GENOMIC AND TRANSCRIPTOMIC ANALYSIS OF THE ANEROBIC FUNGUS ORPINOMYCES STRAIN C1A, A VERSATILE BIODEGRADER OF PLANT BIOMASS

By

MATTHEW BRIAN COUGER

Bachelor of Science in Biochemistry and Molecular Biology

Stillwater Oklahoma

2008

Bachelor of Science in Psychology

Stillwater Oklahoma

2008

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY or EDUCATION
December, 2015

GENOMIC AND TRANSCRIPTOMIC ANALYSIS OF THE ANEROBIC FUNGUS ORPINOMYCES C1A, A VERSATILE BIODEGRADER

Dissertation Approved:

Dr. Mostafa S. Elshahed

Dissertation Adviser

Dr. Wouter Hoff

Dr. Ed Shaw

Dr. Rolf A Prade

Dr. Stephen M. Marek

Name: MATTHEW BRIAN COUGER

Date of Degree: DECEMBER, 2015

Title of Study: GENOMIC AND TRANSCRIPTOMIC ANALYSIS OF THE ANEROBIC FUNGUS ORPINOMYCES STRAIN C1A, A VERSATILE BIODEGRADER OF PLANT BIOMASS
Major Field: MICROBIOLOGY AND MOLEUCLAR GENETICS

Abstract: The anaerobic fungi represent a basal fungal lineage, members of which reside in the rumen and alimentary tract of herbivores. Due to their reported capacity to degrade plant materials, the anaerobic fungi have recently been touted as promising agents for biofuel production. In the first part of this thesis, I present the first reported genomic analysis of a member of the anaerobic gut fungi, *Orpinomyces* sp. strain C1A. The genome of strain C1A was sequenced using a combination of Illumina and PacBio SMRT technologies. The large genome (100.95 Mb, 16,347 genes) displayed extremely low G+C content (17.0%), large non-coding intergenic regions (73.1%), a proliferation of microsatellite repeats (4.9%), and multiple gene duplications. Comparative genomic analysis identified multiple genes and pathways that are absent in Dikarya genomes but present in basal fungal lineages and/or non-fungal Opisthokonts. Analysis of the lignocellulolytic machinery in the C1A genome revealed an extremely rich repertoire, with evidence of horizontal gene acquisition from multiple bacterial lineages. Experimental analysis indicated that strain C1A is a remarkable biomass degrader, capable of simultaneous saccharification and fermentation of the cellulosic and hemicellulosic fractions in multiple untreated grasses and crop residues examined, with the process significantly enhanced by mild pretreatments.
In the second part of my thesis, I analyzed the transcriptomic profiles of C1A when grown on four different types of lignocellulosic biomass (alfalfa, energy cane, corn stover, and sorghum) versus a soluble sugar monomer (glucose). My overall goal was to understand the mechanistic and regulatory basis of biomass deconstruction in anaerobic fungi. Transcriptomic sequencing yielded a total of 468.2 million reads (70.2 GB) that were assembled into 27,506 distinct transcripts. Transcripts belonging to Carbohydrate Active Enzymes (CAZYmes) included 385, 246, and 44 transcripts belonging to 44, 13, and 8 different glycoside hydrolases (GH), carbohydrate esterases (CE), and polysaccharide lyases (PL) families, respectively. Examination of CAZyme transcriptional patterns indicates that strain C1A constitutively transcribes a high baseline level of CAZyme transcripts on glucose. Although growth on lignocellulosic biomass

substrates was associated with a significant increase in transcriptional levels in few GH families, including the highly transcribed GH1 β-glucosidase, GH6 cellobiohydrolase, and GH9 endogluconase, the transcriptional levels of the majority of CAZymes families and transcripts were not significantly altered in glucose grown versus lignocelluosic biomass-grown cultures. Further, strain C1A co-transcribes multiple functionally redundant enzymes for cellulose and hemicellulose saccharification that are mechanistically and structurally distinct. Analysis of fungal dockerin domain (FDD)-containing transcripts strongly suggests that anaerobic fungal cellulosomes represent distinct catalytic units capable of independently attacking and converting intact plant fibers to sugar monomers. Collectively, these results demonstrate that strain C1A achieves fast, effective biomass degradation by the simultaneous employment of a wide array of constitutively-transcribed cellulosomal-bound and free enzymes with considerable functional overlap.

The thesis hence represents the first in-depth evaluation of the genome and transcriptome of a member of this poorly studied group of fungi. Collectively, my work has revealed multiple novel insights into the metabolic capabilities, cell biology, and genomic architecture of anaerobic fungi such as the presence of unique pathways and processes not encountered in higher fungi, genomic features shaped by its unique evolutionary trajectory, extensive lignocellulolytic gene repertoire, and regulatory mechanisms employed to achieve fast and efficient biomass degradation within the herbivore gut.

Would like to thank Mostafa Elshaed, Noah Youseef, Audra Ligenstoffer, and Christ Structymer for all the excellent collaboration over the years. I would also like to thank the Microbiology and Moleucular Genetics Department for my education.

TABLE OF CONTENTS

.

LIST OF TABLES

LIST OF FIGURES

Figure Page

Preface

Currently, the majority of the world's energy production is derived from non-renewable fossil fuels such as crude oil, coal, and natural gas. The heavy reliance on these fuels creates a multitude of problems both in the near future and long term for energy production. Continuous reliance on fossil fuels leads to the depletion of extractable reserves of these non-renewable resources. The production and distribution of these fuels are subject to spasmodic interruptions due to natural disasters or geopolitical events, which can create economic instabilities in developed and emerging economies. More importantly, the use of fossil fuels contributed to the global rise in atmospheric CO_2 levels, a process linked to global warming and the increase in the frequency and intensity of severe weather patterns.

Due to these aforementioned issues, a great emphasis has been placed on the development of alternate renewable energy sources that are sustainable and environmentally benign. Proposed alternate energy sources include solar energy, nuclear energy, wind power, hydroelectric power and biofuel production. Of these biofuels production has a distinct advantage: The final product (ethanol, longer chain alcohols, biodiesel, or alkanes) could directly be utilized using the existing energy infrastructure. Currently, the majority of biofuel production is achieved through saccharification and fermentation of food crops, such as corn or sugarcane. Nevertheless, the production of biofuels from edible crops in a hungry world represents a moral dilemma. Lignocellulosic biomass is defined as raw, non-edible plant biomass that is mainly composed of sugar (cellulose and hemicellulose) and aromatic (lignin) polymers. Lignocellulosic biomass represents a vast, renewable, and underutilized resource for biofuels production. Biofuel derived from lignocellulosic biomass (often referred to as cellulosic or second generation biofuels)

alleviates the moral stigma of using edible crops for biofuel production, as well as the need for the expansion of farming acreage.

Currently, the production of biofuels from lignocellulosic biomass is technically feasible, but too expensive to allow for its widespread utilization and commercialization. In this process, enzymes are utilized to extract sugar from plant polymers, and the produced sugars are then converted into biofuel using dedicated sugar-metabolizing microorganisms. The main plant polymers targeted for biofuel production in lignocellulosic biomass are cellulose and hemicellulose, both of which are structural components of plant cell walls and are chemically bound to a variety of complex macromolecules, mainly lignin. Enzymatic treatment of lignocellulosic biomass is a complex and costly endeavor requiring a mixture of multiple enzymes to depolymerize cellulose and hemicellulose.

Anaerobic fungi reside in the rumen and alimentary tract of ruminant and non-ruminant herbivores. Members of the anaerobic fungi represent a distinct fungal phylum (The Neocallimastigomycota) and have previously been shown to be efficient degraders of plant biomass. As such, I reasoned that the anaerobic fungi could represent promising agents for biofuel production from lignocellulosic biomass. Such reasoning represents the basis of my dissertation where I critically evaluate such an idea using genomic and transcriptomic approaches.

Chapter I of this dissertation describes the history of anaerobic fungi, the current barriers for lignocellulosic biofuel production, and the initial work conducted by prior colleagues in Dr. Elshahed laboratory to demonstrate anaerobic fungi's remarkable ability to degrade various types of lignocellulosic biomass.

In chapter II, I present my analysis of the genomic architecture and inferred metabolic

capabilities of an anaerobic fungal isolate: *Orpinomyces* sp. strain C1A, henceforth referred to as strain C1A. When I started my Ph.D. research, multiple laboratories around the work were attempting to sequence genomes of various anaerobic fungal isolates. However, these efforts largely failed due to the extreme A/T base composition bias and complexity of repeats in anaerobic fungal genomes. Leveraging a combination of emerging technologies, our group produced the first genomic sequence for anaerobic fungi, which gave key insights into the biology and ability to decompose lignocellulosic biomass by anaerobic fungi. In chapter II, I present the first genome sequence produced from anaerobic fungi, and use the genome to highlight the unique evolutionary history of this group of fungi, as well as its enzymatic machinery for lignocellulosic biomass decomposition. This research was published in the journal Applied and Environmental Microbiology.

Chapter III in this dissertation highlights present an extensive analysis of the transcriptional patterns of various biofuel production related genes in strain C1A when grown multiple plant substrates. This extensive quantitative transcriptomic analysis produced critical information on the nature of enzymes putatively expressed by strain C1A to decompose complex lignocellulosic plants, as well as the overall strategy utilized by strain C1A to achieve fast and efficient degradation of plant biomass. This research has been accepted for publication in the journal *Biotechnology for Biofuels.*

CHAPTER I

HISTORY OF ANEROBIC FUNGI AND THEIR POTEINTAL UTILITY AS A NOVEL PLATFORM FOR BIOFUEL PRODUCTION FROM LIGNOCELLULOSIC BIOMASS

History and Functional Assessment of Anaerobic Fungi Anaerobic Fungi are a group of microorganisms that reside in the rumen and alimentary tract of many ruminant, pseudoruminant, and nonruminant herbivores (2,10,26,33) Although flagellated zoospores of anaerobic fungi (AF) were observed as early as 1910, definitive proof that AF are an important constituent of the cow rumen came relatively late (2) due to the morphological similarity between anaerobic fungal spores and ciliate protozoa in the rumen (43). It was not until the 1970's that these organisms where correctly classified as fungi by Orpin (39) based on the presence of chitin in their cell walls. Subsequent research by Orpin (47) demonstrated that these rumen fungi resemble fungi in the Chytridiomycota, with flagellated zoospores typically encountered in soil and aquatic habitats. Microscopic analysis of rumen material also demonstrated that the flagellated zoospores of anaerobic fungi attach to and germinate on ingested plant material in the rumen (26). Experimental evidence subsequently demonstrated their capability to degrade cellulose, a major component of the plant cell wall (10, 55) in an *in-vitro* setting (2). Soon afterwards, enzymatic activities mediating the degradation of various plant polymers were identified in anaerobic fungal cultures, further demonstrating their role in plant biomass degradation in the rumen (9,33,44). This capacity, which ensures that anaerobic fungi are a critical portion of the rumen ecosystems, allow fungi to fully process complex plants cell walls into sugar monomers that could either be fermented by the anaerobic fungi for energy production or taken up by the rumen prokaryotic community (42,44). Due to the complexity of plant cell walls the process of their digestion and fermentation requires many distinct enzymes necessary to dismantle the plants cell walls into usable sugars. These properties make the study of anaerobic fungi extremely important both due to their rich unique biological history and for their potential as a transformative agent in the production of lignocellulosic biofuel.

