataatagtacttatgaagctggcttcaaaactcagtttttactatagttttaatgtctagtattttaatttttggagcatggactctt
actaattttattataaatattagtaaaaatgaagaaaatcaaaaaaagataaataaattatatcaaaaaggattagattatttaaattatctatt
tgaagatttaaagtatattgaaaatgccattaaattctttcaaaaaaatttcagacataaacaggtttatgaatctgtagcacgtgaaaatggtt
tattaccaccatatattgaggaaaagccaaaaactatatattgtgcggaatattatcaaaaagttatagagaaacaaaaatccgctaaagagt
tatattccgatgttactatttatcttactttatccaatggaatgactcttaactatataccaaaatatccatggtattgggatggtgtagagctcttt
actggcattcaacctaattcaaagaaaatctataatccagaatatttatctcctgactatgttccaatgtttaatatatttgatgataataataata
ataataatttttttctaggatataaaactcaagatgttaaagctgaacaagaagctatagaacataatataaacaatacaaactcatcatcaga
ggatgaaaaaaaattatcaactcttttattaacaaaaaataatagtttaaataccaatacatcatcattttcaccagtactccttaccataaatg
aaaataactgtgaagaaactaaaaatcaagaaactattcatgttgaaattgataccaaccataatattgatatccctgttgaaaaagaaaata
aagatttatttgtttcagttccttcagaatattcaattgaaaattttaaggaaaatgaatcaaaaaataatcttaaattatctaattctccttcaca
aatacaagctaataatgaaacaaaaattaaacattatataagacctacaaaatcaacaaccaatttaaataaggaaaaaatgagttcaataa
aagaaaattcatcgtcttctcctttaagaaagtctaaatcattaccaagtcttaaaaataaagataaggatcaacaaaaacaaaaatatataa
aaaaacttaaacataaaaaacaaccaaaaattatatatccaagtataccagttaatcctgatttatcttttgtacccttatataaaaagaatttt
aataaaactaaattaatatctaaaataaaacattctgatatggaaaatacttataaaagattaaattgtaataaaaatcta

>Locus3185v1rpkm22.91

atgaaattctacaacgctttattattattagctgctactttatctttaactcttgccaacaacttagaagtttctgaccttgaagaaaacgatgttg
ctggtactagtcttgatgctggttttggtgaaagttctgaaccagaaatagatcaaaatgaagctccaattgaaccaccaacaattgttccacct
tccttcccaaacactccaagtactccaccaagtactccaccaagtagtggaccaagtgttcctccaccaagcactggaccaagtgttccaccac
caagtagtggaccaagtgtcccaccaccaagcactggaccaagtgttcctccaccaagtagtggtccaagcgtaccaccaagttcaccagtta
ctgataacacttcagcagatagtgttccagaaagtccagtaagtgacaatacttcagctgatggtattaatgctaatggtgtaactgctgattct
ggtgaagcttctgatgattacggtaatgaaagcagcattaataacattgataacactcaagctccagctgatagtgctgatgttagtggagata
gcgctaatgctgataacggtgaaggatctgatgattatggagaagctgataatagtggtgtcactgataatgctggtgttaactctaacgatgtt
acaaatgctgatgctgaaggtgaaaattctgctgatgaagaagaaagttccactggaacaaaggctgctctcggtattgctggtgctgctgctc
tttcttcagctggtatcttcctttgggttaagagatctaagcgtaatgaaggttacgttcaaagtgttcgttctcaaattactatggtt

>Locus1571v1rpkm64.27

atgaaatactacaacgctttatttttattatctgctttatctttaactcttgctaacaacttagaagtttccgattttgaaggaaatgatgttgccaa
ttctggctttgatgctggtttcggtgaaagttctgaaccagatattgatcataatgaagctccattagaaaatccaattgttccgccatcattccca
aatacctcaaatgatactccaatagtaccaccatccttcccaaatacttcaaatgatactccagtagttccaccatcattcccaaatactccagtt
actgataatacttctgatcaaacacaaaatgatagtattccagactttccatctgttgattctgaagttccaccagtaactgatagcactggtaat
gatagcactagtaatgatagtgtaagtggatcagataatactccaagtaatgaagatgccggagaaggttcagatgattatggtaatgatgat
agtattaacaacattgataatgctcaagtaccagctgataacattgaaaacgctagtgttagtggtgataatgttattgctgatactgatgatac
taatgctaatggtgaaggttctgatgattatggagaaactaatgctaatgatacatcagacagtgctaatgttaacaataatgatgttaacgca
gaagctgaaaattcagctgatgaagaagaatcatctactggtactaatgctgctcttggtatagctggagctgctgccctttcttctgctggaatc
ttcctttgggttaagaaatcaaagcgtaataatggttatgtacaaagtgttcgtactcaaattactatggtt

7.4.4. Co-regulated sequences from A. robustus *glucose pulse experiment*

>Locus3686v1rpkm10.56

atgggttatcatgatagatgttttgaagatattccatatttagaaagcaaggatccatatcataataatagacgtgataggttaactgaaagtgc
tataaaccgatacgctgaaactaaccctgaataccaaaaagttttagataaatattatagacctgataaaaaactaccaagagcaataacaac
aacgaagaaatcagcagtagtaaaagaatcttcaacctcaaaaattacaaattcaacaacaaaagaaattccaatgtgctgtgattcccgttg
tagttcagcttcaaactcttcagaaagactttgtcctaactgtaaacaaaaattttacgcttcgaaaaaaaagactagcacaattaaagacaaa
aaagaaaatgattctaatttaatttattggaataaaaataaatttgatgataatcatgtaccaactttcccacccagaccaccaaatgaatgtac
atcccttattttgccagccatgtttacggaagaagaaaaaaagtattatcaatcaaaacaaccaaaatacccaccagttgaaaagattacttca
gttatagaaaataaaccaccaagatcaaaaccatccggtcagtataatataattacaggtgaagaattatatgaaattgca

>Locus6387v1rpkm3.43

atgaaaatatttcgtaatagacccaaaaaaaataataaaaagtttgaaaataattttaaaaacgaattttatatagcaagtaatgaattttcaa
aacaaaaattaatatcattttatggtattgaagattgtgctattacaacaggacttggatcatgcatgaagcttgaaaatatatttggagcttcca
cgatatcaagaggaagtaataatactttagcaaattctaatgagagtttagatgatgttgataaagatattatgttatcgaaaaaaggttttaag
aattacaaatctaataatagccaaatgaatattgcaattgatgatactattagtgagcttaagaaaaaattaaaagaaagaaatacaaaaata
ttaggcagtaataatacgattactattgataatgaaaaaagttatagtgaaactaatcttgaaaataatttaacaagttctgatactatctatcgt
caaagtttagaaattgataattattattcagaaaaagaaagaagaggaagttcctctactagtattgattatttttcaagtaatccagatgtagt
aattgatattaatactttaaaaaaagaaagtaatttatcatcttatttaagtaaaagtaatagtaattctcaaattaataatcataatattcacga
tgaagaaaacaaagatgaagttcctaaaaaggaaattactattgatattaataatgggcaacaaattatg

>Locus5098v1rpkm5.63

atgccattatttagatcagcatcaaatttaaaccctgattttcatagtcccataaatgatttacctattgatgaattaacaaaaagtgaatgtaaa
agaagactttcatttaagaaaagtttttttacatggtcaactcaaaaagatttttggataaaaataaatttgtatttgattccgattctgttatga
aaagtcgtcataacagttttgctataaaatcatcaaaatcatcaaaatcatcaaaatctacttgttctggttcatctataaattatgatgaaagtg
ataatgaatattctagtgatagtgaaggtgatttaaagaaagatggttacttctctatggaagatattgaaaatattgataaaatttcaggttta
aaaaagacatctaaaaataaaagacaattttcagaaagtacattaatttattcttcaaatacaaatactataaatactgtatcattagaatttaa
aacgaaaatttctaataatgatgatataaattatggttttgaagataataatcatattaatgaagctgaagatactggtgataattctggagaaa
attctggtaattcttctaataattcttctaataattctaatggtaattctgacaatactaatgaaaataatgacaacaataatactgataataatat
tgatatatctgatataatagatgaaattaataatatcaatatagataccatagacgatattgataaattgaatcttaatgagttagaagaagaa
gatgaaagtgatgatagtgatcttgaaaatgatcttgaaaatgatgaaattgaaaaagaaaaagataaattgacagattatcctaataatagt
agttataccgttccaaaacctatatttaaaacaaatagtactactaagaagacagttacctttagtgatgatgtagttatcatagaaccaagaa
aaccaagaaaaggtaaaaatttatttaaaagagccattttaaaaattatgaaaaagaaagaagaagaaaaaattgaagaa

>Locus9848v1rpkm1.17

atgaaatacttagaaaaaggaattttaattagtgatattcaaaatagatatgctaatgatttctatccatcacaaaatataaaatatcttaataa
tcaaaatgattttagttcattttcaagtataccatcccaaatacaaaaatcctatatacttccaaaagaaaaaccatcaaaaccaatagaaaaa
ccattgaaaagtcctaaaagattaatttgtacaaacgctacagcaaatagattatatctatcccatactatacaaagtcaacaaagaataaatg
gtcatattaataataaacaacaacaacatcatcatcaaactactaataaaaatatatatttattaaatgatttagttaatgtatataatggtttta
cccataaatttgtaaatgaaaacccatccattagagcctataaaactataggacatacaacactttcatatgcttctattatttctatgaagaata

```
gaccattattagatcgtaaattaataagtggaagaaatcaaaggaaaaagaaggaaaaacaaaaaagaaacaaagataaaaaaggaaat
aataacacatttaaaaacaaaagtataaataataacaatatcaattacaattataataataataatgataataacaaagacgaagatataat
aaattatttaaaacctcttagttcatactttaatcctaaaccagttcaagaaaatgttgaagaagaagttaaagttgaagtattaagaactctttc
agttagtgaaaatggaagattaagtactgaatctttggaaagtttaaaaagcattgactactttaaaattgaagatattattaatgaatatttaa
taagtccttcaataccagatattcccatcagatatgaaaatggttcaaattaccttaattgttttaatcatttatatgatgcctttatttactttcata
atagaaagttattacaatatattgcatggattccaataaatgataataaaaaatatctttcagcatcatctcttttagctaaaaaacatgcccca
attgtatttaaaaagacaattacacaccctcaaagagtttcttctttatacgctaattttattcatactgaacctgaagttaataataataatatta
ataataataataatgataacaatcaaaatgtaaataaaccaaaaggtaaaaaaaaaaactctcgaaaaagcaaatactccaaaaaagtcaaa
tgatctttcaataaatactaaacaagatgttaatgataataataataataatgataataaatcttcatcatcaccaacaaaacaacaagattttt
ctacaccctctatatcaaaaaacaataaatcattacattcaaaagttttagaaaatataatgttaaataatccatcaaaaaaagaagaaaattc
taataacatcattaatgaaaataataagccaaaacaacaagataataaagttgaaataaaatcaacaccaaatctacaaaataataaaaat
gaaacaaaattaaaattaaaaccacaagaaaatactactgaattaagacataaaaaaaaaattaaatgataataatgaagaaataaaatcta
aaaatgaaatgaaatctaaaccaaaaagaagaaaaactaaaactgaaataaatcctaaattgatgccagaagaaattatagaattaacacc
ccaaacaataagatatgaaaaaaaatctatatctaatcgatctccagaaattaaaaatgaaaaaaaatctactaataatatggataagaaatt
caaattaaaaccaaaagaaacaaatgaaataaaacctaaatccataacgaaacacaaaagaaagcattcaaataaaaataatataactga
tatgaaatcaaaattaaaaactaatgaaaataactataattataataaaataatacaatcacaagaaaatataatagagaatgataatctag
accttcaagatagtagaagatcattattatcaataccttcacgaattccaattttccgttcttcaaatattgatccaagtagtattaaatttaatga
acatcaaaatattcaaaatcaaataatatcgaataaaacttctatgattcctacattaggaaaatttgaatttccattttatcatatgtatgatgct
tatgaaacttatggtaaagctttattaagagctccaagagcatggagacaattaaaaagaaatgttgtttctattgcatctatatattcttatgat
aataaacca
```

```
>Locus1925v1rpkm32.64
```

```
atgtctgattcatctttcaatccaagatcaaatgctcttaagaatattcttaaaaacccaaaagaagaaaaaaagagttttggtaaagaagaat
ttttagctagacgtagatatttccttgaaaaacttgaacaggaaggtggtattccaaaaccaaaaagtttaattgataattcttctactaaaaga
aaaccttccaatgaaaaaattaaaacttcttttaattataataataaaaattctacaacttccattaaaaaagaatcaataaaaccaactacttc
aacttcaaccagtagtagtaatactattaaaactaataaaactgaaaaaattactttaccagctcatgattggacaattgcttcagctaaatctg
aatctttaaagaaaattattgatatggcagaaaagaaggatactaatactgttggtaaaagtgtatttaatgcaagacagaaattctttgctga
ttctcaattaaatagtgttgatcaattaatgccacgtccatctactcctccaggttatgtttactttgaaagatgtccttctccttcttatgaaccttt
aaattatagaagaaattcttatggtttatattcttcttcttcttctactactgctctttctccatctttaaaaccttctactccttcttatccattagtttc
tagtccaactccttctgctccttctcctgtaaatacttattctttaccatgggaacatggtaataaaaaggattattctaaagtatcaatgaaacat
gaatctaataataataataataataataataataataataatagtaataataataattcatctaatcatcatggtattgaaatttt
agaaaataaaccacttatatctgaagaaattattccacttcctgaacctgtttcctttaaaaaggaaaaattaaatatagaaaagagtgataaa
cctttatttattgaaattgaaaaaatagaacctttaaaaccaatgtatccaaaaatttttgaatctttaccagaaccaattcaaaaaagtattata
actttaccaatagataaacctcttactgttaatattattaaagttgaggatattaaacctcaatttacaaaagaaatcgttaatataattgaaga
aaaacttccagaacctattaaatttacaaaggaaaaattagatgaacctattgatttaattactataatggctccagtagtagaatctcatctta
caaaaatgatttattttatgccaaaaccaattacttttacaaaagaattattaccaaatgataaattattagaaattcaaactgaaactattcaa
tgtccaattaaacaattaccaaaatttattactattttaccagaacctgtaaaaattacaaatcaaacattaccacaagataaacctttagaaat
tattactgaaaaagtttctttagataaaatggaaattccaaaaattattaatatttcttcttctactactcaagtaaaaagtaacgatgatcacat
ctatggaagaaaaaaatataaatacaatggatttaaataaaactatagatatttcttctttaacccttgaatctaatattgaaggaaaaaaggaa
acaattaaaaaaattaataataataataataatgatgacggtttattaaatacttattattcaagtgaaactcaacctttaattgttctcgat
aatgatactactgatattccaaatgaaaatcaaccattactttcatttaatgataatgaagaattaactgaaagtaataatataaatgttactga
aggaactttaatagattttgatgaaaattcattaacagaaccaaaagaaaaattaactaaaaaaaccaattctgttagttctataactgaatta
ttaaataataatggtactttatctttagatgaacctcttccaagaaaatctattccaatgtctccagattcttataataatgaaaataataataatg
ataatgaagatttaatcgaattttctgaaagtgaaatgtctgataaagttaaaaattggtataatgttgttgataatgctataaataaagaagat
aaagtagaagaaatt
```