Lignocellulosic Biofuels: Current Productions and Obstacles. Currently, the majority of the world's energy is produced from non-renewable fossil fuels (24). This heavy reliance on fossil fuels for energy production is problematic. Extractable reserves of this non-renewable resource are fast diminishing. The use of fossil fuels contributes to global warming and climate change, as well as to geopolitical conflicts (7,62). Development of renewable energy resources has the potential to fulfill current and future energy demand while avoiding issues associated with fossil fuel production outlined above (7,62). These renewable energy sources include nuclear energy, wind turbine energy, hydroelectric energy, solar energy and bioenergy. Each one of these energy sources has great potential, as well as significant drawbacks. Nevertheless, biofuel, the production of alcohols or hydrocarbon derivatives from plant sources, provides two distinct advantages compared to other methods (8). First, the growth of the plant's used for the production of biofuels remove CO_2 from the atmosphere, thus lowering the atmospheric concentration of this greenhouse gas. Second, many of the current biofuel production schemes generate ethanol or alkanes as the final product, both of which are compatible with the current transportation infrastructure and hence could readily be utilized for transportation.

Currently, the majority of biofuel is produced from simple fermentable sugars from edible crops such as corn or sugarcane, a process referred to as "first generation biofuels". First generation biofuels have provided a considerable input to current energy production, particularly as the ethanol component to mobile transportation fuel. However, the process has multiple drawbacks (11). Most current biomass inputs for fermentation for the production of biofuels are also feedstocks, the most common of which is corn in the US. Corn is consumed for food, and hence its utilization for energy production raises legitimate ethical concerns. It is also grown on fertile farmlands and requires extensive application of fertilizers. Another key barrier that

prevents first generation biofuels from providing a greater share of the world's energy is the amount of energy or fuel produced per kg of input (29). The simple sugars used for input for first generation fermentation make up a minority of the carbohydrates in a plant that theoretically could be used for fermentation. The majority of the carbohydrates are stored in long, complex, and interwoven polymers in the stalk, leaves, and other structures within the plant (63).

The production of biofuel from these complex and non-edible plant substrates is a process that has been referred to as lignocellulose biofuel, or second generation biofuels (29). Such process has two major advantages over first generation biofuels. First, the amount of fermentable sugars (g sugar/Kg plant material) stored in lignocellulosic biomass is far greater than fermentable sugars present in corn or other feedstocks (29,63). The second is that many lignocellulosic substrates, such as switchgrass and energy cane (34,59) produce an extremely high yield of plant mass per acre and can be grown on marginal lands (61) that are unsuitable for the production of crops (35). For these reasons, lignocellulosic biofuels hold great promise, and could theoretically provide a large share of energy needs (56,62). Unfortunately, this promise has not yet been realized primarily due to the difficulty and economic inefficiency for biofuel production from lignocellulosic feedstocks (8,62).

Recalcitrance of Plant Cell Wall is the Primary Barrier for Lignocellulosic Biomass Conversion Efficiency. The primary source of fermentable material for biofuel from lignocellulosic biomass comes from the liberation of polymers within the plant cell walls followed by their subsequent depolymerization into sugar monomers. In contrast to first generation biofuel production, which only requires a minimal enzymatic input to process into sugar monomers, lignocellulosic biomass requires a large variety of enzymes to efficiently process lignocellulosic biomass into fermentable sugars (11,21,29). Plant cell walls contains four

major components or polymers Cellulose, Hemicellulose, Pectin and Lignin, which are highly interlinked to provide a recalcitrant and highly stable cell wall (6,55,60,63). Due to the unique structure of each of these polymers and the fact that they are highly interlinked within the cell wall, a separate set of enzymes is required to depolymerize polymer (Figure 1-1) (55,58).

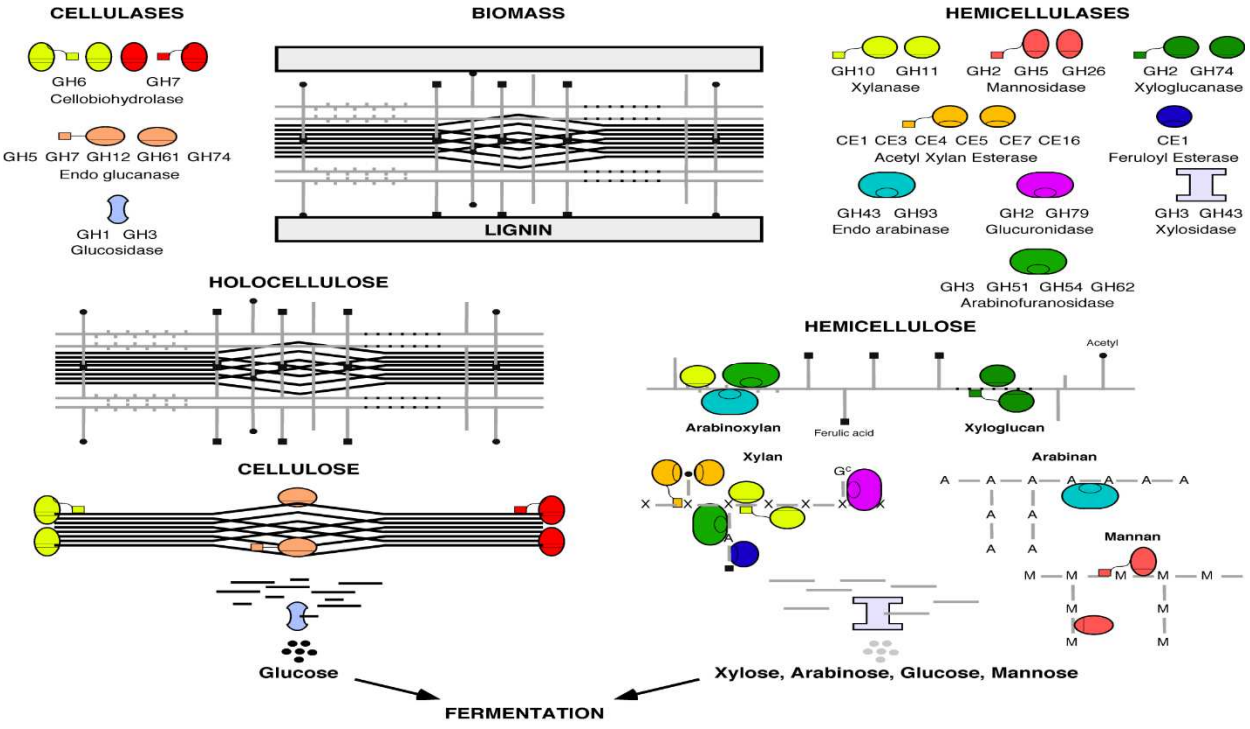


Figure 1-1 Enzymatic Components required for plant cell wall degradation, reproduced with permission from reference (12)

The need for a large amount of unique enzymes for full depolymerization is, obviously, costly (37,51). The required enzymes must be mass produced either in the original microorganism or in a heterologous system (16,17,45), purified and applied to each saccharification cycle (51) in

large quantities. In addition to costs associated with enzyme production, harsh, expensive, and environmentally pernicious pretreatments methodologies are employed to expose plant polymers to enzymatic action. The combined costs of enzymes and pretreatment renders the process of *in vitro* saccharification of lignocellulosic biomass an extremely costly endeavor

Anaerobic Fungi: A Prosing Agent for Lignocellulosic Biofuel Production

Anaerobic fungi possess multiple traits that render them promising agents for biofuel production from lignocellulosic biomass: They are able to efficiently colonize and penetrate plant material as well as the ability for anaerobic fermentative growth on plant mass. Anaerobic fungi have faced considerable evolutionary pressure to maintain a highly efficient system for the decomposition of lignocellulose, as this the major carbon source present in their natural habitat. In addition the fungi, which possess a higher native energy requirement for replication that bacteria would need to be able to effectively compete with the highly evolved bacterial portion of the rumen microbiome, which requires less energy to replicate (31).

To provide a greater understanding of anaerobic fungi and their experimentally demonstrated lignocellulosic capabilities, analysis of their genome and transcriptomic response to growth on biomass would be required. Despite their experimentally-demonstrated lignocellulolytic capabilities, little research, if any, was conducted to analyze the genome of these unique fungi since they were discovered. However, this was not due to lack of interest, but due to a highly abnormal base composition of the anaerobic fungal genomes. Initial research on individual genes from anaerobic fungi demonstrated an extreme A/T bias in genes averaging over 80% within the analyzed genes (20,28). This fact, coupled with the larger genome size compared to many other microorganisms, presented a unique challenge. Sequencing technologies such a Roche 454 could produce the necessary throughput for a genome assembly, but the

technology's inaccuracy in calling homopolymers would hinder accurate base pair calling within the extremely homopolymer-rich anaerobic fungal genomes. Illumina sequence by synthesis sequencing technology on the other hand, has both the capacity to produce a genomic assembly, as well as the ability to sequence the homopolymer regions. However, the short read output of Illumina sequences hinders accurate assembly of produced short reads into large contigs. In my dissertation, I undertook the challenge of producing the first genome assembly of an anaerobic fungal isolate: *Orpinomyces* sp. strain C1A using a combination of sequencing technologies and an innovative genome assembly approach. In addition, I conducted a thorough transcriptomic analysis to understand the transcriptional landscape of strain C1A when grown on soluble substrates e.g. glucose, compared to lignocellulose biomass substrates. This thesis represent the first comprehensive evaluation of the growth and lignocellulosic capabilities of anaerobic fungi using –omics based approaches and provides valuable tools and insights into the physiological characteristics and metabolic capabilities of these fascinating, yet-poorly studies group of microorganisms.

References

1. **Kuhnel S, Schols HA, Gruppen H.** 2011. Aiming for the complete utilization of sugar-beet pulp: Examination of the effects of mild acid and hydrothermal pretreatment followed by enzymatic digestion. Biotechnol Biofuels **4:**14.
2. **Orpin CG.** 1985. Association of rumen ciliate populations with plant particles in vitro. Microb Ecol **11:**59-69.
3. **Georgelis N, Nikolaidis N, Cosgrove DJ.** 2015. Bacterial expansins and related proteins from the world of microbes. Appl Microbiol Biotechnol **99:**3807-3823.

4. **Georgelis N, Nikolaidis N, Cosgrove DJ.** 2015. Bacterial expansins and related proteins from the world of microbes. Appl Microbiol Biotechnol **99:**3807-3823.

5. **Wubah DA, Akin DE, Borneman, WS.** 1993. Biology, fiber-degradation, and enzymology of anaerobic zoosporic fungi. Crit Rev Microbiol **19:**99-115.

6. **Himmel ME, Ding SY, Johnson DK, Adney WS, Nimlos MR, Brady JW, Foust TD.** 2007. Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science **315:**804-807.

7. **Herrera S.** 2006. Bonkers about biofuels. Nat Biotechnol **24:**755-760.

8. **Schubert C.** 2006. Can biofuels finally take center stage? Nat Biotechnol **24:**777-784.

9. **Ljungdahl LG.** 2008. 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.

10. **Bauchop T, Mountfort DO.** 1981. Cellulose fermentation by a rumen anaerobic fungus in both the absence and the presence of rumen methanogens. Appl Environ Microbiol **42:**1103-1110.

11. **Klein-Marcuschamer D, Oleskowicz-Popiel P, Simmons BA, Blanch HW.** 2012. The challenge of enzyme cost in the production of lignocellulosic biofuels. Biotechnol Bioeng **109:**1083-1087.

12. **Tian SQ, Wang ZY, Fan ZL, Zuo LL.** 2012. Comparison of ultrasonic and CO_2 laser pretreatment methods on enzyme digestibility of corn stover. Int J Mol Sci **13:**4141-4152.

13. **Piotrowski JS, Zhang Y, Bates DM, Keating DH, Sato TK, Ong IM, Landick R.** 2014. Death by a thousand cuts: the challenges and diverse landscape of lignocellulosic hydrolysate inhibitors. Front Microbiol **5:**90.

14. **Segato F, Damasio AR, de Lucas RC, Squina FM, Prade RA.** 2014. Genomics review

of holocellulose deconstruction by aspergilli. Microbiol Mol Biol Rev 78:588-613.

15. **Socha AM, Parthasarathi R, Shi J, Pattathil S, Whyte D, Bergeron M, George A, Tran K, Stavila V, Venkatachalam S, Hahn MG.** 2014. Efficient biomass pretreatment using ionic liquids derived from lignin and hemicellulose. **111:**E3587-3595.

16. **Songsiriritthigul C, Buranabanyat B, Haltrich D, Yamabhai M.** 2010. Efficient recombinant expression and secretion of a thermostable GH26 mannan endo-1,4-beta-mannosidase from Bacillus licheniformis in Escherichia coli. Microb Cell Fact **9:**20.

17. **Yamada R, Hasunuma T, Kondo A.** 2013. Endowing non-cellulolytic microorganisms with cellulolytic activity aiming for consolidated bioprocessing. Biotechnol Adv **31:**754-763.

18. **Hill J, Nelson E, Tilman D, Polasky S, Tiffany D.** 2006. Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. Proc Natl Acad Sci U S A **103:**11206-11210.

19. **Simmons BA, Singh S, Liggenstoffer AS, Youssef NH, Wilkins MR, Elshahed MS.** 2014. Evaluating the utility of hydrothermolysis pretreatment approaches in enhancing lignocellulosic biomass degradation by the anaerobic fungus Orpinomyces sp. strain C1A. Proc Natl Acad Sci U S A **104:**43-48.

20. **Li XL, Skory CD, Ximenes EA, Jordan DB, Dien BS, Hughes SR, Cotta MA.** 2007. Expression of an AT-rich xylanase gene from the anaerobic fungus Orpinomyces sp. strain PC-2 in and secretion of the heterologous enzyme by *Hypocrea jecorina*. Appl Microbiol Biotechnol **74:**1264-1275.

21. **Wan C, Li Y.** 2012. Fungal pretreatment of lignocellulosic biomass. Biotechnol Adv **30:**1447-1457.

22. **Gilbert HJ, Hazlewood GP, Laurie JI, Orpin CG, Xue GP.** 1992. Homologous catalytic domains in a rumen fungal xylanase: evidence for gene duplication and prokaryotic origin. Mol Microbiol **6:**2065-2072.

23. **Kumar D, Murthy GS.** 2011. Impact of pretreatment and downstream processing technologies on economics and energy in cellulosic ethanol production. Biotechnol Biofuels **4:**27.

24. **Vertes AA, Inui M, Yukawa H.** 2006. Implementing biofuels on a global scale. Nat Biotechnol **24:**761-764.

25. **Zhang W, Bai FW, Zhong JJ.** 2009. Industrial biotechnology: Current status and future development for the sustainability of human society. J Biotechnol **144:**1-2.

26. **Orpin CG.** 1977. Invasion of plant tissue in the rumen by the flagellate *Neocallimastix frontalis*. J Gen Microbiol **98:**423-430.

27. **Qi M, Wang P, Selinger LB, Yanke LJ, Forster RJ, McAllister TA.** 2011. Isolation and characterization of a ferulic acid esterase (Fae1A) from the rumen fungus *Anaeromyces mucronatus*. J Appl Microbiol **110:**1341-1350.

28. **Chen H, Hopper SL, Li XL, Ljungdahl LG, Cerniglia CE.** 2006. 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.

29. **Sanderson K.** 2011. Lignocellulose: A chewy problem. Nature **474:**S12-14.

30. **Cosgrove DJ.** 2000. Loosening of plant cell walls by expansins. Nature **407:**321-326.

31. **Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, Mackie RI, Pennacchio LA, Tringe SG, Visel A, Woyke T, Wang Z, Rubin EM.** 2011. Metagenomic discovery of biomass-degrading genes and genomes

from cow rumen. Science **331:**463-467.

32. **Li XL, Chen H, Ljungdahl LG.** 1997. Monocentric and polycentric anaerobic fungi produce structurally related cellulases and xylanases. Appl Environ Microbiol **63:**628-635.

33. **Krause DO, Denman SE, Mackie RI, Morrison M, Rae AL, Attwood GT, McSweeney CS.** 2003. Opportunities to improve fiber degradation in the rumen: microbiology, ecology, and genomics. FEMS Microbiol Rev **27:**663-693.

34. **Sticklen MB.** 2008. Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. Nat Rev Genet **9:**433-443.

35. **Sticklen M.** 2006. Plant genetic engineering to improve biomass characteristics for biofuels. Curr Opin Biotechnol **17:**315-319.

36. **Alvira P, Tomas-Pejo E, Ballesteros M, Negro MJ.** 2010. Pretreatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: A review. Bioresour Technol **101:**4851-4861.

37. **Galbe M, Sassner P, Wingren A, Zacchi G.** 2007. Process engineering economics of bioethanol production. Adv Biochem Eng Biotechnol **108:**303-327.

38. **Sharma A, Shinde Y, Pareek V, Zhang D.** 2015. Process modelling of biomass conversion to biofuels with combined heat and power. Bioresour Technol **198:**309-315.

39. **Blumenthal HJ, Roseman S.** 1957. Quantitative estimation of chitin in fungi. J Bacteriol **74:**222-224.

40. **Azarpira A, Lu F, Ralph J.** 2011. Reactions of dehydrodiferulates with ammonia. Org Biomol Chem **9:**6779-6787.

41. **Chen Y, Stevens MA, Zhu Y, Holmes J, Moxley G, Xu H.** 2012. Reducing acid in

dilute acid pretreatment and the impact on enzymatic saccharification. J Ind Microbiol Biotechnol **39:**691-700.

42. **Gordon GL, Phillips MW.** 1998. The role of anaerobic gut fungi in ruminants. Nutr Res Rev **11:**133-168.

43. **Veira DM.** 1986. The role of ciliate protozoa in nutrition of the ruminant. J Anim Sci **63:**1547-1560.

44. **Akin DE, Borneman WS.** 1990. Role of rumen fungi in fiber degradation. J Dairy Sci **73:**3023-3032.

45. **Hemme CL, Mouttaki H, Lee YJ, Zhang G, Goodwin L, Lucas S, Copeland A, Lapidus A, Glavina del Rio T, Tice H, Saunders E, Brettin T, Detter JC, Han CS, Pitluck S, Land ML, Hauser LJ, Kyrpides N, Mikhailova N, He Z, Wu L, Van Nostrand JD, Henrissat B, He Q, Lawson PA, Tanner RS, Lynd LR, Wiegel J, Fields MW, Arkin AP, Schadt CW, Stevenson BS, McInerney MJ, Yang Y, Dong H, Xing D, Ren N, Wang A, Huhnke RL, Mielenz JR, Ding SY, Himmel ME, Taghavi S, van der Lelie D, Rubin EM, Zhou J.** 2010. Sequencing of multiple clostridial genomes related to biomass conversion and biofuel production. J Bacteriol **192:**6494-6496.

46. **Bauer A, Lizasoain J, Theuretzbacher F, Agger JW, Rincon M, Menardo S, Saylor MK, Enguidanos R, Nielsen PJ, Potthast A, Zweckmair T, Gronauer A, Horn SJ.** 2014. Steam explosion pretreatment for enhancing biogas production of late harvested hay. Bioresour Technol **166:**403-410.

47. **Orpin CG.** 1976. Studies on the rumen flagellate Sphaeromonas communis. J Gen Microbiol **94:**270-280.

48. **Park JI, Steen EJ, Burd H, Evans SS, Redding-Johnson AM, Batth T, Benke PI, D'Haeseleer P, Sun N, Sale KL, Keasling JD, Lee TS, Petzold CJ, Mukhopadhyay A, Singer SW, Simmons BA, Gladden JM.** 2012. A thermophilic ionic liquid-tolerant cellulase cocktail for the production of cellulosic biofuels. PLoS One **7:**e37010.

49. **Li XL, Chen H, Ljungdahl LG.** 1997. Two cellulases, CelA and CelC, from the polycentric anaerobic fungus Orpinomyces strain PC-2 contain N-terminal docking domains for a cellulase-hemicellulase complex. Appl Environ Microbiol **63:**4721-4728.

50. **Chen H, Li XL, Blum DL, Ljungdahl LG.** 1998. Two genes of the anaerobic fungus Orpinomyces sp. strain PC-2 encoding cellulases with endoglucanase activities may have arisen by gene duplication. FEMS Microbiol Lett **159:**63-68.

51. **Vicari KJ, Tallam SS, Shatova T, Joo KK, Scarlata CJ, Humbird D, Wolfrum EJ, Beckham GT.** 2012. Uncertainty in techno-economic estimates of cellulosic ethanol production due to experimental measurement uncertainty. Biotechnol Biofuels **5:**23.

52. **Chen Y, Stevens MA, Zhu Y, Holmes J, Xu H.** 2013. Understanding of alkaline pretreatment parameters for corn stover enzymatic saccharification. Biotechnol Biofuels **6:**8.