>Locus10381v1rpkm1.01

```
attaaagcaaatgtttcaattattggaaataaaaatggtactatatttgattttaaaggggatcatagaggaagagtttgttttaattttataaaa
acagaaaatgttgctgtcaaagtagaaaatctcatattacaaaatttttctactcaaggtgtatacattgaaaaagaaataataaaagttattg
ccggtaaaaatgatttacaagtaatatttaataattgcgtattcagaaataatgattatcatatcattcaatatgatatacatacaaataatggtg
aaattttccctataacacaatttatttttaatgaatgtgatttttataataataaggaaagattaatttttgtttatcaacatagtgatatcagagat
tatagaagtattaatatgaatgttactaaatgtaattttatcaataatagaggactatttttcacctttgattcttatttaacaatagatgactgtta
tttttcggatgtagatagagattccgatctttcatttgaaaacgtgttttattattcacctattccatcaatatcaaaagctgtattagatataaaaa
attcaaagtttgaaaatataaatgtcaaaagtgaattaccgttgattactgctaataatcttattttaaagattgaaaatacgtcatttaagaatt
gttattcagcatatggttatttatttaatatattctataaattttcaatcccacatataaatataaataattcaacatttacagaaacatcagcatt
atttcgaggaaaagctcttaagttaatgatatcagataccaattttataatattacaataaataaatatattcccttattatcagatgcaaaata
ttcttctgttgttgttgttaattcaaaatttgataaaataagcttaatgaatggttttgttaatgaagaaacaagttgttcattttataatactgatct
taataatataaaatcaagttcaaattcacttttatatactaaatatcataatatatatattgatggaatgaatattgaaaatgtttcatgttatggt
gatggatcatttatattgtttgaaactggggatgctgaaaatagactaacaataaaaaaatttgaatataaagaaaagtaatttcaatggtccat
ttattaaattagagggaaattatggagaagttatatttgaaaattcaaatctatctgatgttaatacttatggttccattattaaagataaattag
aaaaaata
```

>Locus2800v1rpkm17.14

```
atggattttaattatttatatcaaaaaagattagataataatcttaaaaatgaaattgccgtatttaatggtaattattgttgtttaagagaaactt
gttttacaatacatgatgacatattatcatcaattaatggtggtttttcaattattggagaaaatagaagatttagaggtaatttccaaaatttaa
atatttcattaaaagctcataataatattccaatcttttatcttcaaaattatttattacaagatataaaaaaagtatatagtaatgaaaaacaaa
ataatgtaccatttgctgaaataaaacaatgtgctcttggggcttatttaattaaatttaaaaatatagccaccaatgaaactgattattttgaaa
tgattagtgattataatttccaaatttgtaatatattttatagtaatcaccaaaaattagataaatcttctgttgtttgtcaaattttaagaaatggt
gattgttgtgatgtctatatagcttcaggtgtagaccatgtatttatgttaggacttgcttcatttttcttttgcagagatatatttattaataataat
gaagaagttaaaatagatattgataatgataataataataataattataatgataatttatcttctatgactccattagtaaataatgatgtatta
aaatctcaatacaatacatacactaggataaattattgtactgatgaaacacaatcattaaaaaaatctaaaaataaaaacaaaaacaaaaa
caaaaacaaaaacaaaagaagatttttgaatttttgaatggttgctgttgtggattacttctt
```

>Locus10500v3rpkm0.11

```
atgtttaaagaaaaagaagatagtaaaacaatgattcacttattaagtaagaagccattttctttttgtaaagatgatgatgaggaaaagtcaa
ccccaacttatgaaaccactccaagtattaataaaaataatgaaagacgtcctcctatttacaatagaaaatcaagttcaaatattattattgat
tgtaatatatatcagagtagagatatgcttaatagtagtaatagatctgaagaaagatatgaaagaacagatggtatgaaaagatcacttcgt
aaatcagatgtttctatttcttctattactccaatagttttatttaatatagctaaatctaatggtacatttgctgaaagtttagttacaaaaaatgg
ttttgctattggtaattattatcgttattggcaaaatttatttgatttaagaccttgtccatctatttattctattacgataacttatgaaatattagga
aaagaagaagcaagaaatttccttttaaaacattttgaagaaaaatttatatggcttcatgatagttttccattaggttttatgtataataaaagt
catggaccaatttcacctccatcaactattgaaatgatgtatgcttttggtgaaagatcatggattgatggtcatgcagaatatttagtaagatat
acacgtaatgcaaaaacttatatgaaagatttaagtctttgttcatcagataatttagatgatttgccaccacttgatatggatgtattaattgag
gctgtttcaaatattatttat
```

>Locus4150v1rpkm8.57

```
atgacagttacttcaccaaatgaagaattatgtatcattgattctaaatatgttaatgaaaaggaatctttatattatttaaaatttgattatattc
ctagaactgaatatcttcttcaagatagtaatcataatgatattttcaaaattacctattctggattatgtacatatgatactatctttaaagatatt
aaaagtggtcaagaattatataaatgtgaaagaaaagaacatttgactaaatctaatgaatttaaaattaaagatattgaaaaagatgaaac
aatttttgaagctaaaatatcaaaaaaaaaattcattaactcattataaatatttagttacttttaccaataaagttacaggaaaagatgaatccc
ttgaatttgtattttcatttagtggtcaagaatgtaaagtttattatggtgaaaaaaaaaaggaaggtaaattaatttgtcaatctagtaatataa
gcaaaactgcttatgaaaacaaaatagaaattgcttcaaatgtagatactatgtttatgttaataatttataatgaaattgtccatagtcattatt
tggaaaatgctactatgcaaggtgttgctcttgcaagtacaattagtataatgagt
```

>Locus3804v1rpkm9.92

```
atgacttcaatattaggaaatcaagataacagtagtaccgaatgtattactgattatagtcaagaatttaaaatagaaatggaaaatagtattt
gcaattatttttcgaaatgcaatttagaagaagataaacacgataccaatcgttcagaaactttaatagataataatgaaatagaatatatttct
cccgctccttccttaaaaggtatttcagactttatgataaaccaaattttttcaacagatgaaataatacctatagaatatccatcggatgatgaa
gatacaaagattaaaaatgtaaatgaagaagtaaaaaatgataataataaaagttcagatactaataaaatgagttcatgtaattctgatgtt
actttaaatgttaatgatgtaaatgttactgatagtgaaaatataaatattactgatgtaaatgttactgataatgaaaatataagtaaatctttt
gctgctgaatttattgatttacttgatattaatatgagagaattaaattccaaaaaggatgattttcatgttattggattaaaaaaggttgatgaa
gaaaaagaaaataatgaatccaatagtccaacttcaataaattttaaagatgatactttagataaaaaacaaaaaaagaataaacgttcaaa
acatatttcaatattcacattaaatagaaaatctagtaattcacaactttctcgttctaaatctacaaatattaaatcttcatcaaataaatgtaa
aaatacttcatcagaaattatagatggtagtaatgataatggaaataagcttaaaaaatctaaatccaaattaggtttatcaaaaatttttaaa
aatgttaagaagagaatgtcacatcaatac
```

>Locus5926v1rpkm4.06

```
atgagtaatagtgaaattaatgaagaaaatttaaatgacccagcatggaagccattaaaatccaaaaggcaaatttttgaaggttcctatgga
caatttccaccaggttctatgtcaaacttagaaggaagcagatctaatagtttatctaatattaaaaaatttgatattgaatcaggtattgttgaa
caaagaaaaattgcggtaattgaatcaaaagattcaaatattgatatatcaaaaaaagcagaaccaaatattgaagcatattgcgcatcctca
gaagcaaagcaattaaaacaattattagaacataactcaacttcctcaccatcaacaaaaaaaaaaatattaaataagaaagaattgcctgt
cccaccaaagaaaaatattgcaccaccaccaaaaccattaaataaattatcagataatgataatgaaattgaatatactagtgagaatgtaaa
taaattgtttaataatatgttccaagaaatagaaagtgctattgaacctgaatcaccaatatcagatgacatttttaaaacaaaattaacaccg
aaaaataaaccacttccagaaaaacctaattcagcatcgtcatcaccaaaacttccactaaaaaagaaaataccaccacctgctgttccttcta
aacctactactccaattgtagataattcttctccaaaaacttctaaacctggaacaccgactatgtcaaattcaccaaaaatttctaaattacca
ccaccaccaccagttagagaaaaaccaggaaaatcttctaaaccacaaactcctgttaaagatactcctcctcctcctccaccaaaacaatca
cgtgatgagcaattattatcaacacaattaccaccacaagtaccagttaaaaagacaacaccagcaccaccaacacctaaaaaaacaccaat
taaaaaaacattaccacctccaccattgccagtatcagtcccaacagatcaacaaattaatgatgaatcttcatcaccaacttcagcaacaaca
aatgagaataatgaaattttatctccgacttcagctaaccatacaagaacatcaattattgcgttacagaaaaaggcattaatgttaaagaata
aaaaaagtcttccaccaccaccacctccagttcttaatgataataatgataatactaatcaattatctcgttcattatcacttaatgatggaagta
aaagaacttataataaagatttttcaccttccgtaccatgtacaccaactactaatataccatcattcctccaattaccaacaggatcaccaaaa
gtacaacctaccaatgtaccaaacttatgcgatttagcaattcgtgaagatcaagaagttattactaattctaataattcagataatcttccagtt
aatggaactggagttggaggaggaggattacctagacgtagtcattcttatcgtaatcatgctgtttcaatgtattcatcatcatctaattcaagt
cttccagctccatcattccactctagaacaagatcaattacttattcacatggaaaaggagagtctgaatccgaaaaacttccagattccttcag
aagtttctttaatgaacaagaaattaatgaaaatgatatttcatttattgaagaaaataaaaataacccagaagaaaagaaagaaaaaaaag
ataatgactcttctgttgagaaaactaatgcaagaggatttctctttaatcaagatattaaaaaacaacgccatcaatctaaaagaatatcaac
tactgatgttcctgactctatgaaattattatttggagtttttggaaatgaacaagaattagattctgattctgataatgataataatgattcagat
```

ggagaagaagaaatactaactcataataaaaataataatagtaatgacgtatcttcaacaattattcccgaaaattcatgtgaaatatatttaa
atgaatcttatacaccaaaactttatgtatcctctaataaagataaaaactgtcgtaatagtagtattattaatgcaagtatcgttccacctaatg
aaccaccaattccttcattagataaaaaaccatttagaggaaacactttaagatgggttattaacgtcaattgt

>Locus4155v1rpkm8.56

atgggaagaaaaggtagtatgaacttacaagttatcactgataactttgttaaagatcaaagcaggagaagaaatgttcaaccagcaagtgct
agacttccaacaacaccaacagtaaaagctccaacaagaaaaatgtcatctcctgatattgaattaacttcttctccaaataaaaaagatgaa
ggaacatcatcattaaccattgaaacgattccaaaatcagctataccaaaaggatctggaccaagatcaggacttccattatcagctgctccaa
aatcagctgctccaagtttaagaatttctacattagttacagctggtaataaagctagtggtccgaaaactccaaaaactccaaaaactcctaa
aactccattaagtgcatctgttaaaaagaatccaaacttaactttacaaccaccaacatctagcattccagatattgatgcaccggatacccca
gtcaatgttatttatgcaaatgctttctttgataataaaactcgtgaagatgaagaaacagaagctattaataaaaaagttttaaagcaacaaa
gaaaagaaaatattcgtaatatgcgtcgtgaacgtatgaagcaaccaaaagatattagagttgcattaattcttggccgtttatttgatgcttctc
actatgaattaggtccattatcttattggtctgatccagaaattgaaaataatgaagaattaaaacctaaaattgttgaagaagataatacacc
aaaaccaaagaaaatgattcaaattaatgaagatgaaattgaaaaagatattaaagaaagaaatactgtaatagaaaatcatgataaaact
aataatgttaatgctcatccagcttgggtaccagaagaagctggtagtagtaatatttcatgtagaaaatcagctaccttagttgatggcaaact
tgttaaaaaaggtagaaaaccatatattgctgttgcttcaatttttgccaaattatatgacccagtattaatgaaaaagaaggaagatgaagaa
aaactcagaatgatagaaagacaaaagcaaaattcaattattactgaattaaaaccaatgggaaaaagtattaattctagaaatcaaaacaa
ttcaaaccaaaactatcaacaacaaaataatggtaattataataatagaagccaaaatggccataattatcattatcaaaatagacaatacca
aaattacaataataaaaactttaattcttacaataataataataaatttagaaatgaaaataatagaaatgaaactttccaagttaattttaga
gataatagatcacaaccacagtataatactagaagttatttatttaacaacagaagtactaatcaacaacaaaataataattatatgaataata
attataataaccagaatattaattatggttatgtctaccaatatccagtttatcaatactacgacccaaatatggttcaaagttataatagtaatg
ttgattatgctaattatgattcaagttatgaacaaagttatagtaactactattataattcaaatggcagtaataattatcagaataatagattta
agccaagaatgaataatcaaaatagtaatcaaaataatagaaataataataatgaatcttttaattatagaaaatatgaaaagaaaattaat
attaaatct

>Locus4459v1rpkm7.48

atgagtgaagaaaataaccatcaaactcataaaactataaatcagaattttggtgtggaagtaatacctgtttcagaagttcattgtgatgatg
gagaagataatatttcatgtactccatatggaagcggtgaaatttcaaataattccaatgcttcattatcaaatacatgtgactcatccggttca
actttatgcgttaatcataaagcaaacccattagatgatcccattattcaagaaatgtttaaacaaagaaccgtgaaaatggataaagttgaag
aggaaagaagaattagaatgtataccatgatgcgttatagcaaaaaatataaacatcttcaaactgaaagagctaatataaaatgggataaa
atgcgttgtaatgatgaacatgaagatactgaacgttcattaaaaggaatgaatggccaacctcaaattgtaccagatagtgaagagaattca
ataggtaatcgtagaatgcaaagtcaaagaagtttaaatgctaaacaaatttcttgtaatttcggtattgatttacatgccttagatttgattgat
atggaattaaacaattatcaaagtaatagtagtgataatagcaatgatgatatatcaaaaacaaattctaaatcagaaaataaaaatgatga
gaatgaagataaagataatgaggctaatgataatgaagaaatatatgaaaataatactactaaaaaatatttcaataataatgaaactaattt
agaagatataacaaaatcaaattcaaaacttagtagcagtagcagtgaatacagtaataaagaatttataaaatctcattacaaagtcaatag
tcgattaattaattgtaaaaagaatagttcagtatgtcgttttgatgataatcctgtactaattaataatttcttatatgattctgaaatagatgaa
caagataattctattaaaacaagagatgatactcttcaatttaatgatgaagatgaatgtgattacaatagtggtattgatgattatgttaatagt
aataatgataacgagaatgataacgataataataataatagtaataataataataatgataattgcaataatgtatttcaaaagttttgta
attctgaaacaagtaaaatacagaacaaaaacaaaaaaaaacataaaaagataaaattatttaaaattagtcataaaccaaagggtactat
agaaataaatgatgcaaatacattaaataaaattaattttgataaaggtagtaatgataacaatagttatttcgatgatgacaattcaaataca
ttaattaattcaattttaaatattcagtcgaattcttcaatcaattcaaatagtattgatgatgatttaaatattaatgtcacaccaaaattgaatg
ataaaaaaaagaaacataatggaataattaaattctttatgggaataaaacatcaccattcaaagtcgaataatattaatccagaaaatcaag
ataataataaacaaattattattgataaacataacgaaactttcaaaaaaactaaatcattgagattctcagactcagttaatataattcca