53. **Sun FF, Hong J, Hu J, Saddler JN, Fang X, Zhang Z, Shen S.** 2015. Accessory enzymes influence cellulase hydrolysis of the model substrate and the realistic lignocellulosic biomass. Enzyme Microb Technol **79-80:**42-48.

54. **Hu J, Chandra R, Arantes V, Gourlay K, van Dyk JS, Saddler JN.** 2015. The addition of accessory enzymes enhances the hydrolytic performance of cellulase enzymes at high solid loadings. Bioresour Technol **186:**149-153.

55. **Gilbert HJ.** 2010. The biochemistry and structural biology of plant cell wall deconstruction. Plant Physiol **153:**444-455.

56. **Mabee WE, Saddler JN.** 2010. Bioethanol from lignocellulosics: Status and perspectives in Canada. Bioresour Technol **101:**4806-4813.

57. **Guo M, Li C, Facciotto G, Bergante S, Bhatia R, Comolli R, Ferre C, Murphy R.** 2015. Bioethanol from poplar clone Imola: an environmentally viable alternative to fossil fuel? Biotechnol Biofuels **8:**134.

58. **Cosgrove DJ.** 1999. Enzymes and other agents that enhance cell wall extensibility. Annu Rev Plant Physiol Plant Mol Biol **50:**391-417.

59. **Somerville C, Youngs H, Taylor C, Davis SC, Long SP.** 2010. Feedstocks for lignocellulosic biofuels. Science **329:**790-792.

60. **Cosgrove DJ.** 2005. Growth of the plant cell wall. Nat Rev Mol Cell Biol **6:**850-861.

61. **Dale VH, Kline KL, Wright LL, Perlack RD, Downing M, Graham RL.** 2011. Interactions among bioenergy feedstock choices, landscape dynamics, and land use. Ecol Appl **21:**1039-1054.

62. **Singh A, Pant D, Korres NE, Nizami AS, Prasad S, Murphy JD.** 2010. Key issues in life cycle assessment of ethanol production from lignocellulosic biomass: Challenges and perspectives. Bioresour Technol **101:**5003-5012.

63. **Varner JE, Lin LS.** 1989. Plant cell wall architecture. Cell **56:**231-239.

64. **Hu J, Arantes V, Pribowo A, Saddler JN.** 2013. The synergistic action of accessory enzymes enhances the hydrolytic potential of a "cellulase mixture" but is highly substrate specific. Biotechnol Biofuels **6:**112.

CHAPTER II

GENOME OF THE ANAEROBIC FUNGUS ORPINOMYCES SP. C1A REVEALS THE UNIQUE EVOLUTIONARY HISTORY OF A REMARKABLE PLANT BIOMASS DEGRADER

Abstract

Anaerobic gut fungi represent a distinct early-branching fungal phylum (Neocallimastigomycota), and reside in the rumen, hindgut, and feces of ruminant and non-ruminant herbivores. The genome of an anaerobic fungal isolate, *Orpinomyces* sp. strain C1A, was sequenced using a combination of Illumina and PacBio SMRT technologies. The large genome (100.95 Mb, 16,347 genes) displayed extremely low G+C content (17.0%), large non-coding intergenic regions (73.1%), a proliferation of microsatellite repeats (4.9%), and multiple gene duplications. Comparative genomic analysis identified multiple genes and pathways that are absent in Dikarya genomes but present in early-branching fungal lineages and/or non-fungal Opisthokonts. These included genes for post-translational fucosylation, the production of specific intramembrane proteases and extracellular protease inhibitors, the formation of a complete axoneme and intraflagellar trafficking machinery, and a near-complete focal adhesion machinery. Analysis of the lignocellulolytic machinery in the C1A genome revealed an extremely rich repertoire, with evidence of horizontal gene acquisition from multiple bacterial lineages. Experimental analysis indicated that strain C1A is a remarkable biomass degrader, capable of simultaneous saccharification and fermentation of the cellulosic and hemicellulosic fractions in multiple untreated grasses and crop residues examined, with the process significantly enhanced by mild pretreatments. This capability, acquired during its separate evolutionary trajectory in the rumen, along with its resilience and invasiveness when compared to prokaryotic anaerobes, render anaerobic fungi promising agents for consolidated bioprocessing schemes in biofuels production.

Introduction

Members of the anaerobic gut fungi were originally discovered in sheep (1), but have subsequently been observed in the rumen, hindgut, and feces of ruminant and non-ruminant herbivorous mammals and reptilian herbivores. The observation of flagellated zoospores of anaerobic fungi was reported as early as 1910 (2). However, the accidental discovery and subsequent proof that these flagellated zoospores were actually spores of a new fungal lineage rather than ciliated protozoa came relatively late (1). Anaerobic gut fungi belong to the phylum Neocallimastigomycota, an early-branching fungal lineage, for which no current genome analysis has yet been reported. With the exception of the Microsporidiae, few genomes belonging to non-Dikarya fungal lineages have been sequenced and analyzed (3, 4). Therefore, analysis of a Neocallimastigomycota genome and comparative genomic analysis to early-branching and Dikarya fungal genomes could identify salient characteristics associated with fungal evolution and diversification.

In addition to their distinct phylogenetic position, anaerobic fungi appear to be habitat-restricted, and are the only known fungal group that lives within the rumen and gut of herbivores (5). This evolutionary trajectory in a distinct habitat resulted in multiple metabolic and structural adaptations. For example, members of the Neocallimastigomycota have adapted a strict anaerobic lifestyle. Similar to other anaerobic eukaryotes (e.g. *Trichomonas vaginalis*, (6, 7)), their mitochondria have undergone a reductive evolution process to a hydrogenosome, an organelle whose main function is ATP production via substrate level phosphorylation and hydrogen production (6, 8, 9). Anaerobic fungi also reproduce asexually via the release of motile flagellated zoospores from zoosporangia that develop during rhizoidal fungal growth (10, 11). Finally, anaerobic fungi are highly fibrolytic microorganisms, producing a wide array of

cellulolytic, hemicellulolytic, glycolytic, and proteolytic enzymes (5, 12-15). It has been established that in anaerobic gut habitats, these organisms play a role akin to their aerobic counterparts in soils and streams. By attaching themselves to plant materials, they colonize and excrete extracellular enzymes that mobilize the structural plant polymers to be available to other microbes.

Therefore, analysis of Neocallimastigomycota genomes could not only lead to better understanding of the genomic features and metabolic characteristics of an early-branching fungal lineage, but also lead to the identification of metabolic, physiological, and genome-wide adaptations that enabled the survival and establishment of Neocallimastigomycota as core members of the highly eutrophic, prokaryotes-dominated herbivorous rumen and gut. Here we report on the sequencing and analysis of the draft genome and transcriptome of the anaerobic fungal isolate *Orpinomyces* sp. strain C1A (henceforth C1A). We identified multiple unique features within the genome, and reason that these genomic features are a reflection of two important factors: its placement within a phylogenetically distinct early-branching phylum in the Mycota, and its adaptation to the animal rumen gut during its separate evolutionary trajectory from the Mycota. We further demonstrate that one of these evolutionary adaptations, the presence of remarkably efficient lignocellulolytic machinery coupled to anaerobic fermentative metabolism of hexose and pentose monomers, renders this microorganism an extremely promising agent for lignocellulolytic conversion in consolidated biological processing (CBP) schemes for biofuels production.

Materials and Methods:

Culturing, DNA Sequencing, and Genome Assembly.

Culturing. Strain C1A was isolated from the feces of an Angus steer on a cellobiose-switchgrass medium using previously described protocols (16). *Orpinomyces* strain C1A was grown in an anaerobic, rumen fluid-free basal medium that was reduced by cysteine-sulfide and dispensed under a stream of 100% CO_2 as previously described (17). Cellobiose (3.75 g/L) was used as the substrate. C1A cultures were scaled up for nucleic acids extraction in 1-liter batches prepared in 2-liter Schott bottles equipped with the stoppered top of a Balch tube to maintain strict anaerobic conditions during fungal growth. Culturing was conducted using the techniques described by Bryant and modified by Balch and Wolfe (18, 19). After autoclaving, the Schott bottles were cooled to room temperature and the gas phase was replaced by vacuuming and re-pressurization with 100% CO_2 (19). The medium was then amended with penicillin, streptomycin, and chloramphenicol from an anaerobic stock solution in order to provide final concentrations of 50 µg/ml, 20 µg/ml, and 50 µg/ml of each antibiotic, respectively. The medium was then prewarmed at 39°C for approximately 3-4 hours and inoculated with 50 ml of an actively growing culture of *Orpinomyces* strain C1A. The cultures were incubated at 39°C for approximately 3-4 days and the fungal cells were harvested during late log phase by centrifugation at 10,000 rpm for 30 minutes.

DNA extraction and sequencing.

Illumina sequencing. High molecular weight genomic DNA was extracted using a modified CTAB method for isolation of nucleic acids in anaerobic fungi (20). Four micrograms of high molecular weight DNA was used to generate libraries for Illumina Sequence by Synthesis (Illumina-SBS) genome sequencing (21) using the standard Illumina TruSeq DNA protocol

(http://genome.med.harvard.edu/documents/illumina/TruSeq_DNA_SamplePrep_Guide_15005180_A.pdf). Post adaptor ligation size-selected fragments used for flow-cell cluster generation had a mean size of 293 bp as reported by the Agilent 3200 Bioanalyzer. Illumina sequencing was conducted using the services of a commercial provider (Ambrys Genetics, Aliso Viejo CA, USA) on a HiSeq 2000 sequencing platform using 100bp paired-end chemistry. Illumina sequencing yielded 29.2 GB in 146,385,792 quality-filtered paired-end reads (106).

PacBio sequencing. DNA used for Pacific Biosciences SMRT sequencing (22) was isolated using EpiCentere Plant DNA extraction kit (Epicentre Corp., Madison, WI, USA) according to the manufacturer's specification. Selected inserts of 5-10 Kb read-size were prepared from 10 µg of extracted high molecular weight DNA by ligation to the SMRTbell sequencing adapter. SMRT sequencing was conducted using the services of a commercial provider (Expression Analysis, Durham, NC, USA) on a PacBio RS sequencing platform using the second generation C2 sequencing chemistry with eight Zero Mode Wavelength (ZMW) SMRT cells. PacBio sequencing yielded 984.8 MB of quality-filtered data in 463,832 raw long reads (106), with an average read length of 2,124 bp. An additional 26.9 MB of DNA sequence data were harbored in 16,949 reads (average read length of 1586bp) that reached a circular consensus (CCS) during sequencing, and were used for long read error correction.

Genome assembly. All computational assemblies were conducted using the SGI UV 1000 cache coherent Non Unified Memory Architecture (cc-NUMA) high performance computing system Blacklight. Blacklight is an Extreme Science and Engineering Discovery Environment (XSEDE) community-shared computational resource dedicated for high memory footprint jobs such as de novo assembly. Blacklight is housed at the Pittsburg Super Computer Center (http://psc.edu/).

We initially attempted to utilize Illumina paired-end sequencing as the sole mechanism for C1A genome sequencing. Illumina quality filtered reads were assembled with Velvet 1.1.07 (23), using a kmer value of 63 and a minimum coverage cutoff of 7. The resulting assembly was highly fragmented, with an extremely large number of contigs in the final assembly (82,325 contigs), a large proportion of the final assembly (32.4%) harbored in extremely short contigs (300-900 bp), and a low N50 (1666 bp).

Therefore, we sought to improve the assembly by using a hybrid SRMT-Illumina strategy that leverages short read high accuracy data formed from Illumina sequencing to correct errors encountered in long reads produced by SMRT sequencing (24). This hybrid approach has two steps: 1. SMRT read correction, where insertion/deletion errors present in SMRT read outputs are removed to produce corrected reads with sufficient accuracy and quality scores, and 2. De novo assembly of the corrected SMRT reads, either independently or in conjugation with Illumina reads.

SMRT reads were corrected using the PacBioToCA package (http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=PacBioToCA) in wgs 7.0 (25, 26). The high fidelity read set used for the correction was produced by using a combination of Circular Consensus SMRT reads, Illumina paired-end reads with sufficient overlap to merge into a single extended accurate read using Fast Length Adjustment of Short Read (FLASH) (27), and Illumina paired reads without sufficient overlap for extension. Error correction resulted in a total of 570.1 Mbp, in 394,300 long corrected SMRT (C-SMRT) reads with an average phred quality score of 58.5 (28). These C-SMRT reads, which had an N50 of 1686bp, and ranged between 500bp to 10,932bp in length, were subsequently used for de novo assembly. All sequence data included in the final assembly had an average quality score of 59.7. The final

assembly was a marked improvement when compared to the Illumina only assembly, as evident from the improved N50/N90 values, the increase in the number of genes with PASA transcript alignment (see below), and the increase in the average length of gene models. More importantly, the long C-SMRT reads allowed for the identification of a large number of introns previously undetected using Illumina assembly, with the extremely low GC content (8.1%) and the large number of microsatellites within these introns which probably hindered their detection and assembly from short Illumina paired-end reads.

We used the core eukaryotic genes (CEGMA) to test the completion of the final assembly (29). Due to the unique nature of C1A e.g. absence of gluconeogenesis, anaerobic fermentative mode of metabolism, 22/454 genes are not expected to be present, we identified 408 out of 432 genes within the final assembly, suggesting a sequence completion of ≈ 94.4%.

RNA sequencing and gene calling.

RNA sequencing, assembly, and quantitative analysis. RNA for RNA-seq analysis (30) was isolated from a log-phase strain C1A subculture grown and propagated on rumen fluid-free basal medium with cellobiose or cellulose using Masterpure® Yeast RNA Purification kit (Epicentre Corp., Madison, WI, USA). RNA sequencing libraries were generated using the Illumina TruSeq RNA sample protocol. Illumina sequencing was conducted using the services of commercial providers (Ambrys Genetics, Aliso Viejo CA, USA using a HiSeq 2000 sequencing platform using 100bp-paired end chemistry for cellobiose treatment, and Centrillion biosciences Palo Alto, CA using Illumina Miseq Platform for cellulose treatment). Transcripts generated on cellobiose were used in gene calling efforts as described below. All quality-filtered reads were assembled into transcript candidates using the de novo transcriptome assembly program Trinity (31). The Trinity de novo assembly was executed using the standard non-strand specific library

settings with the addition of the "--jaccard_clip" option which minimizes the formation of fusion transcripts, which are often present in fungal genomes by checking for logical spatial orientation of paired reads on the assembled transcript. Transcripts with a base pair length greater than 300bp were considered valid candidates for downstream analysis. Expression levels for the transcriptome assembly were calculated by mapping all pair-ended RNA-seq reads with bowtie (31) using the Trinity de novo assembly as the reference index. Quantitative expression estimation of alignment values were calculated in transcripts per million values with the RNA-Seq by Expectation-Maximization (RSEM) package (32). The cellobiose C1A transcriptome was utilized for gene calling through transcript alignment. However, while genomic data was used for genomic analysis in this study, in rare cases within pathway analysis (Results section below), transcriptome data were consulted in case a single/few gene(s) within a specific pathway was inexplicably missing from the genome.

Gene calling. Contigs larger than 1000 base pairs produced by the assembly were used as input contigs for gene model generation and downstream analysis. Gene calling was conducted using a combination of ab initio gene models prediction using GlimmerHMM (33) and Augustus (34), and transcript alignments from cellobiose treatment using PASA (Program to Assemble Spliced Assemblies) (35). Training parameters for the ab initio programs were generated using the de novo assembled transcripts aligned to the genome assembly using GMAP (36). Additional gene hint parameters available for the Augustus program using the unassembled RNA-seq read data were generated using the software's recommended protocol (http://bioinf.uni-greifswald.de/augustus/binaries/readme.rnaseq.html).

AB initio gene-calling algorithms produced 60,595 gene models, which were combined with 14,009 PASA high quality transcript assemblies and 38,647 Trinity transcripts to genome GMAP

alignments. The final single consensus gene model for each theoretical locus was produced by information-based source-weighted integration using EvidenceModeler (EVM) (35). A final 16,347 consolidated consensus gene models were generated by EVM.

To examine the impact of assembly fragmentation on the overall number of genes identified in the C1A genome, we extracted 300 bp from both ends of every contig in the C1A final genome assembly. To identify sequence redundancy, we queried these ends against NT and NR GenBank databases using BlastN and BlastX respectively. The results show a unique first hit for each contig end, suggesting that the large number of contigs in the final assembly did not result in any false duplication of predicted genes. A similar result was obtained when performing the same analysis against the COG database.

Annotation.

Annotation strategy overview. Annotation of gene models and gene transcripts was achieved using a combination of command line bioinformatics programs, manual curation, and automated online annotation suites. Closest relative homologs were assigned to each consensus gene and de novo transcript model using the BLAST+ (37) module Blastp for the gene models and Blastx for putative transcripts against the nr database. Identified homologs with an e-value of e^{-4} or less were considered sufficient for evaluation of functional activity assignment. PFAM domains were identified using the hmmscan module of the HMMER suite (38) against the PFAM 26.0 database of conserved protein families (39). All domains identified by hmmscan having a full sequence e-value of e^{-4} or less were assigned to the gene models for functional annotation. The integrated microbial genomes system (IMG) was used for automated gene calling and annotation of both the genome gene models and the transcripts. The resulting Blast and IMG results were manually curated and used for analysis of various cellular processes within the genome, as well as for

confirmation of the presence/ absence of key metabolic genes using reciprocal Blastp against the genome. All gene calling and annotation computational work was conducted using the Oklahoma State University high performance computing clusters Pistol Pete and Spur.

Carbohydrate Active Enzymes (CAZymes) identification and analysis. Identification of CAZy genes in the C1A genome, as well as in multiple genomes that were used for comparative analysis, was achieved via PFAM domain identification and analysis as described previously (40). The Carbohydrate Active Enzyme database classification system (41) was used to classify glycoside hydrolase, carboxyl esterase (CE), pectate lyase (PL), and carbohydrate-binding modules (CBM). All models harboring a dockerin domain with e^{-4} significance value or less were considered eligible for further analysis and classification. Potential secreted peptides and transmembrane proteins candidates were identified using SignalP 4.0 (42) and TMHMM (43).

Additionally, more stringent controls were conducted to guard against any possibly *in silico* gene number inflation within the CAZyme dataset due to the fragmented assembly. To guard against *in silico* inflation of gene numbers due to improper assembly, we identified genes with near identical (>97%) amino acid identity. Each cluster of genes with near identical amino acid sequence was aligned; and genes with less than 80% alignment to the parent model were identified as possible artifacts and removed from the assembly. To guard against possible *in silico* inflation of gene numbers due to gene fragmentation between two contigs, we manually examined each CAZyme gene model to identify genes with incomplete pfam CAZy domain. Genes with incomplete pfam domains were removed from the assembly. Collectively, these additional quality control approaches resulted in the removal of 76 GH, 5 CE, and 5 PL genes.

Repeats identification. Repetitive DNA sequences (DNA repeats) are defined as sequences present in more than a defined number of copies and that have no apparent biological function

(44, 45). DNA repeats can be classified into simple sequence repeats (SSRs), and complex repeats. SSRs (also known as tandem repeats) are classified, depending on the length of the repeated unit, into microsatellites (basic unit length ranging from 1-6 bp) (44, 45), and minisatellites (basic unit length ranging from 15->150 bp repeated 2-100 times) (45). Longer tandem repeats constitute satellites (centromeres) and telomeric repeats. Complex repeats, on the other hand, result from transposable elements (TE). TEs are further classified into class I retrotransposons (including Long terminal repeats (LTR), and non-LTR), and class II DNA transposons.

Microsatellite (SSRs) in the C1A genome were identified using PHOBOS (46) with the --minPerfection 100 flag to detect only perfect repeats. SSRs identified have the following minimum number of repeats: Mononucleotide with at least 10 repeats; dinucleotides with at least 6 repeats; tri-, tetra-, penta- and hexanucleotides with at least 5 repeats.

Complex repeats were identified in C1A genome as previously described (47) using a combination of RepeatScout (48), RepeatMasker (49), LTR_FINDER(50), and BLASTx against RepBase (http://www.girinst.org/repbase/index.html). Briefly, a consensus repeat library was created using the default parameters of RepeatScout. This library was filtered by removing short sequences (<100 bp), and those repeats with significant hits to Uniprot proteins (except repeats with significant hits to transposable elements). The filtered consensus library was then compared to RepBase database using BLASTx for manual annotation and classification. LTRs were identified in the genome using LTR finder. Similar to other TEs, candidate LTRs were compared to the RepBase for classification. Finally, all candidate repeats classified by RepBase were used to mask the genome in RepeatMasker to identify the number of occurrences and the percentage genome coverage of each TE class.

Gene duplication was identified by running local Blastp using C1A proteins as both the subject and the query. Only the first and second hits were examined. The second hit is a protein with similarity to the query protein present anywhere in the genome. Percent similarities cutoff of 40% or more were used.

Non-coding RNA identification. Ribosomal RNAs were identified using local Blastn search with sequences corresponding to published 5.8S (5.8S fragment in Genbank accession number AJ864475.1), 18S (GenBank Accession number AY546684) *Spizellomyces punctatus*.1), 28S (28S fragment in Genbank accession number AJ864475.1), ITS1 (Genbank accession number AF170191.1), ITS2 (Genbank accession number JN943062.1) as the database. Transfer RNAs were identified using tRNAscan-1.4 (51).