>Locus2461v1rpkm21.07

```
atgtcaactacatctgatgatacttttccaaatgaaaagagacgctctttaacttttgctgataagtttgaagttaaatttttctttaaagcagaag
ctccaagagaagttgcttattctgaaactagcttcactactggtagtgaaagctattatgacgatgacgaatattggtatcaagaagattcttat
gatgatgaagaagaagaagattattatgatgatgattattatgaatatgatgattatgatgaacaagattatgataattatttacctcaaagaa
attatcatcattatactactgccgaatatgatgattactatcattatgattataattattataataatggtgtaaatctttccattcaatattctatga
atgaatctgatgatagtgaagaagaaattagaggtcgtcgtagagaattatttggaagttttagaggaggagaagttgaagtagaagaagaa
gatgatgatgttcaaaatgcaaatgaaagatcaacgacaagaagattagctttagctaaattagaaaaaggtttattccgtcgtcaacaattac
gtgattctgaatcctttagtgaaactgaagaatctgaagtttctgtttctaattcagaaacagatccatttgatgaaccattaccatctttaccatc
cttatcatccttaccatccttaccttctaaatcaaaaactgattttctttccgttgatcttaaacgatctaattctcctgtaactataaaggcaacat
ttactacaagagatttaccaatttctattcctatttcgaatcataatgcttataaggtaaaaaataatttaccaatacttaatgaattttctgaaag
atttaaaccatcttccctttctttattggatatagaagataataatccactttcatctttagatgatacacttttatcttgtactccactttctatgata
aataaatcccctttagtcgttaatttatctacttccccaaaagaccatgttaaagatatgttaaaaaaatctttattttccaatcctgctacaccaa
taggaaataaaactaaatctcttgatcttaaatgtgttactactttaaattcaaaagatttagttaaatccctaaccagtaataaactttccaata
ataaatcttcgaatacaactaataatcatcctaaactttctaatttattaagtaatattcataataatactaatagtagaggtgttaataataaag
gtgttaagaaaagatcaagatcttattcttcaccatcttctctccaacaacttttaccttgttat
```

>Locus6328v1rpkm3.50

```
atgtcatcctataaatcaacattcttatctaattctttaacaagacaaaaattaaatcaaatgagtcaaaattcaagtatgaaaaatgggcatat
tggaaaatataatatatatctaaataaaaagaattctaaaaatagcttcaaagaagaaaattcaatagccaaatcctatgaaccaaaaagttc
cataaaagatagttcttgttggcttttatcattagatttaataaattttattcaatctctttgtattgggaaacttagttctcaagcaaaatggtctat
attagaattatatactaatgctaaaaatattaaaactaattttgattcaaaattagataaaaggataatcgaaaaatataaattagcacaaaaa
agaaatagtaaaaaagcagtaggctcacctcaagctatggagttatgtagtccaaagagtaataataatgcttctttatcacaatctctaccta
atagttcactctcagttgaactttctaatgctatcagtggtaataattcttgtgtattaactagaagtgaagaacatagagctaaactccaagca
gaattcttttattctcatcaaaacacactttcttataaatttgtacaagcagcacaacatttctataaaagtttagttacaagaaatactgttaaa
actggagacaaaaataatgatatatggccttcaagtgagtcaaaaataacaaaaatgaagacattagattcattttctaataatgctaatgata
atgataatgatattcaaacctttggaattagtattaaatctaaaaataatcataataataatttttctttatcaacttcattacctgcaaattccttt
ataaattattctatctctccaataaataataatataacacaagcttcttctacaatttcatctgatttaccattcaaacttgatgaagatattattac
agaagaaaaagaagatgaagaaaatgatgataactatatttattccgatgatgataatatggaagaagatcaaagctatggaaattgtatgg
atgaccaactttataattatcttcaaaataaaaatattaagacgtactat
```

>Locus3598v1rpkm11.14

```
atgacaatgaaaataataaataatcaaaagattcaacatagatcacaaagaccaattcaacttacaagacaatttagagaatctgttataggt
ggtagtaaatatactagtgttgatttagattattatcctacaattacaaaagatatctataaattaaattctgtgttgattagtataccttacaaag
aagaaggaagtagacctgaaactccatataatgaaaaaggagcttcacctttattatttgatccttatgaagtattcaatatgatttctaataag
aataattcaaattcttccagttttgaaaatttaacaaatataaaaaatgatgatgaaaatgattataataaattgaaaagtccattaacgcaag
gttttacaatttctgatgatgatgataacaagaatgaagatgattataataaattaaaaagtccattaacgcaaggttttactatttctgaagaa
gatgatgatgattcaagttctttaggttatcataaacattataaaatgtcaagttctgatactactttttctaatacaagtgatggttataataata
caagtcgtattaattcatttaaaataaataatcatatttttgaagaagcaaatactctattaagtataattgaagaaaatgaagattgtagtgag
gaaaatgatgcttctatgtatgatgaaatgttaactaaattagaaatatctatggaatatgaaattactcaaattcataacttttatgtcaatgaa
aaaaagcctattattaatgagttagaaagaagatgtattggtggaaactttgaccat
```

>Locus1765v1rpkm37.42

atggaagacgacagattaaatgaaaagagagctcaagaaattcttgatcgtttagataaagatgattttgtttttactatgaatgatttagaag
gagaagaatatgttactttaccaaatggtacaaaaatgttatgtgatattatgaatgataaaatagatggaaacttgacaaaaaagataaaag
atgaaaaatctaaaccaattgatgaccagatacttcaatttgatgattatataaatacatatgatatggaccataaagatgataataatttatta
tttaaaattgatgatactgaaggagatgaagcatttgacgaaactgaaaagtattatggtagtgatgaagaaggattagcaattaaacttacta
atgataaaaatgaatctgttccattatcatcaactacaaccaagccaataactattcaatcaaatagaccacaaatatcagctagttctttaagt
gttagtgatgatggtcatttagtggcaaattcaccaaatataccagtaactccaccaccttcattagctaaaagtcaaaatatgtttaaaatattt
aaaagtaaagatgaaatatatcaatctgatagatcatctcatccagtaccagatggtgaattaaatgaaaaagatattcaagaacttttggca
aaacttgttattcatcaagataaacaaaaa

>Locus2820v1rpkm17.04

atgtctttattatcacatttttcacagtctttattttcttctcttttcaataagaaaaaagaaactaataataaagataatactacaagagaaattg
atacaagttcgactaattctcttagttctattccatcattctctatattaaaccattccttaaccaacaatattaataataattataatcttaatgatt
ataattcaaccatagttacaagagaaaatatatttttggaaaagatattgataccaatgatacccatgaaattataaattataatttaaaaaa
attaagaaatggtcaaaaattaggtagtatggatgagaaagagaagaagaatgaagaagatataataattcatccaatagaagaattatca
agtgaaccaataattgaaagtaatattattgaaagagaattaattaataaatcgattcatactaaaaatgttaatgaaatagaaaatgataat
gataataataatgataatgaaagtttaatggaaagtgataatgatgataatgataaaactgaaactgaaattgaaactgaaatagaagatga
aaatgaaaatgaaaataccagtgatgaggatgatgatgaggatgaaactgaaaatgaaagtaatacaagtattgattcagatactaataaaa
atcaggatcagaatcagaatcagaatcagaataagaatcaacaagaaaaaagtaaaataaaatctaaacataattcaattgattcccttaca
gatagtattattgctgaagatttagatttagataaattaatttataatgtaccaccttcttttacaagaattaataattctaatataaaagaaggtt
ttccaatgttaagatttacagattctatgagtgtacctgcaagtatgacaaaatttcgtcaaactcctatgggtcgatcctatgctacaagaaata
ttttaagtttaagtcgttcttataattcctcacctaaaaatacaacgacaacgacaactccaaatacaaatacaacaactacaaatccaccctat
cttacaattcctggtattaaaacaatagaattacctaaatcatcattaaaatcaaatcatcatttaagaaaacgttctactttagctttatctttat
ctttatcaaatagtttttcagcaccattatctacatcacttttagctattcaacaaccatcaatgattaatgatgataatgatgatgatgatattca
aaatattactaatgattcagaagattcttatgatccttttgatgatgaagaagaagaatatgatatttttatatttaatgaa

>Locus8708v1rpkm1.60

atgacaagtattgaatattctagtggacctgataaggtaccattacatttttatataagtcgtaattcaacaacccataataattatcaaagaatt
tatgaaagagaaaagccaaaatgtccttgtcggaaatattatcatggagatgatagtgaatgtttacattgcaaaaattcaaaatcaaaatcat
cccctgaagaagatagtcgaggattgattcatgatttattaaaaaatgtaacaaaacaagaagtaaaaactggattttcaaataataaaaatc
catatgttgtttatgataaaaatattgatgaaaatgatcattttagagatgattttcattggttaactacatataatgataagtataaagatccaa
aatataaagataaaggaaataatatattaactattgttgaagatggttatacaagaggaattaaaaatatatgggatcctactatgcatactaa
agatgatgatataagtgtaatgaaaaaagattatacattaaaaagaaataatactagcatttgtacaaaagataatgttattgatacaaattca
gggtattgtactaattcatcgaaaattcatacttataaagatttagcagattatgtggataatgattatgattatatagataaagatagaaaaat
tgactcttattataataaagatgttttcccttcatatatagaagatgatggttttacaagaaaaggaaatataaatagttatatggatttagcacc
aaaattcgaagataatgaaaaaaagagcattaatttaccttcagttaatgaaaaatctactaaaccatttgaacaattcaaaactattacacaa
aatgattacactttcttaccagtaaatatttatgaaagattaaaagtaaatgttgataaaagtaacgctgataaaaatcaaatatatagtggatt
atcatatattccaaataaatctgatccaaatgattttataacagaacatatggataaatatagagataataaagaaagtaagaaagttgaaga
acaattaaaacatccgtttgtttgtgattatatgttagatgatggatacactaaaggtaacagaaacagtaaaggtttaatatgtaatggacatg
gaaagaaagaattaaaggatgaagatggattttatatttgcccatgtcaatatcaatcaagattatatttaaagagacatggagaattaactaa
actcccagtaaatccacaaaactctaaaccagatactcaaatatcaaattatcttaaaatgacaaaaaataaatct

>Locus3694v1rpkm10.53

atggataatttaccaccaccaaaatatgaacatataatgaatatgtttaatgataataatagtaataataataatgtaaaaattttaacacctca
gagagtatatgaaatcccaagaaataattttattacagtaaaagattctatatttggagaagcttttattaaaaataatcttgatgagttcaaat
ggtcacaggaagatgaaaaaattcaagaatttattataaattttctagaatctaaaggaaatattatgcctactaatttttcaaatgattttatta
aaaattataataaatttgaaactgcattagaaaaaacaggtttatataaatattttgaaaattcaagttttggaaatcattggacggtagttaaa
aatttaaaatcatattataatgatgtactttgtaatgaagaattatgtaataaaacatgttctttatattttcgtgcatttaatgaagcattaaaat
ctaatatttatgaaggtgattttgaaataacaagtagaggaaaaattgctttaaatggtaaagatataacaagaagaattattagtgaaaatg
aggaagaagctgttgatagtaataatattttaattaatttctgtttatattgtgatattgatacttattggaataaaattttggaaggaataaaaa
atatgttatttttaaatagcagtagcatttataataatgttaatttattagaaaaagaaaaagtttgtcatttaaaggatagtaaatgttgttatca
atccacttttcatagacatattaaatttaatttagttccatatattaattatttaatgtctaatttagatata

>Locus3429v1rpkm12.14

atgtcaagtgcacaagttatgccaattagtacagatttaaagagaaaaccagttagtcttttagacaacgtaaatacaaaacaaattaaattgg
ataatctaagttcaaaaacatcctcttataataaaactaatactacttttgaaaactatataccaaactatacccatcttgataatacccaaaag
gaaaagtttaatagtagtaaaattagtgaatatcttaccaatattatattaaataataatggtaatgttgcaagtcaaactcattttgtaactgaa
gcaccagaaccagttatagttgattattgtgaaacttgtaattatgtacttcaacattgtaaaggatgtttagatcattctaaactttgtggtaatc
ctcattgtggtaatcgtcaaaaaggtactactttctgtgttgcttgtaaaggtttagatgaatatcatccaatttgtcgtaattgtgttgaattttca
ccatatacagcccaaactaatcaatgtgctcattgtaaaggttatttctgtgctttttctctttctaatccttcccttaaatatacttgtgacaaatgt
gaaagtatggtttgttggcgttgccgtaacgtttgtaaccataataatgatactactggtagtagttcaccatcttctaaatcatccgttaatatcg
atattttaatgaag