Identification of proteases, protease inhibitors, and transporters. All proteins were compared to the MEROPS database using Blastp to identify potential proteases and protease inhibitors. Membrane transporters in the C1A genome were identified by Blastp comparison against the transporter classification database (TCDB) sequences available at (http://www.tcdb.org/seqfile/tcdb) using the GBLAST2 program (http://www.tcdb.org/labsoftware.php).

Identification of hydrogenosomal proteins. We bioinformatically predicted proteins potentially imported to the hydrogenosomal matrix in strain C1A using a combination of motif search and Mitoprot v1.0 (52). First, C1A proteins were examined for the presence of an N-terminal mitochondrial targeting sequence corresponding to previously predicted motif similar to ML(S|T|A|C|G|R){0,1}X{0,19}RXF(I|L|F|S|A|G|Q), ML(S|T|A|C|G|R){0,1}X{0,19}R(F|N|E|S|G) (I|L|F|S|A|G|Q), MTLX{0,19}RXF(I|L|F|S|A|G|Q), MTLX{0,19}R(F|N|E|S|G) (I|L|F|S|A|G|Q), MSLX{0,19}RXF(I|L|F|S|A|G|Q), or

MSLX{0,19}R(F|N|E|S|G) (I|L|F|S|A|G|Q), where X is any amino acid except tryptophan. Numbers between braces refer to the previous residue repeat number and parenthesis means that any of the residues enclosed is possible at that position (7). Mitochondrial import probabilities of proteins harboring this N-terminal motif were then predicted using Mitoprot v1.0 (52), where an arbitrary probability of 0.6 was used as the cutoff. Using these criteria, we identified 21 potential intra-hydrogenosomal proteins. Further, Mitoprot was also used to predict the mitochondrial import probabilities of proteins with similarity to known mitochondrial matrix proteins that did not have the above mitochondrial-targeting motif. An additional 25 potential hydrogenosomal proteins were identified using these criteria.

Comparative analysis of the C1A genome to basal fungi, Dikarya, and Opisthokonta genomes.

We used local Blastp comparison of C1A proteins against all Mycota proteins, as well as against Dikarya proteins. A Mycota Blastp database was created by downloading proteins of all sequenced fungal genomes available from Genbank and IMG (total of 116 fungal genomes). Of those, 4 belonged to early-branching lineages (*Allomyces macrogynus*, *Batrachochytrium dendrobatidis*, *Spizellomyces punctatus*, and *Rhizopus oryzae*). The remaining fungal genomes constitute the Dikarya Blastp database. The number and identity of Blastp first hits of C1A proteins against both databases at different E-value cutoffs (e^{-5}, e^{-10}, e^{-15}, e^{-20}, e^{-25}, e^{-30}, e^{-35}) were then used to specify C1A proteins that are; general fungal proteins (present in both early-branching and Dikarya fungi), early-branching fungi-specific proteins (present only in early-branching fungi but not Dikarya fungi), and C1A-specific proteins (specific to C1A and absent from other fungal genomes). The later (C1A-specific proteins) were compared to the nr database excluding Mycota using the flag –negative_gilist. Proteins with no hits in the nr database were

considered C1A hypothetical proteins. Functional annotation of various C1A-specific and early-branching fungi-specific proteins identified were conducted using PANTHER classification outline (53).

Lignocellulolytic capabilities of strain C1A

Plant materials and pretreatment. Samples of mature Kanlow switchgrass (*Panicum virgatum var Kanlow*), mature *Sorghum bicolor*, and mature energy cane (*Saccharum officianarum* var Ho02) were obtained from Oklahoma State University experimental plots in Stillwater, OK. Dried alfalfa was obtained from a local farm and ranch supplier. Samples of Bermuda grass (*Cynodon dactylon*) were obtained from residential lawn clippings in Guthrie, OK. Samples of corn stover from *Zea mays* were obtained from the Industrial Agricultural Products Center at the University of Nebraska in Lincoln. Untreated wood samples, including cedar (*Juniperus* sp.), oak (*Quercus* sp.), and pine (*Pinus* sp.) were obtained from a local lumberyard in Stillwater, OK. Cottonwood (*Populus deltoides)* and willow (*Salix babylonica)* wood samples were harvested from live trees growing in the Stillwater area. All samples were dried at 45°C overnight, milled, and sieved to a final particle size of 2 mm as previously described (54).

Sodium hydroxide (NaOH) treatments were conducted by heating 4g of dried plant material in 40 ml of a 1% NaOH solution inside a sealed serum bottle at 50^0C for 12 hours (55). Acid treatment was conducted by heating 4g of dried plant material in 40 ml of 0.5% H_2SO_4 inside a sealed serum bottle for 1 hour (56). Hydrothermolysis-treated switchgrass was prepared by mixing 60g of switchgrass with distilled water to achieve a 10% dry matter mixture (54). This mixture was placed inside 1L benchtop pressure reactor (Parr Series 4520, Parr Instrument Company, Moline, IL, USA) that was heated to 200°C and agitated at 500 rpm (54). The

switchgrass/water mixture was held at 200°C for 10 minutes and then cooled in an ice bath (54). All of the treated switchgrass samples were recovered from pretreatment incubations by filtration. The sodium hydroxide and acid treated switchgrass were washed with deionized water as previously described (55, 56). All of the pretreated switchgrass samples were dried at 45°C for approximately 48 hours before they were used in the experiments described below.

Growth of strain C1A on plant materials. Experiments to evaluate the growth of strain C1A on different treated and pretreated plant materials were conducted under strict anaerobic conditions in 160-ml serum bottles. All experiments were conducted in triplicate, and unless otherwise specified, 0.5g of plant material was used as the substrate. Experiments were conducted in a previously described rumen fluid-free basal medium (17). The medium was prepared under strict anaerobic conditions using 100% CO_2 and the techniques of Bryant (18), as modified by Balch and Wolfe (19). Once the basal medium was prepared it was autoclaved for 20 minutes at 121°C and 15 psi of pressure and then cooled. Each serum bottle was then amended with the appropriate type of plant biomass inside an anaerobic chamber (Coy Laboratory Products Grass Lake, MI). After the serum bottles were amended with plant materials they were removed from the glove bag and the headspace was re-pressurized with 15 psi of 100% CO_2 (19). Five milliliters of an actively-growing culture of strain C1A (approximately 2.6 mg of fungal biomass) was used as an inoculum and added to 45ml media in 160 ml serum bottles. In all experiments, serum bottles were incubated at 39^{O}C in a non-shaking incubator. Substrate-unamended controls were included in all experiments to account for any product carryover from the inoculum. Triplicate bottles were sacrificed at different time intervals to quantify substrate loss and product formation.

Analytical methods. Fatty acids and ethanol in supernatant fractions were quantified using an HPLC with a refractive index detector (1100 Series, Agilent, Santa Clara, CA, USA) and an Aminex HPX-87H column (Biorad, Sunnyvale, CA, USA), which was heated to 60^0C. The mobile phase was 0.01 N H_2SO_4, with a flow rate of 0.6 ml per minute. Sugars in supernatant fractions were also quantified using an HPLC with a refractive index detector (1100 Series, Agilent, Santa Clara, CA, USA). The HPLC was equipped with an Aminex HPX-87P column (Biorad, Sunnyvale, CA, USA), which was heated to 85°C. Distilled water was used as the mobile phase at a flow rate of 0.6 ml per minute.

The amount of plant material consumed in serum bottles was calculated by subtracting the time final from the time 0 dry weights of each plant material. Since the time final pellets contained a mixture of plant and fungal biomass, the amount of fungal biomass at time final was indirectly quantified using formate concentrations as previously described (12). The amounts of cellulose, xylan, hemicellulose, and lignin in the different plant substrates were determined using the standard NREL procedures (57). The procedure included the addition of 3mL of 72% sulfuric acid to each sample and incubation at 30±3°C for 1 hour, stirring every 5-10 min. The samples were then diluted with 84 mL of deionized water, capped, and autoclaved for 1 hour to 121°C. The cooled solution was filtered, and this filtrate was used to determine carbohydrate content and soluble lignin. The remaining solids were washed and dried to constant weight at 105°C to determine acid-insoluble residue (AIR) and then ashed at 575°C for 24 hours (57). Analyses of resulting carbohydrates within the filtrate were done by HPLC with refractive index detection (RID) (Agilent 1100 Series, Santa Clara, CA) on an Aminex HPX-87P column at 85°C with a mobile phase of deionized water pumped at 0.6mL/min for 30 min (57). Twenty microliters of each sample were analyzed for cellobiose, glucose, xylose, galactose, arabinose and mannose.

Contributions of structural constituents to the total biomass composition were determined using the NREL summative mass closure procedure (58). The acid-soluble lignin (ASL) content was determined using a UV spectrophotometer set at a wavelength of 205 nm, as has been previously used to determine ASL in switchgrass (59). As recommended in the NREL procedure, ASL in corn stover was measured at 320 nm, whereas a 240 nm wavelength was used for the remaining biomass types (57).

Nucleotide sequence accession numbers. The final genome assembly is available in the IMG genome database with accession number 2518645524 and in Genbank database with accession number PRJNA176180. The final transcriptome is available in the IMG database with accession number 2510461071 and in Genbank database with accession number PRJNA176180.

Results

Isolation and general genomic features. Strain C1A was isolated from the feces of an Angus steer on a cellobiose-switchgrass medium using previously described protocols (16). The isolate displayed polycentric growth and effectively colonized switchgrass. Phylogenetic analysis using the nuclear ribosomal internal transcribed spacer II (ITS-II) region supported the placement of strain C1A as a member of the genus *Orpinomyces*, and phylogenetic analysis using a concatenated set of 42 housekeeping genes supported the basal, early-diverging position of the Neocallimastigomycota (106)

We sequenced the C1A genome using a combination of paired-end short read Illumina technology (≈290X coverage) and Single Molecule Real time (SMRT) Pacific Biosciences technology (≈10X coverage). The transcriptomes of strain C1A grown on cellobiose and cellulose were also sequenced using Illumina technology. The C1A genome displayed several interesting features (Figure 2-1): It had the lowest GC content (17.0%) when compared to available genomes of all free-living microorganisms sequenced to-date (Figure 2.1a). This value is lower than those observed in the notoriously AT-rich *Dictyostelium* spp. and *Plasmodium* sp. within the microeukaryotes, and is only surpassed by a few Proteobacterial obligate endosymbionts e.g. 'Candidatus *Zinderia insecticola*' (13.5%) and 'Candidatus *Carsonella ruddii*' (16.6%). The GC content was higher in protein-coding genes (26.8%), compared to non-coding regions (14.8% in intergenic and 8.1% in introns) but still resulted in a marked codon usage skew (106). The C1A genome was also characterized by a relatively large proportion of non-coding intergenic regions (73.1%) (Figure 2.1b). Further, non-coding regions displayed massive proliferation of simple sequence repeats (SSRs) in the C1A genome. The 249,194 SSRs constituted 4.9% of the entire genome, as well as 5.8% and 6.0% of the intergenic region and the

introns, respectively (Figure 2.1c). These values vastly surpass the number of SSR repeats that were observed in previously analyzed fungal genomes by at least one order of magnitude (47, 60). Homopolymeric A or T mono-repeats represented the majority of observed repeats (68.6% and 60.6% of total SSR numbers and length, respectively), with 3,589 identified cases of ≥ 50 bp stretches of A or T within the assembly (Figure 2.1d).