>Locus4432v1rpkm7.57

atgaaaagttcattaaaaaccaaaatttctcaccttttttaaaggaaaatcaaataaattagaaaaatttccaacttatgatacttcgtcatcaag
agttttagacaatcagtctactccaagaagtatagaaactgaaaaattaggaagagaaaaagttcatggaaataagaaaagtcacaaaaaa
cctaagactaatttaaatttaattgatactaaggttaaaaaatatcatgaatttcctgttttaccttcaccaattcatgatcctaatgaatctgcta
aattaacagctcaagaatttgctaaagctgtaggtattaaaattttacacagaacagatgaagaagaagatgaagagtgcgattgtgaatatt
gtagatcagctcgttttaatagtactaatcttaatacagttagtactattgatccaagtttattagatgaagcttctactccaactcaacagattcc
aaatgtttctttaaatcaactttcagttaatgcttctatttcatctactaccaattcaaatgcatctactattaatgataaccaaaatattaatattc
caccattcccagcctttaatatgagttttagtaacgatagattaaataaatgttctagtaacacttctgtaagtaccaattctaatcactctattat
aaatagcaagacaattggtactccaggttatcactgtcatcgcaataacagaaagaattctatttcaaaagttattgatatgtcattatttattcc
accaacagaagaagaaatgaagagtagagttcatagttcttccttatccatcaattctactccagaagtaaataatattcaaccatgtgcttca
actgaaaaacttatgggtggtagtttagatagacacaaaaatgttggttgtaagaagaattattatggtattggtgttgcttcttcaatgtctact
cgtgaaagatcaattagtacaagtattgctagttccagtagaaataatatgagaatgatgaacaatgaatatagaacttctccaatattaaaag
aatccagtattggtcattcttctcgtattcaatttaaacaacaaccattccatcgttgtgattctggaactgaactttcaaatggaaatactagtac
cggtattaatacaaatattagtaataatggtaatgctaatattaatccttcttctccaaataaaagaccatatccaccttctcttgcttcttcttcct
cttcaactaccctttcttcagcactttctttaactcatatttcacagccaaatcttaatcaggctccatcttgttcctcaagatcttccatatctaaaa
aaccaatctctcataataattctttacaatctgttaagtattctccaaaaccaagtaaagcttcaacaccggcctgtaatatgtcccctgcatcat
caaactcatccattactccaataaaaccacataacagttttaaaattacaagatctgttactatatctgaaggtactcgtcgtgctgaaattgatc
ttcaaccaatacaacaagatgaaattaaagtatatactaaaggtcgttttactattactcatgaatattctagacgtccatctgttaattcaaatc
attctaataac

>Locus3047v1rpkm14.79

atggatctttctattcatcatagcagaactctcctttatacatcaccagttactaatgaagaagttaaattaagagaagtaaagatacttcatggt
cgtttattacatataattattagtgttttaaatatcgcttatgaaaaggaagttggcgcaagacttttatatagaaacttaagtggtaaaggaata
atggcacatcttaaagcttcatatgtaggatctgaaaagaatgttgatgaaagccaaccaaatattgatatttttgaagtaaagtatgatatgac
taagaatagatcaaacagagctccatggtccccagtggttaatttagttccatattatcgagttaagggagaatattatgaagataatccaaat
gaaagagaagagacttatagggtagctgaagattgtattgttcgtattctttgtaaaaaaggatacgtttttaaagcaccaattgtttctaaattt
gaaagtgatttagcatgggaaatttttgaaaaggaagcaggattagataaaactaatgaagctaaaaaaaatttaatgagacataattctga
aagtggtgatttcatttctaaagataataaagttttctgggaagttaattttaaggatgctcttcttcaatgtagtggtaatagtcaaaaggcttg
gaaatatttctgtcataaaacttcatgcattccaaatcctaatttagaatcaaaaaaatcatataataatcaaaactctttagatgattatgatta
ttgtgatgatgttaaagattggaaaaagagattaaatcaaaatacagcatctaaaagtagctcatggaaaattccaggatgtgaagaattttat
aaacaagatatgttccaaaatccatttgcatatgtcaaagaaccaaaagaagaagatagtgatgaagatgattctatgagaaaatctaagact
agtaaatcaaaatctaataagaaatcaattccagtttccagtaaacattcaaataagactaaaagtaacgaatatcaaaatgctattaatcca
aaaacaattagtaatcaaattaataataaatctgctaagataatgaat

>Locus10725v1rpkm0.93

atgtcattgcaaggtttattaagaaatataaattttaaaattcaagataatgaattaaataggtttaagtctattgaccataatattcgtcatgcc
tacttcactcaaattgtatatcttctttatcaacttagacagttttggaaaatgaataaattgaagcatttagatcaagaaatgtttaataaggaa
tatttgaatttatatacagatgttcctgaaataaaaatggaagaaaagaaaaaagttcatgaaatgacgaaagaagaacatgatttagataca
cgaaataaaatgtttcaacttttaacagaaaaaattaatggaatagatattgaaaatcattcatttaatacaaaagttatcgcagatagattgtt
tgataatttaaataaatctgatattaactcttttaaaatagtatggatggttatggatacaagaatgtcacatagtagtctcttagcaaaattaat
atttgataaatgtttagatttaaaaatatttaaaccagaagaaatatacctttattcaactaataaagaaataacagaaaaattaaattatcca
agtaatttaaagtttacaagtgataaggcaaatttaccaaaagatgttgatgtggtagtggtatcagtttcagcaccagtacttaaaagatgttc
attttatgttcaaaaaatatttggaagtaataatgaaaatgtagttaatcgtttacctatgattatacccattatgccatctatgccatgtttaaaa
ttacgatttgcttttgggtggcatagaacattaataccatgggtgaataaagattatatacgtaatcatcttgtttctgaaaataattcatctgttg
atatcatggaaagcagaacgtattcc

>Locus3994v1rpkm9.18

atggtttctgttcataatcctttaaacaaaaaacattcttcatatcttgtaagaaattttacggaaaaaaagaatatgaaaaagatggtatcatt
aaataaatttagaagaaaaaataagcatttggtatatgatctgtataatcgttcagaaaatgattttatagaaggatcaccaataaattcattg
aaaaggaattcttattctttttcccaatatgaggataatattataccaaagactgagaataattctcgtagaagtcttaagaaaagtaaatcttt
gattactgataccaggcttttagaagaaataccatgtgattataatagtgaaggtgaaagtacaattatgtataaatcagataaagaagaaaa
ttcctttgaatttcaatatggtacactatttggaaatgaaaaaatggctgcttgtgcaagtaattttctagaaaattgggatacatttatcagtgat
attagtggagaagatgatattaatagatctgaacattaccctagtatagaaaatcaaatcaataaagaaaaacaaatggttcctaataattca
ggttatttaacacctacaagtcaaaaggatgattcaagttcatatgaaggtaatacaattttcaatgaaaatattgaggatgcaatgatatgtg
aaaaaaatagtgaacaagctgtaaatgaaagtaattcatcagagttttcaccgattattattgatgattttgatgatatatgtccttttgaaaatg
atgattatgaagatgaaagaattagaaaaattttgaatgcagatattttacctagtgattacattggaaaaataccatctattccagaaaaacc
aatgggtgatttaggtaaacttatgaaaaaatatcatttt

>Locus2473v1rpkm20.87

atgtaccaaaatcaaagtttatatcaaaatttaggtaataattacaatagtaatagtaatacccaaaatctttatcaagatttaaatcaatattc
aaaaccagcaaatactcaaaatctttatgatacctcttctcctcaacaaaatcaacctccacaaataaataattcttatacaatagatattggaa
catccgttttccctggtcaaaatgctcaacaacctccaccacaaggtggttttaataactttaatggtggtgctcctttaagtagtccacttcctca
attccaaggtgcttcttcaatgggttctccattattgggttcccctccacttatgggttctcctccatcaatgggttctcctccaccacttggtaata
attatggttaccaagctcaatctgtaaactctgctttcactaatccttcagttataatgcctcaacctcaagcaggtttaggtttctctcaacaacc
aaacttaccacctaatatgaatactcctcgtttaggttttggagctggtggtccaccaggtggtccaggtggtctaggtggtcctggtggtcctgg
tggtcctggtagtttcagtatgatgcctcctcaaacaaatcctatgatgtctcctcaagctagttttggtggtcctgctcaacctccaccatgttac
catggaccaccaggttgg

>Locus1203v1rpkm68.48

atggtttaccaaaactctcgtaatggatttgtcaagagtttaaataactctttcatgaacaatgctgatgttaaacaacatgttcaacgtaagca
acaattaaaccaaagaagccaaaatcaaaagccaatgagaggtcaagatttagctgcttttaatgaagctgttcgtgaattatgttcaatgtac
gatgctgttaactacttcggtaattctgattatgctaagaagaatggttggagagaaattaacaaggaacaattaaatacagccactatttatg
ctattcgttatgctccagttgttccaaagagatctgcttcctcaaagagacaacaaggaagacaacaaaaccaaaaacaacaaaaccaaaga
caacaacaaccacaacaacaatacaaaagacaacaacaacaacaacaaccaagacaaaataacagaagacaaagtggtagaagaattgg
aggctgcccatacaaaccaaatgttccacgtgtttacagttataatactactccattaccaaatgatttatacccaagtgttaacaatttcaacca
cttatatgaagcttttaattacttcggtaattctcaagtaagaaatgttaagtcatggattgaaagttctcgtactcacaaattgaactctatcgct
tccgtctttgctaagagaaatgctccatatgttccagttagacaatatcgttctaataaaaaaaatcaaccacaacaacaaaaaagacaacaa
taccaaaagaaacaacaacaacaagaaattgttccaaaacaacaatcaaacaaaagacaaaacaaaagacaacaaaagaaacaaatgat
taaccaagaaccactttcaaatgtctgggaagattactacgtttacccaaatcaacgtatgttcaag

>Locus6738v1rpkm3.07

atgagtcaaattaataatgaatctataccttcgtcaaataatatttgtgacatagaaaaagaaaaagaaattattaaaaataattataataata
acaataataatagtaatcaatcttcaccaaattcttctttaaacactttagcatcaaattcttctcaatatttaattacatctaataaaagaaaaa
attccctaacttcttcaatattatcaaatacaaataccttaaataattctcgttctactattggattaaaccatcaccaaaaaacatattcatcctc
cagtcaacaacaacaacaacaacaacaacaacaacaacaaaatctacatcattctaaaatgagttcaccttcagaaattagtgaaagttatat
tgcaaaacaagcttttgcaaatatttcaggaagtactccaagttcagtacaatcatcagaaaattctttaagttcttcaactcctactattaaaag
aagtttagatagtgtatcaagttcaaaaaataatattcattctaaaaataatactagtacaaaaagaagttcaacttcttctaaaaataattata
ctaaagcttttgatttaaatgataaacaattaagtattacacttagttgttcaagtgcaagtgcttctagtagtaataattcaagtaataaatcaa
gtcgtaataatattgctaaaggaataatagaagaggacgaaaggggagattgtaatattaataataaaacagaaacaggagaagaggataa
taaaggaactaaaaaatttggatttaaacattttaaaaatactatatcaaaaacatttagtaaagcttctttccgtaaattttcagtttctagtaa
aaaaacaattaatagtaattcggaatcacttggtgaacaagaatttaatttcaattcaaaagaagaacaagaagaatttaataatttcgttaat
acattaggttcgaaaaaagcaaatagtagacgtagacccacattaaatggaatatttcaaaatccggatatgccatcacaaccatcatcaagt
aataataatgaggataatttagaaaatattcaatcacgtcgaaaaagtattgataataaaaattttgaaacagctttaaatcatttatgtgaagt
attggattatgaagaaccaatggtattggcaaaatatttaaaagcagcaaatggtgatgaaaatttagctttgaaaaattatttaaaagatgct
aataaaaatactaaaaaagcaaataatcattcagataaagatgctccaatgaatataaatgaaatgaataatattgaaggaaatacgacaa
gacaacctcctgtaataatagataagaagaatagacgagtaaaa

7.4.5. Co-regulated sequences from N. californiae *glucose pulse experiment*

>Locus12489v1rpkm1.81

atgtcaagtgcatctattcaagttatgacaattaataacgaactgaagagaaaaccggtgagtcttctcgataatatgaatgctaaacaaatta
aacttgatttgtccggtttaaaaacaaaaatgagtagtactaatgatacttttgaaaactatataccaacctatacccaatatgataataatacc
caaaaagaaaagtttaccaatagtaaaattagtgaattatatacaaatcgtattttaaatactcataatgaaacacattttgttactgaagcacc
tgccccagttcctgtagattactgtgaaacttgtagttgtatgcttcaacattgtaaaggttgtttagatcatactaaactttgtggtaatcctcatt
gtggtaatcgtcaaaagggtactacgttttgtgttgcctgtaaaggtttagatgaatatcatcctatttgtcgtagttgtgtagagtatcctccttac
tctgctcaaacaaatcaatgtggtcattgtaagggttatttctgtgctgcctccctctctaatccttccttaaaatatacttgtagaaaatgtgata
ctatggtttgttggcgttgtcgtaccatgtgtcatcatgaggataatactaatgaagatgaagctattgctgctgctaaagcttctgttaatattga
tatattactgaag

>Locus5354v2rpkm3.75

atgtctgttgaaattgaccaaagtgaagaagaagcttttaactcttccaagagtaaaagagccgataagcttcgtgaatattatttcccaacaa
aagagccatctcaacaagaaacagagaaaattccgaaaaagaagagaccttctattgtttcctctgttgatccaaatcaacttccacaagtga
attattctactttaggtattgatgttgaacaattagattcaactatttcctttaatgtggagtcctttgttaatgatacattaggtgaatttggttgga
cgccatctcatgaacttcaattaagtcaaacacttaatttggatgattttaagctagaaacgactgcctatttagataatcagtgccttgaactta
ttcaaaaggcctacaaaagtacttctttaacctgtcgtggtggtaaagtttatgacacaaaggaactcacagaccttgttacattagaaaaaat
acaaaatgatttcgattctttaagagttattactgagacagttgaaacaataattccaaaaagtcaaacattagataatttgaagggttacctta
aagatgtaaaggaagacgaaaagattaatgaatttagtgaagtatgtgatattgtaagagatttatcgcaagtacttaaacaagttgaggata
attatattgataataattcagacaatgaagcatttgcaattatagcattctataaattaagacaaatggaaattaatggaataacaaaacaaca
tttagatgttttaaatgaagcaattacttcatatgaaaataaagcaaatccaaaagattttgaaaaattaaagaaaatccaaatattcgctaatt
ttattggtgaacatctttcatca

>Locus8268v1rpkm4.54

atggttgttaattctaaaaagttaattataaaaatttttttaaaaaaggaattcattaaaaaagaattacctaattattatgatattttatgtaaat
accaaaataaaaactattattatagtattccaaatactaaaaagttaataaagtcaaatataaaaaatcgtaaaattagaaaaatgttacttga
aaatatgctagctttagatgctaatttatataaaaaaattaaatatatgaaaggatataaattatcatcaaataacattaataaaaataggaga
aataagttaatgataaaaaataaattaaatatgatttcgtcaactaaaaattcaaaagattgttgtgtagaaaaaactttacctaaattaacctt
taaaattcaaccaatgatgaaaaatagcgaaaacattattaaaactgaacaacctcatacccgtaacttaaatagtaataatagtgctttttta
actggaaatgttgaagaagaaaaagcttatcattatcgttcaaaatccgttacttttgtcaaggaaagcttggaacaaattcgtttattccatatc
tttgatcctccaaatagtattacaatgaccccaattcattattctactggaaaaatagattctaattctcttttcagattacataaccgtaatagaa
atgataaatttaaagttatatcatttacttcctctgaattaaagaaaaacgcattatatgatcaccctgttcgtagtagtcatcgttcaaaattaa
atataccttataatttaaagatggacttttcttttggaagtaattcaccatccccaaaatcctcatccccttcttccccaaagtctacacctaaatt
agaaatgtcaaataatcttccaatagaaaagaataatacctctaattataaattaagtgaagatcgttcaagtcaaaaggaaatgaatttaag
ttatgaattaggaaacccaaaatctcctgtagtaccaaataactatagttatacaaccaatgtcatggataattctcttaattttgtaaataataa
taaagaagttatctataatacttctgcacttacagtaaccccacataaaataacacctttagttcaattaata

>Locus5687v1rpkm9.00

atgtcaagtacaacaattcaaatcatggcaatcaataatgaattaaagagaaaaccagttggtttactgggtaatatgaatactaaacaagtt
agacttgatttagctggtttaaaagctagaaataatactaatactacttttgaaaactatataccatcctatacccaatatgaaaatacccaaga
aaagttagccaatagtaaaattagtgatttatttaatcaacaaatattagattcaactaacggaactcattttgttactgaggctccagaacctgt
tgaagtggattattgtgaaacttgtaatcatgtacttcaacattgtaagggttgtttagatcatactaaactttgtggtaatcctcattgtggtaat
cgtcaaaagggaacaactttttgtattgcttgtaaaggtttagatgaattccatcctatttgtcgtaattgtgttgaatacccaccatactctgctc
caacgaatcaatgtggtcattgtaagggttatttctgttcagcagctctttcaaatccttcattaaaatatacttgtagtaaatgtgattcaatggt
atgttggcgttgtcgtaccatgtgtaatcataacactttaaatactaataatgagaatgatactataacttcgaaggcctccgttaatattgatat
tttaatgaag

>Locus3346v1rpkm21.30

atgggatattttgaaaaacaattatatagaagaataaacagttcacccaaagaggttgtcctctttgatcagtcttatgtttatcaaaaatcgaa
gagcctctacattgaatatgttgaaggattaaagtttccttattttttcattgtcaagggaatccaagataaagaattctttaaatgtgtttataaa
tgtagtaaactatttttttatacgatggagggtgttcctattttcaactatgacaataattcctctccaaagaaaatatatgctggtgataaagag
gataaattaattgctaccctcaccagaaagtattcatggaaagctaaaaaatataaggtagagtatgttaatttactcactcaaaaaaaggaa
attctcgatatgaatcttgataagggctatcgtacctgtggtgtttttcttggtcgagaaagtgataccaccccaaaaatctgtagaatggttgcc
tttaaaaaggaaaagaattatggaccacgttatctagtagaaatatcgagtggagtagataatatgtttatgattgcccttggtatatcatttgct
attcttagaacgaaggcagaaatttatgaaagagaaactttcattaatgaatgtataaaatctcataattattccatgatagctaatattagtga
a

>Locus2561v1rpkm32.14

atggataaaataccaccatatctttatacccaagcattagcaaaacaattaaatatagaaaatggttctaaaattgaatataaagcaaatgact
ttattattgctttaaataattttaaaatggatccagaaaaatttgctaactttgaaagaaaaagatatagtcgtcttattgtccgtgaaattattca
cgctatgggtttcacatctactgaagtaattgcccaattaaaagatacggaagatatgttaaagatgtcaccatccttcattaagaatgatggtt
ccaatacatttagatatataccaaatgttttatctgatattgactggaatcaattgagtaaagtttcaaaattagaagattatgtaagtgaattat
atgattcgaaatttataggtttgtcaccattgacagtgtttgctaaaaatattgttgatattaaaactaaagagaagttatttaaagatttgggat
tttattataaagattttaattgtataaaagatcaagaagatgttgaagctattaaggatgttttagagaaacatcgtttagaatgttttaaacaat
tggatgaaaaaacaaaggaaacggtcactaacattgctatgaaatatttccttaaaagtaaaagtattggatttttaacggatgctggtaaggt
tattccacttcaaacatttgaagatatgttccatccaggtagtagtattaatcatattcaattcgataaatatgatgaaattcgtgatgatccaga
aaagaagacagaatttcttaagggtacctttatcacgaaggaaaatatttcagattattataatgaagaagctttaatgtactatactcaaggg
gattctattagtaatgaagaatttttagaaaccattgctaaatctaatgctcatggtttaataggtcctagtatggttgaagcattaaaaacttta
ggttggacagaaaagggtaaagaaagtgaatccaaacgtatttattatttcgatgaaaaggaagttccttatccagaacaaaatacctttaaa
tatgttaatatgaaaatttatgaaatcactatgtcacaacaaacttcagaagaatctgaaaattcatcagaaaagaaagatgaaacaaaaga
aaatgaaactaaacaaaaaatagtaaaagaagaatta

8. References

1 de Jong, E., Higson, A., Walsh, P. & Wellisch, M. Bio-based chemicals value added products from biorefineries. *IEA Bioenergy, Task42 Biorefinery* (2012).

2 Naik, S. N., Goud, V. V., Rout, P. K. & Dalai, A. K. Production of first and second generation biofuels: A comprehensive review. *Renewable and Sustainable Energy Reviews* **14**, 578-597, doi:10.1016/j.rser.2009.10.003 (2010).

3 Jørgensen, H., Kristensen, J. B. & Felby, C. Enzymatic conversion of lignocellulose into fermentable sugars: challenges and opportunities. *Biofuels, Bioproducts and Biorefining* **1**, 119-134, doi:10.1002/bbb.4 (2007).

4 Zeng, Y., Zhao, S., Yang, S. & Ding, S. Y. Lignin plays a negative role in the biochemical process for producing lignocellulosic biofuels. *Curr Opin Biotechnol* **27**, 38-45, doi:10.1016/j.copbio.2013.09.008 (2014).

5 Olson, D. G., McBride, J. E., Shaw, A. J. & Lynd, L. R. Recent progress in consolidated bioprocessing. *Curr Opin Biotechnol* **23**, 396-405, doi:10.1016/j.copbio.2011.11.026 (2012).

6 Flint, H. J. The rumen microbial ecosystem-some recent developments. *Trends in Microbiology* **5**, 483-488, doi:10.1016/S0966-842X(97)01159-1 (1997).

7 Windham, W. R. A., D E. Rumen fungi and forage fiber degradation. *Applied and Environmental Microbiology* **48**, 473-476 (1984).

8 Nicholson, M. J., Theodorou, M. K. & Brookman, J. L. Molecular analysis of the anaerobic rumen fungus Orpinomyces - insights into an AT-rich genome. *Microbiology* **151**, 121-133, doi:10.1099/mic.0.27353-0 (2005).

9 Bootten, T. J., Joblin, K. N., McArdle, B. H. & Harris, P. J. Degradation of lignified secondary cell walls of lucerne (Medicago sativa L.) by rumen fungi growing in methanogenic co-culture. *J Appl Microbiol* **111**, 1086-1096, doi:10.1111/j.1365-2672.2011.05127.x (2011).

10 Ljungdahl, L. G. The cellulase/hemicellulase system of the anaerobic fungus Orpinomyces PC-2 and aspects of its applied use. *Ann N Y Acad Sci* **1125**, 308-321, doi:10.1196/annals.1419.030 (2008).

11 Haitjema, C. H., Solomon, K. V., Henske, J. K., Theodorou, M. K. & O'Malley, M. A. Anaerobic gut fungi: Advances in isolation, culture, and cellulolytic enzyme discovery for biofuel production. *Biotechnol Bioeng* **111**, 1471-1482, doi:10.1002/bit.25264 (2014).

12 Sanderson, K. US biofuels a field in ferment. *Nature* **444**, 673-676, doi:10.1038/444673a (2006).

13 Lee, R. A. & Lavoie, J. M. From first- to third-generation biofuels: Challenges of producing a commodity from a biomass of increasing complexity. *Animal Frontiers* **3**, 6-11, doi:10.2527/af.2013-0010 (2013).

14 Naylor, R. L. *et al.* The Ripple Effect: Biofuels, Food Security, and the Environment. *Environment* **49**, 30-43, doi:10.3200/ENVT.49.9.30-43 (2007).

15 Himmel, M. E. *et al.* Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* **315**, 804-807, doi:10.1126/science.1137016 (2007).

16 Brethauer, S. & Studer, M. H. Biochemical Conversion Processes of Lignocellulosic Biomass to Fuels and Chemicals - A Review. *Chimia (Aarau)* **69**, 572-581, doi:10.2533/chimia.2015.572 (2015).

17 Mosier, N. *et al.* Features of promising technologies for pretreatment of lignocellulosic biomass. *Bioresour Technol* **96**, 673-686, doi:10.1016/j.biortech.2004.06.025 (2005).

18 Bayer, E. a., Chanzy, H., Lamed, R. & Shoham, Y. Cellulose, cellulases and cellulosomes. *Current opinion in structural biology* **8**, 548-557 (1998).

19 Saha, B. C. Hemicellulose bioconversion. *J Ind Microbiol Biotechnol* **30**, 279-291, doi:10.1007/s10295-003-0049-x (2003).

20 Shallom, D. & Shoham, Y. Microbial hemicellulases. *Current Opinion in Microbiology* **6**, 219-228, doi:10.1016/s1369-5274(03)00056-0 (2003).

21 Boerjan, W., Ralph, J. & Baucher, M. Lignin biosynthesis. *Annu Rev Plant Biol* **54**, 519-546, doi:10.1146/annurev.arplant.54.031902.134938 (2003).

22 Sun, Y. & Cheng, J. Hydrolysis of lignocellulosic materials for ethanol production: a review. *Bioresource Technology* **83**, 1-11, doi:10.1016/S0960-8524(01)00212-7 (2002).

23 Agbor, V. B., Cicek, N., Sparling, R., Berlin, A. & Levin, D. B. Biomass pretreatment: fundamentals toward application. *Biotechnology advances* **29**, 675-685, doi:10.1016/j.biotechadv.2011.05.005 (2011).

24 Alvira, P., Tomas-Pejo, E., Ballesteros, M. & Negro, M. J. Pretreatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: A review. *Bioresour Technol* **101**, 4851-4861, doi:10.1016/j.biortech.2009.11.093 (2010).

25 Balan, V., Chiaramonti, D. & Kumar, S. Review of US and EU initiatives toward development, demonstration, and commercialization of lignocellulosic biofuels. *Biofuels, Bioproducts and Biorefining* **7**, 732-759, doi:10.1002/bbb.1436 (2013).

26 Zhang, L., Xu, C. & Champagne, P. Overview of recent advances in thermo-chemical conversion of biomass. *Energy Conversion and Management* **51**, 969-982, doi:10.1016/j.enconman.2009.11.038 (2010).

27 Kumar, A., Jones, D. D. & Hanna, M. A. Thermochemical Biomass Gasification: A Review of the Current Status of the Technology. *Energies* **2**, 556-581, doi:10.3390/en20300556 (2009).

28 Dimarogona, M., Topakas, E. & Christakopoulos, P. Cellulose degradation by oxidative enzymes. *Comput Struct Biotechnol J* **2**, e201209015, doi:10.5936/csbj.201209015 (2012).

29 Vasudevan, P., Sharma, S. & Kumar, A. Liquid fuel from biomass: An overview. *Journal of Scientific and Industrial Research* **64**, 822-831 (2005).

30 Ma, F. & Hanna, M. A. Biodiesel production: A review. *Bioresource Technology* **70**, 1-15, doi:10.1016/S0960-8524(99)00025-5 (1999).

31 Steen, E. J. *et al.* Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* **463**, 559-562, doi:10.1038/nature08721 (2010).

32 Sitepu, I. R. *et al.* Oleaginous yeasts for biodiesel: current and future trends in biology and production. *Biotechnol Adv* **32**, 1336-1360, doi:10.1016/j.biotechadv.2014.08.003 (2014).

33 Tseng, H. C. & Prather, K. L. Controlled biosynthesis of odd-chain fuels and chemicals via engineered modular metabolic pathways. *Proc Natl Acad Sci U S A* **109**, 17925-17930, doi:10.1073/pnas.1209002109 (2012).

34 Kawaguchi, H., Hasunuma, T., Ogino, C. & Kondo, A. Bioprocessing of bio-based chemicals produced from lignocellulosic feedstocks. *Curr Opin Biotechnol* **42**, 30-39, doi:10.1016/j.copbio.2016.02.031 (2016).

35 Ahring, B. K., Traverso, J. J., Murali, N. & Srinivas, K. Continuous fermentation of clarified corn stover hydrolysate for the production of lactic acid at high yield and productivity. *Biochemical Engineering Journal* **109**, 162-169, doi:10.1016/j.bej.2016.01.012 (2016).

36 Abdel-Rahman, M. A., Tashiro, Y. & Sonomoto, K. Lactic acid production from lignocellulose-derived sugars using lactic acid bacteria: overview and limits. *J Biotechnol* **156**, 286-301, doi:10.1016/j.jbiotec.2011.06.017 (2011).

37 Abdel-Rahman, M. A., Tashiro, Y. & Sonomoto, K. Recent advances in lactic acid production by microbial fermentation processes. *Biotechnol Adv* **31**, 877-902, doi:10.1016/j.biotechadv.2013.04.002 (2013).

38 Liang, L. *et al.* Repetitive succinic acid production from lignocellulose hydrolysates by enhancement of ATP supply in metabolically engineered Escherichia coli. *Bioresour Technol* **143**, 405-412, doi:10.1016/j.biortech.2013.06.031 (2013).

39 Okino, S. *et al.* An efficient succinic acid production process in a metabolically engineered Corynebacterium glutamicum strain. *Appl Microbiol Biotechnol* **81**, 459-464, doi:10.1007/s00253-008-1668-y (2008).

40 Salvachua, D. *et al.* Succinic acid production from lignocellulosic hydrolysate by Basfia succiniciproducens. *Bioresour Technol* **214**, 558-566, doi:10.1016/j.biortech.2016.05.018 (2016).

41 Wang, J. *et al.* Enhanced succinic acid production and magnesium utilization by overexpression of magnesium transporter mgtA in Escherichia coli mutant. *Bioresour Technol* **170**, doi:10.1016/j.biortech.2014.07.081 (2014).

42 Lynd, L. R., van Zyl, W. H., McBride, J. E. & Laser, M. Consolidated bioprocessing of cellulosic biomass: an update. *Curr Opin Biotechnol* **16**, 577-583, doi:10.1016/j.copbio.2005.08.009 (2005).

43 van Zyl, W. H., Lynd, L. R., den Haan, R. & McBride, J. E. Consolidated bioprocessing for bioethanol production using Saccharomyces cerevisiae. *Adv Biochem Eng Biotechnol* **108**, 205-235, doi:10.1007/10_2007_061 (2007).

44 Tsai, S. L., DaSilva, N. A. & Chen, W. Functional display of complex cellulosomes on the yeast surface via adaptive assembly. *ACS Synth Biol* **2**, 14-21, doi:10.1021/sb300047u (2013).

45 la Grange, D. C., den Haan, R. & van Zyl, W. H. Engineering cellulolytic ability into bioprocessing organisms. *Appl Microbiol Biotechnol* **87**, 1195-1208, doi:10.1007/s00253-010-2660-x (2010).