Comparative gene content with basal and Dikarya fungi. Gene calling resulted in the identification of 16,347 protein-coding genes, a number surpassed by only few fungal genomes This large number could partly be attributed to gene duplication (Figure 2.2a), since 3,252 gene pairs share >90% sequence similarity.

In addition, comparative genomic analysis indicated that only 48.4% of C1A genes has at least one ortholog in all examined Dikarya genomes (n=116), 9.5% of C1A genes has at least one ortholog within examined early-branching fungal (n=4), but not in Dikarya genomes, and that 42.2% (6,886) of C1A genes are unique and have not been previously encountered within the Mycota (Figure 2.2b), These unique C1A genes were either hypothetical proteins (n=5,666), genes with non-fungal, eukaryotic orthologs (n=578), or genes with bacterial orthologs (n=642). Eukaryotic, non-fungal C1A genes were mostly encoding cellular processes e.g. receptors and nucleic acid-binding proteins, highlighting the distinct early-branching fungal position of the Neocallimastigomycota, while C1A genes with prokaryotic orthologs were mainly involved in metabolic processes, e.g. hydrolases, transporters, transferases, and phosphatases, highlighting the potential role of horizontal gene transfer in shaping C1A metabolic capabilities.

Genomic analysis and comparative genomics reveal multiple differences between Neocallimastigomycota and Dikarya. Analysis of genes involved in information processing (replication, transcription, and translation), as well as cytoskeletal structure and intracellular

trafficking mechanisms revealed all salient features associated with such processes in eukaryotic cells. More importantly, comparative genomic analysis identified multiple cellular processes in which either only the C1A genome, or all early-branching fungal genomes, possess features that appear to be absent from Dikarya genomes, but mostly associated with non-fungal Opisthokonts (choanoflagellates such as *Monosiga brevicollis, Capsaspora owczarzaki,* an independent unicellular Opisthokonta lineage, and Metazoa) and higher non-Opisthokonts eukaryotes (Table 2.1). Five different examples are highlighted. 1. One gene in the C1A genome encodes metalloprotease site-2-protease (S2P) family, and seven different genes encode various components of the γ-secretase complex, including aspartyl protease presenelin. Both of these types of intramembrane proteases are represented in Metazoa (mostly Chordata, Nematoda, and Arthropoda) with few representatives in plants, and have no representation in Dikarya (Table 2.1). 2. The genome contains two fucosyltransferase genes that mediate fucosylation, a post-translational modification process that is typically observed in Chordata, Arthropoda, and Viridiplantae, but not in the Dikarya (Table 2.1) 3. The C1A genome possesses a near-complete focal adhesion machinery (Table 2.1). Focal adhesions are large multiprotein intracellular assemblies that mediate cell anchorage and mechanical adhesion to the extracellular matrix. They also act as a signaling milieu where signaling proteins are concentrated at sites of integrin binding and connect the cell's cytoskeleton to the extracellular matrix. FA appears to be absent from filamentous fungi, and more common in other eukaryotes (Amoebozoa, Metazoa (sponges, placozoans, and cnidarians), and Holozoa (61). 4. The C1A genome possesses a complete axoneme and intraflagellar trafficking machinery proteins. Axoneme acts as a scaffold for other protein complexes including motor proteins (e.g. kinesin, and dynein) essential for intraflagellar transport of proteins. Ciliated and flagellated eukaryotic cells are known to possess an axoneme,

as do Neocallimastigomycota, and other early-branching fungi that produce flagellated zoospores. This feature is absent in other Dikarya fungi that produce non-flagellated spores. 5. Finally, the C1A genome encodes various extracellular protease inhibitors, some of which (serpins) have not previously been encountered in the Dikarya, (e.g. serpins of family I4 are present mostly within eukaryotic metazoan phyla (Arthropoda, Chordata, and Nematoda), and have also been identified in Bacteria and Archaea, but have not previously been encountered in Fungi). Several identified serpins have dockerin domains, confirming their cellulosomal destination and their potential role in combating plant proteases, as previously suggested (62).

Hydrogenosomal structure and function. Anaerobic fungi lack mitochondria, but possess a double-membrane hydrogenosome whose main function is ATP production via substrate level phosphorylation and hydrogen production (6, 8, 9). The C1A genome encodes a near-complete hydrogenosomal protein import system with components of the TOM outer membrane transport system (4 out of 7 genes), the SAM sorting and assembly complex (4 out of 5 genes) for protein insertion in the outer membrane, the MIA intermembrane space import and assembly complex (2 out of 3 genes), small TIMs (2 out of 4 genes), the TIM22 complex for protein insertion in the inner membrane (6 out of 6 genes), and the inner membrane transport system and associated motor (TIM23 complex, 10 out of 11 genes) (Figure 2.3). By comparison, the hydrogenosomal import machinery of *Trichomonas vaginalis* has been reduced to few outer membrane proteins (Tom40, Sam50, Hmp35, and Hmp36), few inner membrane proteins (Tim17/22/23, Tim 44, and PAM16, 18), and one highly modified intermembrane small TIMs (63). Further, examining the phylogenetic affiliation of mitochondrial import proteins in strain C1A clearly demonstrates their fungal origin, since their closest relatives are consistently those from fungal mitochondria.

Using two different bioinformatic criteria, we identified 46 intra-hydrogenosomal proteins in the C1A genome. Candidate proteins for import into the hydrogenosomal matrix included several hypothetical proteins, Fe-S cluster assembly and maturation proteins, peptidases, intralumenal chaperones and co-chaperones, as well as pyruvate metabolism and energy production enzymes (106).

Hydrogenosomes are the site of multiple metabolic processes for pyruvate metabolism, ATP production via substrate level phosphorylation, and regeneration of reduced electron carriers, e.g. NADPH and NADH. The C1A genome encodes the genes required for mixed-acid fermentation, the predominant pathway for pyruvate metabolism in anaerobic fungi, as previously suggested (64-66) (Figure 2.3). Genomic reconstruction suggests that pyruvate produced from sugar catabolism in the cytosol could either be metabolized cytosolically, or imported and metabolized in the hydrogenosome. In the cytosol, pyruvate could either be converted to acetyl CoA and formate via cytosolic pyruvate formate lyase (PFL) (a cytosolic PFL-activating enzyme is also encoded by the genome), converted to D-Lactate via cytosolic D-lactate dehydrogenase, or used to produce TCA intermediates required for anaplerotic reactions via an incomplete cytosolic TCA cycle. In the cytosol, acetyl-Co produced could be converted to ethanol via aldehyde dehydrogenase/ alcohol dehydrogenase. In the hydrogenosome, pyruvate could be imported from the cytosol or could be produced from malate via the action of the hydrogenosomal malic enzyme (with the production of CO_2). Hydrogenosomal pyruvate could be then metabolized to acetyl-CoA and formate by a hydrogenosomal PFL. Acetyl CoA produced in the hydrogenosome could subsequently be converted to acetate via the combined action of hydrogenosomal acetate:succinate CoA transferase/ succinyl-CoA synthase to produce ATP via substrate level phosphorylation. The genome encodes a hydrogenosomal acetyl CoA

hydrolase. A similar enzyme in *T. vaginalis* was shown to possess an acetate:succinate CoA transferase activity. The *Orpinomyces* acetyl-CoA hydrolase homologue is most likely performing a similar transferase activity that, coupled to the succinyl-CoA synthase activity, could convert acetyl-CoA to acetate.

In addition, the hydrogenosomal components contain elements for NADH recycling coupled to H_2 production (Fe-only hydrogenase large subunit, NADH dehydrogenase (complex I) subunits E, and F). The 2 subunits of NADH dehydrogenase most probably function to re-oxidize NADH produced in the lumen (e.g. during fatty acid degradation) and transfer electrons to Fe-only hydrogenase. Since PFL mediates pyruvate metabolism without the production of reduced equivalents, H_2 production via the hydrogenase enzyme is thought to be minor and to be only required to cope with the NADH produced from other intra-lumenal reactions, e,g, 3-hydroxyacyl-CoA dehydrogenase of fatty acid metabolism (67). Elements of hydrogenosomal NADPH recycling are also present in the genome. NADPH produced from NADP-dependent reactions, e.g. malic enzyme, could possibly be used by the NADP-requiring fatty acid synthesis reactions, e.g. 3-oxoacyl-[acyl-carrier protein] reductase, or recycled by NADPH:quinone reductase to a quinone, where electrons could then transfer to succinate dehydrogenase to reduce fumarate to succinate. Finally, the genome also encodes subunits α, β, γ, and δ of F_0F_1-type ATP synthase that is thought to pump protons to the cytosol keeping the luminal pH slightly alkaline. ATP synthase is likely functioning in conjunction with an ADP/ATP carrier.

Lignocellulolytic repertoire of strain C1A. Prior research efforts have identified multiple genes involved in plant biomass degradation in several Neocallimastigomycota isolates genome (5, 14, 15, 68-90). To provide an overall view of the plant biomass degradation machinery of an anaerobic rumen fungus, we analyzed the lignocellulolytic machinery in the C1A genome. Such

an analysis revealed an extremely rich repertoire that consisted of 357 glycoside hydrolases (GH) genes, 24 polysaccharide lyases (PL), and 92 carbohydrate esterases (CE) (Figure 2.4a). Principal-coordinate analysis demonstrated the unique position of the GH catalytic machinery, compared to multiple fungal and bacterial genomes of distinct habitats, ecological roles, phylogenetic affiliation, and oxygen preferences (Figure 2.4b). For instance, compared to aerobic fungal biomass degraders of industrial and ecological relevance such as *Trichoderma reesei*, *Postia placenta*, *Aspergillus oryzae*, and *Myceliophthora thermophila*, the C1A genome shows an expansion of cellulolytic families GH6, GH9, GH45, GH48 and hemicellulolytic families GH10, GH11, and GH43, as well as the reduction or absence of families GH7, GH16, GH18, GH28, and GH61.