46 Mazzoli, R., Lamberti, C. & Pessione, E. Engineering new metabolic capabilities in bacteria: lessons from recombinant cellulolytic strategies. *Trends Biotechnol* **30**, 111-119, doi:10.1016/j.tibtech.2011.08.003 (2012).

47 den Haan, R., van Rensburg, E., Rose, S. H., Gorgens, J. F. & van Zyl, W. H. Progress and challenges in the engineering of non-cellulolytic microorganisms for consolidated bioprocessing. *Curr Opin Biotechnol* **33**, 32-38, doi:10.1016/j.copbio.2014.10.003 (2015).

48 Lambertz, C. *et al.* Challenges and advances in the heterologous expression of cellulolytic enzymes: a review. *Biotechnology for Biofuels* **7**, 1-15, doi:10.1186/s13068-014-0135-5 (2014).

49 van Rensburg, E., den Haan, R., Smith, J., van Zyl, W. H. & Gorgens, J. F. The metabolic burden of cellulase expression by recombinant Saccharomyces cerevisiae Y294 in aerobic batch culture. *Appl Microbiol Biotechnol* **96**, 197-209, doi:10.1007/s00253-012-4037-9 (2012).

50 Desvaux, M. Clostridium cellulolyticum: model organism of mesophilic cellulolytic clostridia. *FEMS Microbiol Rev* **29**, 741-764, doi:10.1016/j.femsre.2004.11.003 (2005).

51 Higashide, W., Li, Y., Yang, Y. & Liao, J. C. Metabolic engineering of Clostridium cellulolyticum for production of isobutanol from cellulose. *Appl Environ Microbiol* **77**, 2727-2733, doi:10.1128/AEM.02454-10 (2011).

52 Guedon, E., Desvaux, M. & Petitdemange, H. Improvement of Cellulolytic Properties of Clostridium cellulolyticum by Metabolic Engineering. *Applied and Environmental Microbiology* **68**, 53-58, doi:10.1128/aem.68.1.53-58.2002 (2002).

53 Lutke-Eversloh, T. & Bahl, H. Metabolic engineering of Clostridium acetobutylicum: recent advances to improve butanol production. *Curr Opin Biotechnol* **22**, 634-647, doi:10.1016/j.copbio.2011.01.011 (2011).

54 Chung, D. *et al.* Cellulosic ethanol production via consolidated bioprocessing at 75 degrees C by engineered Caldicellulosiruptor bescii. *Biotechnol Biofuels* **8**, 163, doi:10.1186/s13068-015-0346-4 (2015).

55 Cha, M., Chung, D., Elkins, J. G., Guss, A. M. & Westpheling, J. Metabolic engineering of Caldicellulosiruptor bescii yields increased hydrogen production from lignocellulosic biomass. *Biotechnology for Biofuels* **6**, 85-92, doi:10.1186/1754-6834-6-85 (2013).

56 Kuck, U. & Hoff, B. New tools for the genetic manipulation of filamentous fungi. *Appl Microbiol Biotechnol* **86**, 51-62, doi:10.1007/s00253-009-2416-7 (2010).

57 Sambrook, J. & Russell, D. W. *Molecular cloning: a laboratory manual*. (CSHL press, 2001).

58 Peng, X. N., Gilmore, S. P. & O’Malley, M. A. Microbial communities for bioprocessing: lessons learned from nature. *Current Opinion in Chemical Engineering* **14**, 103-109, doi:10.1016/j.coche.2016.09.003 (2016).

59 Zuroff, T. R., Xiques, S. B. & Curtis, W. R. Consortia-mediated bioprocessing of cellulose to ethanol with a symbiotic Clostridium phytofermentans/yeast co-culture. *Biotechnology for Biofuels* **6**, 59-70, doi:10.1186/1754-6834-6-59 (2013).

60 Minty, J. J. *et al.* 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* **110**, 14592-14597, doi:10.1073/pnas.1218447110 (2013).

61 Wilkinson, S., Smart, K. A., James, S. & Cook, D. J. Bioethanol Production from Brewers Spent Grains Using a Fungal Consolidated Bioprocessing (CBP) Approach. *BioEnergy Research* **10**, 146-157, doi:10.1007/s12155-016-9782-7 (2016).

62 Trinci, A. P. J. *et al.* Anaerobic fungi in herbivorous animals. *Mycological Research* **98**, 129-152, doi:10.1016/S0953-7562(09)80178-0 (1994).

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

64 McCann, J. C., Wickersham, T. A. & Loor, J. J. High-throughput Methods Redefine the Rumen Microbiome and Its Relationship with Nutrition and Metabolism. *Bioinform Biol Insights* **8**, 109-125, doi:10.4137/BBI.S15389 (2014).

65 Brulc, J. M. *et al.* Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc Natl Acad Sci U S A* **106**, 1948-1953, doi:10.1073/pnas.0806191105 (2009).

66 Theodorou, M. K. *et al.* Anaerobic fungi in the digestive tract of mammalian herbivores and their potential for exploitation. *Proc Nutr Soc* **55**, doi:10.1079/pns19960088 (1996).

67 Bauchop, T. M., Douglas O. Cellulose Ferementation by a Rumen Anaerobic Fungus in Both the Absence and Presence of Rumen Methanogens. *Applied and Environmental Microbiology* **42**, 1103-1110 (1981).

68 Marvin-sikkema, F. D., Richardson, A. J., Stewart, C. S., Gottschal, J. A. N. C. & Prins, R. A. Influence of Hydrogen-Consuming Bacteria on Cellulose Degradation by Anaerobic Fungi. *Applied and Environmental Microbiology* **56**, 3793-3797 (1990).

69 Gordon, G. L. & Phillips, M. W. The role of anaerobic gut fungi in ruminants. *Nutrition Research Reviews* **11**, 133-168, doi:10.1079/NRR19980009 (1998).

70 Joblin, K. N. Isolation, Enumeration, and Maintenance of Rumen Anaerobic Fungi in Roll Tubes. *Applied and Environmental Microbiology* **42**, 1119-1122 (1981).

71 Orpin, C. G. Studies on the rumen flagellate Neocallimastix frontalis. *J Gen Microbiol* **91**, doi:10.1099/00221287-91-2-249 (1975).

72 Mountfort, D. O. & Orpin, C. G. *Anaerobic Fungi: Biology, Ecology, and Function*. (CRC Press, 1994).

73 Ozkose, E., Thomas, B. J., Davies, D. R., Griffith, G. W. & Theodorou, M. K. *Cyllamyces aberensis* gen.nov. sp.nov., a new anaerobic gut fungus with branched sporangiophores isolated from cattle. *Canadian Journal of Botany* **79**, 666-673, doi:10.1139/cjb-79-6-666 (2001).

74 Dagar, S. S. *et al.* A new anaerobic fungus (Oontomyces anksri gen. nov., sp. nov.) from the digestive tract of the Indian camel (Camelus dromedarius). *Fungal Biol* **119**, 731-737, doi:10.1016/j.funbio.2015.04.005 (2015).

75 Griffith, G. W. *et al.* Buwchfawromyces eastonii gen. nov., sp. nov.: a new anaerobic fungus (Neocallimastigomycota) isolated from buffalo faeces. *MycoKeys* **9**, 11-28, doi:10.3897/mycokeys.9.9032 (2015).

76 Ho, Y. W. & Barr, D. J. S. Classification of Anaerobic Gut Fungi from Herbivores with Emphasis on Rumen Fungi from Malaysia. *Mycological Society of America* **87**, 655-677, doi:10.2307/3760810 (1995).

77 Brookman, J. L., Mennim, G., Trinci, a. P., Theodorou, M. K. & Tuckwell, D. S. Identification and characterization of anaerobic gut fungi using molecular methodologies based on ribosomal ITS1 and 18S rRNA. *Microbiology* **146**, 393-403 (2000).

78 Tuckwell, D. S., Nicholson, M. J., McSweeney, C. S., Theodorou, M. K. & Brookman, J. L. The rapid assignment of ruminal fungi to presumptive genera using ITS1 and ITS2 RNA secondary structures to produce group-specific fingerprints. *Microbiology* **151**, 1557-1567, doi:10.1099/mic.0.27689-0 (2005).

79 Youssef, N. H. *et al.* The genome of the anaerobic fungus Orpinomyces sp. strain C1A reveals the unique evolutionary history of a remarkable plant biomass degrader. *Appl Environ Microbiol* **79**, 4620-4634, doi:10.1128/AEM.00821-13 (2013).

80 Grigoriev, I. V. *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res* **42**, D699-704, doi:10.1093/nar/gkt1183 (2014).

81 Nordberg, H. *et al.* The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res* **42**, D26-31, doi:10.1093/nar/gkt1069 (2014).

82 Haitjema, C. H. *et al.* A parts list for fungal cellulosomes revealed by comparative genomics. *Nature Microbiology* **2**, doi:10.1038/nmicrobiol.2017.87 (2017).

83 Solomon, K. V., Henske, J. K., Theodorou, M. K. & O'Malley, M. A. Robust and effective methodologies for cryopreservation and DNA extraction from anaerobic gut fungi. *Anaerobe* **38**, 39-46, doi:10.1016/j.anaerobe.2015.11.008 (2016).

84 Brownlee, A. G. Remarkably AT-rich genomic DNA from the anaerobic fungus Neocallimastix. *Nucleic Acids Research* **17**, 1327-1335 (1989).

85 Chen, H., Hopper, S. L., Li, X. L., Ljungdahl, L. G. & Cerniglia, C. E. Isolation of extremely AT-rich genomic DNA and analysis of genes encoding carbohydrate-degrading enzymes from Orpinomyces sp. strain PC-2. *Curr Microbiol* **53**, 396-400, doi:10.1007/s00284-006-0098-2 (2006).

86 Gardner, M. J. A status report on the sequencing and annotation of the P. falciparum genome. *2001* **118**, 133-138, doi:10.1016/S0166-6851(01)00390-5 (2001).

87 Roberts, R. J., Carneiro, M. O. & Schatz, M. C. The advantages of SMRT sequencing. *Genome Biology* **14**, 405-408, doi:10.1186/gb-2013-14-6-405 (2013).

88 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA* **74**, 5463-5467 (1977).

89 Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145, doi:10.1038/nbt1486 (2008).

90 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

91 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).

92 Anderson, S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research* **9**, 3015-3027, doi:10.1093/nar/9.13.3015 (1981).

93 Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333-351, doi:10.1038/nrg.2016.49 (2016).

94 von Bubnoff, A. Next-generation sequencing: the race is on. *Cell* **132**, 721-723, doi:10.1016/j.cell.2008.02.028 (2008).

95 Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, doi:10.1038/nature07517 (2008).

96 Schirmer, M. *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res* **43**, e37, doi:10.1093/nar/gku1341 (2015).

97 Claesson, M. J. *et al.* Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res* **38**, e200, doi:10.1093/nar/gkq873 (2010).

98 Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biology* **14**, R51, doi:10.1186/gb-2013-14-5-r51 (2013).

99 Rank, D. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133-138, doi:10.1126/science.1162986 (2009).

100 Feingold, E. *et al.* The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640, doi:10.1126/science.1105136 (2004).

101 Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **44**, D7-19, doi:10.1093/nar/gkv1290 (2016).

102 Finn, R. D. *et al.* InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res* **45**, D190-D199, doi:10.1093/nar/gkw1107 (2017).

103 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, doi:10.1016/s0022-2836(05)80360-2 (1990).

104 Jirtle, R. L. & Skinner, M. K. Environmental epigenomics and disease susceptibility. *Nat Rev Genet* **8**, 253-262, doi:10.1038/nrg2045 (2007).

105 Razin, A. & Riggs, A. D. DNA Methylation and Gene Function. *Science* **210**, 604-610 (1980).

106 Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41-45, doi:10.1038/47412 (2000).

107 Tsompana, M. & Buck, M. J. Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin* **7**, 16, doi:10.1186/1756-8935-7-33 (2014).

108 Goldberg, A. D., Allis, C. D. & Bernstein, E. Epigenetics: a landscape takes shape. *Cell* **128**, 635-638, doi:10.1016/j.cell.2007.02.006 (2007).

109 Handelsman, J. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews* **68**, 669-685, doi:10.1128/MMBR.68.4.669-685.2004 (2004).

110 Degnan, P. H. & Ochman, H. Illumina-based analysis of microbial community diversity. *ISME J* **6**, 183-194, doi:10.1038/ismej.2011.74 (2012).

111 Fierer, N. *et al.* Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl Environ Microbiol* **73**, 7059-7066, doi:10.1128/AEM.00358-07 (2007).

112 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57-63, doi:10.1038/nrg2484 (2009).

113 Morozova, O., Hirst, M. & Marra, M. A. Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet* **10**, 135-151, doi:10.1146/annurev-genom-082908-145957 (2009).

114 Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research* **27**, 4636-4641, doi:10.1093/nar/27.23.4636 (1999).

115 Lukashin, A. V. & Borodovsky, M. GeneMark. hmm: new solutions for gene finding. *Nucleic Acids Research* **26**, 1107-1115, doi:10.1093/nar/26.4.1107 (1998).

116 Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59, doi:10.1186/1471-2105-5-59 (2004).

117 Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78-94, doi:10.1006/jmbi.1997.0951 (1997).

118 Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nature Biotechnology* **20**, 508-512, doi:10.1038/nbt0502-508 (2002).

119 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, doi:10.1093/bioinformatics/btp120 (2009).

120 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

121 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).

122 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652, doi:10.1038/nbt.1883 (2011).

123 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323 (2011).

124 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

125 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).

126 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 1-21, doi:10.1186/s13059-014-0550-8 (2014).

127 Leis, S. *et al.* Finding a robust strain for biomethanation: anaerobic fungi (Neocallimastigomycota) from the Alpine ibex (Capra ibex) and their associated methanogens. *Anaerobe* **29**, 34-43, doi:10.1016/j.anaerobe.2013.12.002 (2014).

128 Nicholson, M. J., McSweeney, C. S., Mackie, R. I., Brookman, J. L. & Theodorou, M. K. Diversity of anaerobic gut fungal populations analysed using ribosomal ITS1 sequences in faeces of wild and domesticated herbivores. *Anaerobe* **16**, 66-73, doi:10.1016/j.anaerobe.2009.05.003 (2010).

129 Henry, T., Iwen, P. C. & Hinrichs, S. H. Identification of Aspergillus Species Using Internal Transcribed Spacer Regions 1 and 2. *Journal of Clinical Microbiology* (2000).

130 Solomon, K. V. *et al.* Early-branching gut fungi possess a large, comprehensive array of biomass-degrading enzymes. *Science* **351**, doi:10.1126/science.aad1431 (2016).

131 Li, G. J. *et al.* Fungal diversity notes 253–366: taxonomic and phylogenetic contributions to fungal taxa. *Fungal Divers.* **78**, doi:10.1007/s13225-016-0366-9 (2016).

132 Harhangi, H. R. *et al.* Xylose metabolism in the anaerobic fungus Piromyces sp. strain E2 follows the bacterial pathway. *Arch Microbiol* **180**, 134-141, doi:10.1007/s00203-003-0565-0 (2003).

133 Martin, J. *et al.* Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC genomics* **11**, 663-670, doi:10.1186/1471-2164-11-663 (2010).

134 Kamenetskii, F. Simplification of the empirical relationship between melting temperature of DNA, its GC content and concentration of sodium ions in solution. *Biopolymers* **10**, 2623-2624, doi:10.1002/bip.360101223 (1971).

135 Sharp, P. M., Tuohy, T. M. F. & Mosurski, K. R. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research* **14**, 5125-5143, doi:10.1093/nar/14.13.5125 (1986).

136 Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42**, D490-495, doi:10.1093/nar/gkt1178 (2014).

137 Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **37**, D233-238, doi:10.1093/nar/gkn663 (2009).

138 Beg, Q. K., Kapoor, M., Mahajan, L. & Hoondal, G. S. Microbial xylanases and their industrial applications: a review. *Applied Microbiology and Biotechnology* **56**, 326-338, doi:10.1007/s002530100704 (2001).

139 Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669-680, doi:10.1038/nrg2641 (2009).

140 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208, doi:10.1093/nar/gkp335 (2009).

141 Bailey, T. L. & Elkan, C. Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 28-36, doi:citeulike-article-id:878292 (1994).

142 Pelechano, V. & Steinmetz, L. M. Gene regulation by antisense transcription. *Nat Rev Genet* **14**, 880-893, doi:10.1038/nrg3594 (2013).

143 Wagner, E. G. H. & Simons, R. W. Antisense RNA control in bacteria, phage, and plasmids. *Annual Reviews Microbiology* **48**, 713-742 (1994).

144 Britto-Kido Sde, A. *et al.* Natural antisense transcripts in plants: a review and identification in soybean infected with Phakopsora pachyrhizi SuperSAGE library. *ScientificWorldJournal* **2013**, 219798, doi:10.1155/2013/219798 (2013).

145 Donaldson, M. E. & Saville, B. J. Natural antisense transcripts in fungi. *Mol Microbiol* **85**, 405-417, doi:10.1111/j.1365-2958.2012.08125.x (2012).

146 Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564-1566, doi:10.1126/science.1112009 (2005).

147 Arthanari, Y., Heintzen, C., Griffiths-Jones, S. & Crosthwaite, S. K. Natural antisense transcripts and long non-coding RNA in Neurospora crassa. *PLoS One* **9**, e91353, doi:10.1371/journal.pone.0091353 (2014).

148 Theodorou, M. K., Brookman, J. & Trinci, A. P. J. in *Methods in Gut Microbial Ecology for Ruminants* (eds Harinder P.S. Makkar & Christopher S. McSweeney) Ch. 2.4, 55-56 (IAEA, 2005).

149 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**, 2725-2729, doi:10.1093/molbev/mst197 (2013).

150 Chenna, R. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research* **31**, 3497-3500, doi:10.1093/nar/gkg500 (2003).

151 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539, doi:10.1038/msb.2011.75 (2011).

152 Theodorou, M. K., Davies, D. R., Nielsen, B. B., Lawrence, M. I. G. & Trinci, A. P. J. Determination of growth of anaerobic fungi on soluble and cellulosic substrates using a pressure transducer. *Microbiology* **141**, 671-678, doi:10.1099/13500872-141-3-671 (1995).

153 Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676, doi:10.1093/bioinformatics/bti610 (2005).

154 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

155 Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Research* **28**, 304-305 (2000).

156 Li, L., Stoeckert, C. J. J. & Roos, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research* **13**, 2178-2189, doi:10.1101/gr.1224503.candidates (2003).

157 Lam, K. K., LaButti, K., Khalak, A. & Tse, D. FinisherSC: a repeat-aware tool for upgrading de novo assembly using long reads. *Bioinformatics* **31**, 3207-3209, doi:10.1093/bioinformatics/btv280 (2015).

158 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, 563-569, doi:10.1038/nmeth.2474 (2013).

159 Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875, doi:10.1093/bioinformatics/bti310 (2005).

160 Kang, H. S. & Brady, S. F. Arimetamycin A: improving clinically relevant families of natural products through sequence-guided screening of soil metagenomes. *Angew Chem Int Ed Engl* **52**, 11063-11067, doi:10.1002/anie.201305109 (2013).

161 Curran, K. A. & Alper, H. S. Expanding the chemical palate of cells by combining systems biology and metabolic engineering. *Metab Eng* **14**, 289-297, doi:10.1016/j.ymben.2012.04.006 (2012).

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

163 Durand, R., Rascle, C., Fischer, M. & Fèvre, M. Transient expression of the β-glucuronidase gene after biolistic transformation of the anaerobic fungus Neocalimastix frontalis. *Current Genetics* **31**, 158-161, doi:10.1007/s002940050190 (1997).

164 Nagpal, R., Puniya, A. K., Sehgal, J. P. & Singh, K. Survival of anaerobic fungus Caecomyces sp. in various preservation methods: a comparative study. *Mycoscience* **53**, 427-432, doi:10.1007/s10267-012-0187-y (2012).

165 Yarlett, N. C., Yarlett, N., Orpin, C. G. & Lloyd, D. Cryopreservation of the anaerobic rumen fungus Neocallimastix patriciarum. *Letters in Applied Microbiology* **3**, 1-3, doi:10.1111/j.1472-765X.1986.tb01533.x (1986).

166 Sakurada, M., Tsuzuki, Y., Morgavi, D. P., Tomita, Y. & Onodera, R. Simple method for cryopreservation of an anaerobic rumen fungus using ethylene glycol and rumen fluid. *FEMS microbiology letters* **127**, 171-174 (1995).

167 Brownlee, A. G. A rapid DNA isolation procedure applicable to any refractory filamentous fungi. *Fungl Genetics Newsletter* **35** (1988).

168 Phillips, M. W. & Gordon, G. L. R. Growth characterstics on cellobiose of three different anaerobic fungi isolated from the ovine rumen. *Applied and Environmental Microbiology* **55**, 1695-1702 (1989).

169 Simbolo, M. *et al.* DNA qualification workflow for next generation sequencing of histopathological samples. *PLoS One* **8**, e62692, doi:10.1371/journal.pone.0062692 (2013).

170 Brookman, J. L., Ozkose, E., Rogers, S., Trinci, A. P. J. & Theodorou, M. K. Identification of spores in the polycentric anaerobic gut fungi which enhance their ability to survive. *FEMS microbiology ecology* **31**, 261-267, doi:10.1111/j.1574-6941.2000.tb00692.x (2000).

171 Brookman, J. L. & Nicholson, M. J. in *Methods in Gut Microbial Ecology for Ruminants* (eds H. P. S. Makkar & C. S. McSweeney) Ch. 4.3, 139-150 (Springer, 2005).

172 Lis, J. T. & Schleif, R. Size fractionation of double-stranded DNA by precipitation with polyethylene glycol. *Nucleic Acids Research* **2**, 383-390, doi:https://doi.org/10.1093/nar/2.3.383 (1975).

173 Ahn, S. J., Costa, J. & Emanuel, J. R. PicoGreen quantitation of DNA: Effective evaluation of samples pre- or post-PCR. *Nucleic Acids Research* **24**, 2623-2625, doi:10.1093/nar/24.13.2623 (1996).

174 Theodorou, M. K., Williams, B. a., Dhanoa, M. S., McAllan, A. B. & France, J. A simple gas production method using a pressure transducer to determine the fermentation kinetics of ruminant feeds. *Animal Feed Science and Technology* **48**, 185-197, doi:10.1016/0377-8401(94)90171-6 (1994).

175 Agapakis, C. M., Boyle, P. M. & Silver, P. A. Natural strategies for the spatial optimization of metabolism in synthetic biology. *Nat Chem Biol* **8**, 527-535, doi:10.1038/nchembio.975 (2012).

176 Zhou, K., Qiao, K., Edgar, S. & Stephanopoulos, G. Distributing a metabolic pathway among a microbial consortium enhances production of natural products. *Nat Biotechnol* **33**, 377-383, doi:10.1038/nbt.3095 (2015).

177 Zuroff, T. R. & Curtis, W. R. Developing symbiotic consortia for lignocellulosic biofuel production. *Appl Microbiol Biotechnol* **93**, 1423-1435, doi:10.1007/s00253-011-3762-9 (2012).

178 Dien, B. *et al.* Chemical composition and response to dilute-acid pretreatment and enzymatic saccharification of alfalfa, reed canarygrass, and switchgrass. *Biomass and Bioenergy* **30**, 880-891, doi:10.1016/j.biombioe.2006.02.004 (2006).

179 Seppälä, S., Solomon, K. V., Gilmore, S. P., Henske, J. K. & O'Malley, M. A. Mapping the membrane proteome of anaerobic gut fungi identifies a wealth of carbohydrate binding proteins and transporters. *Microbial Cell Factories* **15**, 212, doi:10.1186/s12934-016-0611-7 (2016).

180 Ali, B. R. *et al.* Cellulases and hemicellulases of the anaerobic fungus Piromyces constitute a multiprotein cellulose-binding complex and are encoded by multigene

families. *FEMS Microbioogy Letters* **125**, 15-21, doi:https://doi.org/10.1111/j.1574-6968.1995.tb07329.x (1995).

181 Couger, M. B., Youssef, N. H., Struchtemeyer, C. G., Liggenstoffer, A. S. & Elshahed, M. S. Transcriptomic analysis of lignocellulosic biomass degradation by the anaerobic fungal isolate Orpinomyces sp. strain C1A. *Biotechnol Biofuels* **8**, 208, doi:10.1186/s13068-015-0390-0 (2015).

182 Amore, A., Giacobbe, S. & Faraco, V. Regulation of Cellulase and Hemicellulase Gene Expression in Fungi. *Current Genomics* **14**, 230-249, doi:10.2174/1389202911314040002 (2013).

183 Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42**, D199-205, doi:10.1093/nar/gkt1076 (2014).

184 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**, 27-30 (2000).

185 Jeffries, T. W. Utilization of xylose by bacteria, yeasts, and fungi. *Advances in Biochemical Engineering/Biotechnology* **27**, 1-32, doi:10.1007/BFb0009101 (1983).

186 Bruinenberg, P. M., de Bot, P. H. M., van Dijken, J. P. & Scheffers, W. A. The role of redox balances in the anaerobic fermentation of xylose by yeasts. *European Journal of Applied Microbiology and Biotechnology* **18**, 287-292, doi:10.1007/BF00500493 (1983).

187 Schomburg, I. *et al.* BRENDA: A resource for enzyme data and metabolic information. *Trends in Biochemical Sciences* **27**, 47-49, doi:10.1016/S0968-0004(01)02027-8 (2002).

188 Yarlett, N., Orpin, C. G., Munn, E. A., Yarlett, N. C. & Greenwood, C. A. Hydrogenosomes in the rumen fungus Neocallimastix patriciarum. *The Biochemical Journal* **236**, 729-739 (1986).

189 Hackstein, J. H. P., Akhmanova, A., Boxma, B., Harhangi, H. R. & Voncken, F. G. J. Hydrogenosomes: Eukaryotic adaptations to anaerobic environments. *Trends in Microbiology* **7**, 441-447, doi:10.1016/S0966-842X(99)01613-3 (1999).

190 Muller, M. The hydrogenosome. *Journal of General Microbiology* **139**, 1879-2889, doi:10.1099/00221287-139-12-2879 (1993).

191 Muller, M. *et al.* Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol Mol Biol Rev* **76**, 444-495, doi:10.1128/MMBR.05024-11 (2012).

192 Hrdý, I. *et al.* Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* **432**, 618-622, doi:10.1038/nature03149 (2004).

193 Dodds, D. R. & Gross, R. a. Chemicals from Biomass. *Science* **318**, 1250-1251, doi:10.1126/science.1146356 (2007).

194 Berlin, A. *et al.* Inhibition of cellulase, xylanase and beta-glucosidase activities by softwood lignin preparations. *J Biotechnol* **125**, 198-209, doi:10.1016/j.jbiotec.2006.02.021 (2006).

195 Sanderson, K. Lignocellulose: A chewy problem. *Nature* **474**, S12-S14, doi:10.1038/474S012a (2011).

196 Martinez, D. *et al.* Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina). *Nat Biotechnol* **26**, 553-560, doi:10.1038/nbt1403 (2008).

197 James, T. Y. *et al.* Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443**, 818-822, doi:10.1038/nature05110 (2006).

198 Gancedo, J. M. Yeast carbon catabolite repression. *Microbiology and molecular biology reviews* **62**, 334-361 (1998).

199 Gorke, B. & Stulke, J. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nat Rev Microbiol* **6**, 613-624, doi:10.1038/nrmicro1932 (2008).

200 Nakari-Setala, T. *et al.* Genetic modification of carbon catabolite repression in Trichoderma reesei for improved protein production. *Appl Environ Microbiol* **75**, 4853-4860, doi:10.1128/AEM.00282-09 (2009).

201 Coradetti, S. T., Xiong, Y. & Glass, N. L. Analysis of a conserved cellulase transcriptional regulator reveals inducer-independent production of cellulolytic enzymes in Neurospora crassa. *Microbiologyopen* **2**, 595-609, doi:10.1002/mbo3.94 (2013).

202 Bakir, U., Yavascaoglub, S., Guvencb, F. & Ersayinb, A. An endo-beta-1,4-xylanase from Rhizopus oryzae: production, partial purification and biochemical characterization. *Enzyme and Microbial Technology* **29**, 328-334, doi:10.1016/S0141-0229(01)00379-9 (2001).

203 Todero Ritter, C. E., Camassola, M., Zampieri, D., Silveira, M. M. & Dillon, A. J. Cellulase and Xylanase Production by Penicillium echinulatum in Submerged Media Containing Cellulose Amended with Sorbitol. *Enzyme Res* **2013**, 240219, doi:10.1155/2013/240219 (2013).

204 Hrmová, M., Biely, P. & Vršanská, M. Specificity of cellulase and β-xylanase induction in Trichoderma reesei QM 9414. *Archives of Microbiology* **144**, 307-311, doi:10.1007/BF00410968 (1986).

205 Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95** (1999).

206 Payne, C. M. *et al.* Fungal cellulases. *Chem Rev* **115**, 1308-1448, doi:10.1021/cr500351c (2015).

207 Wong, D. W. S. Structure and action mechanism of ligninolytic enzymes. *Appl. Biochem. Biotechnol.* **157**, 174-209, doi:10.1007/s12010-008-8279-z (2009).

208 Orpin, C. G. The Role of Ciliate Protozoa and Fungi in the Rumen Digestion of Plant Cell Walls. *Animal Feed Science and Technology* **10**, 121-143, doi:10.1016/0377-8401(84)90003-8 (1984).

209 Akin, D. E. & Bennert, R. Degradation of Polysaccharides and Lignin by Ruminal Bacteria and Fungi. *Applied and Environmental Microbiology* **54**, 1117-1125 (1988).

210 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

211 Shi, J. & Walker, M. G. Gene set enrichment analysis (GSEA) for interpreting gene expression profiles. *Curr. Bioinf.* **2**, 133-137, doi:10.2174/157489307780618231 (2007).

212 Teunissen, M. J., Op den Camp, H. J. M., Orpin, C. G., Huis in 't Veld, J. H. J. & Vogels, G. D. Comparison of growth characteristics of anaerobic fungi isolated from ruminant and non-ruminant herbivores during cultivation in a defined medium.

Journal of General Microbiology **137**, 1401-1408, doi:10.1099/00221287-137-6-1401 (1991).

213 Manavalan, T., Manavalan, A. & Heese, K. Characterization of Lignocellulolytic Enzymes from White-Rot Fungi. *Curr. Microbiol.* **70**, 485-498, doi:10.1007/s00284-014-0743-0 (2015).

214 Schwarz, W. H. The cellulosome and cellulose degradation by anaerobic bacteria. *Applied Microbiology and Biotechnology* **56**, 634-649, doi:10.1007/s002530100710 (2001).

215 Dewi Puspita, I., Kamagata, Y., Tanaka, M., Asano, K. & Nakatsu, C. H. Are Uncultivated Bacteria Really Uncultivable? *Microbes and Environments* **27**, 356-366, doi:10.1264/jsme2.ME12092 (2012).

216 Segal, D. J. & Meckler, J. F. Genome engineering at the dawn of the golden age. *Annu Rev Genomics Hum Genet* **14**, 135-158, doi:10.1146/annurev-genom-091212-153435 (2013).

217 Behling, R., Valange, S. & Chatel, G. Heterogeneous catalytic oxidation for lignin valorization into valuable chemicals: what results? What limitations? What trends? *Green Chem.*, 10.1039/C1035GC03061G, doi:10.1039/C5GC03061G (2016).

218 Xu, J., Xie, X., Wang, J. & Jiang, J. Directional liquefaction coupling fractionation of lignocellulosic biomass for platform chemicals. *Green Chem.*, 10.1039/C1035GC03070F, doi:10.1039/C5GC03070F (2016).

219 Karkas, M. D., Matsuura, B. S., Monos, T. M., Magallanes, G. & Stephenson, C. R. J. Transition-metal catalyzed valorization of lignin: the key to a sustainable carbon-neutral future. *Org. Biomol. Chem.* **14**, 1853-1914, doi:10.1039/C5OB02212F (2016).

220 Beckham, G. T., Johnson, C. W., Karp, E. M., Salvachua, D. & Vardon, D. R. Opportunities and challenges in biological lignin valorization. *Curr. Opin. Biotechnol.* **42**, 40-53, doi:10.1016/j.copbio.2016.02.030 (2016).

221 Hofrichter, M. Review: lignin conversion by manganese peroxidase (MnP). *Enzyme Microb. Technol.* **30**, 454-466, doi:10.1016/S0141-0229(01)00528-2 (2002).

222 Eggert, C., Temp, U. & Eriksson, K.-E. L. The ligninolytic system of the white rot fungus Pycnoporus cinnabarinus: purification and characterization of the laccase. *Appl. Environ. Microbiol.* **62**, 1151-1158 (1996).

223 Eggert, C., Temp, U. & Eriksson, K.-E. L. Laccase is essential for lignin degradation by the white-rot fungus Pycnoporus cinnabarinus. *FEBS Lett.* **407**, 89-92, doi:10.1016/S0014-5793(97)00301-3 (1997).

224 Levasseur, A. *et al.* The genome of the white-rot fungus Pycnoporus cinnabarinus: a basidiomycete model with a versatile arsenal for lignocellulosic biomass breakdown. *BMC genomics* **15**, 486, doi:10.1186/1471-2164-15-486 (2014).

225 Liers, C., Arnstadt, T., Ullrich, R. & Hofrichter, M. Patterns of lignin degradation and oxidative enzyme secretion by different wood- and litter-colonizing basidiomycetes and ascomycetes grown on beech-wood. *FEMS microbiology ecology* **78**, 91-102, doi:10.1111/j.1574-6941.2011.01144.x (2011).

226 Spies, D. & Ciaudo, C. Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis. *Comput Struct Biotechnol J* **13**, 469-477 (2015).

227 Zhao, Q.-Y., Gratten, J., Restuadi, R. & Li, X. Mapping and differential expression analysis from short-read RNA-Seq data in model organisms. *Quant. Biol.* **4**, 22-35, doi:10.1007/s40484-016-0060-7 (2016).

228 Vikman, P., Fadista, J. & Oskolkov, N. RNA sequencing: current and prospective uses in metabolic research. *J. Mol. Endocrinol.* **53**, R93-R101, doi:10.1530/JME-14-0170 (2014).

229 Hui, P. Next generation sequencing: Chemistry, technology and applications. *Top. Curr. Chem.* **336**, 1-18, doi:10.1007/128_2012_329 (2014).

230 McGettigan, P. A. Transcriptomics in the RNA-seq era. *Curr. Opin. Chem. Biol.* **17**, 4-11, doi:10.1016/j.cbpa.2012.12.008 (2013).

231 Bhadauria, V., Popescu, L., Zhao, W.-S. & Peng, Y.-L. Fungal transcriptomics. *Microbiol. Res.* **162**, 285-298, doi:10.1016/j.micres.2007.06.006 (2007).

232 Couturier, M. *et al.* Enhanced degradation of softwood versus hardwood by the white-rot fungus Pycnoporus coccineus. *Biotechnology for Biofuels* **8**, 1-16, doi:10.1186/s13068-015-0407-8 (2015).

233 Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M. & Henrissat, B. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnol. Biofuels* **6**, 41, doi:10.1186/1754-6834-6-41 (2013).

234 Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat Meth* **10**, 623-629, doi:10.1038/nmeth.2483 (2013).

235 Stanke, M., Tzvetkova, A. & Morgenstern, B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* **7** **Suppl 1**, S11.11-18 (2006).

236 Sannigrahi, P., Ragauskas, A. J. & Tuskan, G. A. Poplar as a feedstock for biofuels: a review of compositional characteristics. *Biofuels, Bioprod. Biorefin.* **4**, 209-226, doi:10.1002/bbb.206 (2010).

237 Herpoel, I., Moukha, S., Lesage-Meessen, L., Sigoillot, J. C. & Asther, M. Selection of Pycnoporus cinnabarinus strains for laccase production. *FEMS Microbiol. Lett.* **183**, 301-306, doi:10.1016/S0378-1097(99)00616-3 (2000).

238 Geng, X. & Li, K. Degradation of non-phenolic lignin by the white-rot fungus Pycnoporus cinnabarinus. *Appl. Microbiol. Biotechnol.* **60**, 342-346, doi:10.1007/s00253-002-1124-3 (2002).

239 Morgenstern, I., Robertson, D. L. & Hibbett, D. S. Characterization of three mnp genes of Fomitiporia mediterranea and report of additional class II peroxidases in the order hymenochaetales. *Appl. Environ. Microbiol.* **76**, 6431-6440, doi:10.1128/AEM.00547-10 (2010).

240 Rohr, C. O., Levin, L. N., Mentaberry, A. N. & Wirth, S. A. A first insight into Pycnoporus sanguineus BAFC 2126 transcriptome. *PLoS One* **8**, e81033/81031-e81033/81014, doi:10.1371/journal.pone.0081033 (2013).

241 Wariishi, H., Akileswaran, L. & Gold, M. H. Manganese peroxidase from the basidiomycete Phanerochaete chrysosporium: spectral characterization of the oxidized states and the catalytic cycle. *Biochemistry* **27**, 5365-5370, doi:10.1021/bi00414a061 (1988).

242 Banci, L., Bertini, I., Dal ozzo, L., Del Conte, R. & Tien, M. Monitoring the Role of Oxalate in Manganese Peroxidase. *Biochemistry* **37**, 9009-9015, doi:10.1021/BI972879+ (1998).

243 Wariishi, H., Valli, K. & Gold, M. H. Oxidative cleavage of a phenolic diarylpropane lignin model dimer by manganese peroxidase from Phanerochaete chrysosporium. *Biochemistry* **28**, 6017-6023, doi:10.1021/bi00440a044 (1989).

244 Piontek, K., Smith, A. T. & Blodig, W. Lignin peroxidase structure and function. *Biochem Soc Trans* **29**, 111-116 (2001).

245 Ruiz-Duenas, F. J. *et al.* Manganese Oxidation Site in Pleurotus eryngii Versatile Peroxidase: A Site-Directed Mutagenesis, Kinetic, and Crystallographic Study. *Biochemistry* **46**, 66-77, doi:10.1021/bi061542h (2007).

246 Jung, S.-J., Kim, S.-H. & Chung, I.-M. Comparison of lignin, cellulose, and hemicellulose contents for biofuels utilization among 4 types of lignocellulosic crops. *Biomass Bioenergy* **83**, 322-327, doi:10.1016/j.biombioe.2015.10.007 (2015).

247 DeMartini, J. D. *et al.* Investigating plant cell wall components that affect biomass recalcitrance in poplar and switchgrass. *Energy Environ. Sci.* **6**, 898-909, doi:10.1039/c3ee23801f (2013).

248 Zhou, G., Taylor, G. & Polle, A. FTIR-ATR-based prediction and modelling of lignin and energy contents reveals independent intra-specific variation of these traits in bioenergy poplars. *Plant Methods* **7**, 9, doi:10.1186/1746-4811-7-9 (2011).

249 Yin, D. *et al.* Structure-function characterization reveals new catalytic diversity in the galactose oxidase and glyoxal oxidase family. *Nat. Commun.* **6**, 10197, doi:10.1038/ncomms10197 (2015).

250 Hernandez-Ortega, A. *et al.* Stereoselective Hydride Transfer by Aryl-Alcohol Oxidase, a Member of the GMC Superfamily. *ChemBioChem* **13**, 427-435, doi:10.1002/cbic.201100709 (2012).

251 Cragg, S. M. *et al.* Lignocellulose degradation mechanisms across the Tree of Life. *Curr. Opin. Chem. Biol.* **29**, 108-119, doi:10.1016/j.cbpa.2015.10.018 (2015).

252 Alberts, B. *et al. Molecular Biology of the Cell*. 4 edn, (Garland Science, 2002).

253 Kell, D. B., Swainston, N., Pir, P. & Oliver, S. G. Membrane transporter engineering in industrial biotechnology and whole cell biocatalysis. *Trends Biotechnol* **33**, doi:10.1016/j.tibtech.2015.02.001 (2015).

254 Hector, R. E., Qureshi, N., Hughes, S. R. & Cotta, M. A. Expression of a heterologous xylose transporter in a Saccharomyces cerevisiae strain engineered to utilize xylose improves aerobic xylose consumption. *Appl Microbiol Biotechnol* **80**, doi:10.1007/s00253-008-1583-2 (2008).

255 Young, E., Poucher, A., Comer, A., Bailey, A. & Alper, H. Functional survey for heterologous sugar transport proteins, using Saccharomyces cerevisiae as a host. *Appl Environ Microbiol* **77**, doi:10.1128/aem.02651-10 (2011).

256 Ha, S. J. *et al.* Energetic benefits and rapid cellobiose fermentation by Saccharomyces cerevisiae expressing cellobiose phosphorylase and mutant cellodextrin transporters. *Metab Eng* **15**, doi:10.1016/j.ymben.2012.11.005 (2013).

257 Boyarskiy, S. & Tullman-Ercek, D. Getting pumped: membrane efflux transporters for enhanced biomolecule production. *Curr Opin Chem Biol* **28**, doi:10.1016/j.cbpa.2015.05.019 (2015).

258 Dunlop, M. J. *et al.* Engineering microbial biofuel tolerance and export using efflux pumps. *Mol Syst Biol* **7**, doi:10.1038/msb.2011.21 (2011).

259 Saier, M. H., Jr., Tran, C. V. & Barabote, R. D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res* **34**, D181-186, doi:10.1093/nar/gkj001 (2006).

260 Petersen, T. N., Brunak, S., Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**, doi:10.1038/nmeth.1701 (2011).

261 Lagerström, M. C. & Schiöth, H. B. Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov* **7**, doi:10.1038/nrd2518 (2008).

262 Kroeze, W. K., Sheffler, D. J. & Roth, B. L. G-protein-coupled receptors at a glance. *J Cell Sci* **116**, 4867-4869, doi:10.1242/jcs.00902 (2003).

263 Fredriksson, R., Lagerström, M. C., Lundin, L. G. & Schiöth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol* **63**, doi:10.1124/mol.63.6.1256 (2003).

264 Kolakowski, L. F. GCRDb: a G-protein-coupled receptor database. *Receptors Channels* **2** (1994).

265 Krishnan, A., Almén, M. S., Fredriksson, R. & Schiöth, H. B. The origin of GPCRs: identification of mammalian like Rhodopsin, Adhesion, Glutamate and Frizzled GPCRs in fungi. *PLoS ONE* **7**, doi:10.1371/journal.pone.0029817 (2012).

266 Pin, J. P., Galvez, T. & Prézeau, L. Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacol Ther* **98**, doi:10.1016/s0163-7258(03)00038-x (2003).

267 O'Hara, P. J. *et al.* The ligand-binding domain in metabotropic glutamate receptors is related to bacterial periplasmic binding proteins. *Neuron* **11**, doi:10.1016/0896-6273(93)90269-w (1993).

268 Felder, C. B., Graul, R. C., Lee, A. Y., Merkle, H. P. & Sadee, W. The Venus flytrap of periplasmic binding proteins: an ancient protein module present in multiple drug receptors. *AAPS PharmSci* **1**, doi:10.1208/ps010202 (1999).

269 Marin-Rodriguez, M. C. Pectate lyases, cell wall degradation and fruit softening. *J Exp Bot* **53**, doi:10.1093/jxb/erf089 (2002).

270 Käll, L., Krogh, A. & Sonnhammer, E. L. L. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**, doi:10.1016/j.jmb.2004.03.016 (2004).

271 Krogh, A., Larsson, B., Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, doi:10.1006/jmbi.2000.4315 (2001).

272 Gonzalez, R. & Scazzocchio, C. A rapid method for chromatin structure analysis in the filamentous fungus Aspergillus nidulans. *Nucleic Acids Research* **25**, 3955-3956 (1997).

273 Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* **2010**, pdb prot5384, doi:10.1101/pdb.prot5384 (2010).

274 Simon, J. M., Giresi, P. G., Davis, I. J. & Lieb, J. D. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat Protoc* **7**, 256-267, doi:10.1038/nprot.2011.444 (2012).

275 Mondo, S. J. *et al.* Widespread adenine N6-methylation of active genes in fungi. *Nat Genet*, doi:10.1038/ng.3859 (2017).