Detailed phylogenetic analysis (106) suggests that the GH machinery in strain C1A has evolved from an ancestor with relatively limited cellulolytic capability to a robust cellulolytic and hemicellulolytic organism through the acquisition of genes from multiple bacterial lineages, many of which are known to be prevalent in the bovine rumen. Overall, 247 (69.2%) of GH genes were most closely related to bacterial orthologs, and 141 (39.5%) of GH genes were most closely related to bacterial orthologs from lineages that are prevalent in the bovine rumen. Such lineages include families Lachnospiraceae, Clostridiaceae, Eubacteriaceae, and Ruminococcaceae within the order Clostridiales, family Streptococcaceae within the order Bacillales, family Prevotellaceae within the order Bacteroidetes, as well as the phylum Fibrobacteres. Cellulose degradation machinery in strain C1A consists of GH5, GH8, GH9, and GH45 endoglucanases and GH6, and GH48 cellobiohydrolases. GH8 is an exclusively prokaryotic gene family (41), and phylogenetic analysis of GH5, and GH9 endoglucanases indicate their close affiliation with endoglucanases from multiple bacterial sources, including the

ruminal genera *Clostridium, Ruminococcus, and Eubacterium*. On the other hand, strain C1A also possesses the distinctively eukaryotic fungal-affiliated GH45 endoglucanases that have rarely been observed in bacterial genomes. Similar to endoglucanases, a dual prokaryotic/eukaryotic origin of strain C1A cellobiohydrolases was observed. Strain C1A possesses multiple GH48 reducing-end cellobiohydrolases, hallmarks of cellulosomal cellobiohydrolases, as well as multiple GH6 non-reducing end cellobiohydrolases that are common in fungal genomes, but rarely observed in anaerobic cellulolytic bacteria.

Unlike cellulose metabolism, hemicellulose degradation machinery in strain C1A appears to be entirely of prokaryotic origin. The C1A genome contains all genes required for the degradation of xylans (glucuronoarabinoxylans, and arabinoxylans), mannans (galactoglucomannans, and glucomannans), and mixed β-(1-3, 1-4) glucans. Strain C1A appears to be highly adapted to the degradation of xylans, the prevalent hemicelluloses in grasses (order Poales) (91). This is evident by the identification of 109 different xylanases, xylosidases, arabinofuranosidases, and α-glucuronosidases belonging to families GH10, GH11, GH39, GH43, and GH67, in addition to multiple glucuronoarabinoxylan-, and arabinoxylan-debranching enzymes (acetylxylan esterases, ferulic acid esterases, and polysaccharide deacetylases). Phylogenetic analysis of GH10 xylanases suggests their close affiliation with multiple bacterial lineages, including the ruminal genera *Butyrivibrio, Clostridium,* and *Eubacterium*. Phylogenetic analysis of GH11 xylanases suggests that they have been solely acquired from *Fibrobacter* species, important constituents of rumen microbiota. A similar bacterial origin was also observed for GH39 and GH43 xylosidase/arabinofuranosidases as well as GH67 α-glucuronosidases, with potential bacterial donors being members of the genera *Clostridium, Ruminococcus, Butyrivibrio, Cellulosilyticum, Eubacterium,* and *Provotella*. Finally, GH26 mannosidases, and

GH16 β-(1,3-1,4)-glucanase, mediating the breakdown of mannans and mixed glucans also had similar bacterial origin, with several sequences affiliated with ruminal lineages e.g. *Acetovibrio, Fibrobacter*, and *Streptococcus*.

Anaerobic fungi produce cellulosomes: extracellular structures that harbor multiple extracellular enzymes bound to scaffoldins (2). Cellulosomal-bound genes in anaerobic fungi usually harbor a fungal dockerin domain (FDD) that is similar in structure to carbohydrate binding module family 10 (CBM10) (5). We identified a total of 220 genes with FDD; 108 of which contained dual glycoside hydrolase-fungal dockerin domains (GH-FDD). GH-FDD genes identified suggest that cellulosomal-bound enzymes play a role in the degradation of cellulose and hemicellulose; but not chitin, starch, or pectin. Within the remaining FDD-containing genes, we identified multiple putative activities that could either aid in biomass degradation (e.g. polysaccharide deacetylases, tannase, lipases, swollenin and expansin module proteins), or act as cellulosomal preservation and defense mechanisms (e.g. protease inhibitors (serpins)), as well as multiple conserved hypothetical and hypothetical proteins.

Carbohydrate-binding modules (CBM) are non-catalytic domains that are often encountered in lignocellulolytic enzymes and promote the association of the enzyme with the substrate. A total of 103 genes harboring carbohydrate-binding modules (CBM) domains belonging to 6 different CBM families were identified. The majority (75.7%) of CBMs were members of the exclusively fungal CBM1 domain. Twenty-six genes with CBM domains were associated with GH enzymes, and 7 were associated with PL enzymes (106). Within GH-CBM dual domain genes, CBM1 domains were associated with several GH10 and GH11 xylanases, CBM18 with GH18 chitinases, and CBM48 with GH13 amylases. No CBM domains were identified in GH genes putatively involved in cellulose metabolism in the C1A genome. No

CBM2 or CBM3 domains, the prevalent CBM modules in bacterial plant biomass-degradation genes and in rumen anaerobic cellulosomal bacteria, respectively, were identified in the C1A genome (106).

Comparative transcriptomic analysis of strain C1A was conducted on cellobiose-grown versus microcrystalline cellulose-grown cultures (Figure 2.5). A total of 172 GH genes were expressed under both conditions, while 39 and 4 GH genes were identified only in cellobiose-grown and cellulose-grown cultures, respectively. In cellulose-grown cultures, transcripts belonging to GH5 cellulases, as well as GH9 and GH48 cellobiohydrolases were drastically upregulated compared to cellobiose-grown cultures. GH8 and GH45 cellulases were only slightly upregulated and their overall transcriptional levels were relatively low (Figure 2.5a, b). GH1 and GH3 β-glucosidases, essential for substrate degradation under both conditions, were either not significantly changed or only slightly upregulated in cellulose-grown cultures (Figure 2.5a, b).

Analysis of expression profiles of all GH genes identified under both conditions revealed that while several cellulase and cellobiohydrolase genes were clearly upregulated in cellulose-grown cultures, the majority of such genes were not significantly (>2-fold) affected by the growth condition (Figure 2.5c), and few were even significantly downregulated. Interestingly, few of the genes upregulated in cellulose-grown cultures belong to GH families associated with the degradation of plant polymers other than cellulose e.g. GH10 and GH11 xylanases, GH18 chitinases, and GH26 mannosidases.

Strain C1A is an effective, versatile biomass degrader. Strain C1A grew readily on untreated, as well as mild acid-, mild alkali-, and hydrothermolysis-treated switchgrass, with the concurrent utilization of cellulose and hemicellulose fractions, but not lignin (Figure 2.6). Dry weight losses

of substrate ranged between 18.6% (28.7% of non lignin fraction) in untreated switchgrass to 40.8% (53.9% of non lignin fraction) in NaOH-treated switchgrass. Further, adjustments to the inoculum/substrate ratios resulted in an increase in the amount of switchgrass metabolized up to 42.8% and 58.4% of the dry weight of untreated and NaOH-treated switchgrass, respectively. Strain C1A performed extremely well on NaOH-treated switchgrass, since this method of pretreatment retains the majority of the hemicellulose content (92, 93), which is degradable by strain C1A. Strain C1A also grew well on hydrothermolysis-treated switchgrass, presumably due to the fact that the removal of hemicellulose resulted in a greater accessibility to cellulose fibers. End product analysis indicated that lactate, acetate, and formate are the main end product of plant biomass degradation. Only minor amounts of ethanol were produced, ranging between 0.045-0.096 mg ethanol/mg biomass.

In addition to switchgrass, we tested the capability of strain C1A to utilize several other types of energy crops (e.g. alfalfa, sorghum, energy cane), agricultural residues (e.g. corn stover), and grasses (e.g. Bermuda grass). We chose these specific plant materials due to the variations in the percentages of cellulose, hemicellulose, and lignin in these plants. The results (Figure 6e) demonstrate the versatility of strain C1A, since it was able to metabolize all different types of examined plant biomass. Within both untreated and NaOH-treated experiments, strain C1A was most effective in metabolism of corn stover, with 40.6% and 62.3% dry weight loss, 51.0% and 75.8% loss in cellulose fraction, and 43.0% and 74.3% loss in hemicellulose fractions in untreated and NaOH-treated corn stover, respectively.

Discussion

Analysis of the C1A genome revealed thoroughly eukaryotic information processing, cytoskeletal structure, and intracellular trafficking machineries. On the other hand, we identified multiple cellular processes in which the C1A genome possesses features that appear to be absent from Dikarya genomes, but mostly associated with early-branching fungi and non-fungal Opisthokonts (Table 2.1). These observations suggest that such features have evolved prior to fungal separation from an Opisthokonta ancestor and were subsequently lost during the evolution of Dikarya, but were retained in the Neocallimastigomycota. The rationale behind the retention of some of these features in the Neocallimastigomycota could be attributed to their unique habitat and evolutionary trajectory. For example, the possession of protease inhibitors to guard against plant, ciliate, and bacterial proteases is extremely beneficial in the rumen habitat. The possession of an axoneme and an intraflagellar-trafficking machinery is required for the motility of flagellated zoospores produced by the Neocallimastigomycota, but not the Dikarya. However, the rationale behind other observed differences e.g. retention of specific intramembrane proteases, post-translational fucosylation capabilities, or the majority of focal adhesion proteins in the Neocallimastigomycota is not entirely clear.

Many of the observed structural, metabolic, and genomic traits within the C1A genome are not shared with other early-branching fungal relatives or non-fungal Opisthokonts, and hence could be regarded as Neocallimastigomycota-specific adaptations to the anaerobic gut environment. The mitochondrial reductive evolution to a hydrogenosome, the apparent substitution of ergosterol with tetrahymanol in the cell membrane (since oxygen is required for squalene epoxidation, steroid ring demethylation, and ring unsaturation during ergosterol biosynthesis (94)), and the sole dependence on mixed acid fermentation pathway for pyruvate

metabolism and energy production in strain C1A are clear adaptations to anaerobiasis. The development of cellulosomes, and the acquisition of many GH enzymes could be viewed as an adaptation to improve the access, speed, and efficacy of biomass degradation.

In addition to metabolic adaptations to an O_2-independent mode of metabolism and organelle development via reductive evolution and gene acquisition, evolution of anaerobic fungi in the rumen and gut of herbivores appears to have triggered multiple genome-wide patterns. These include the possession of a large genome, the presence of large intergenic regions, the low (17.0%) G+C content, and the occurrence of a high level of gene duplication and microsatellite repeats (Fig 2.1). We argue that these genome-wide patterns are due to genetic drift, triggered by the low effective population sizes, bottlenecks in vertical transmission, and the asexual life style of anaerobic fungi. Species with low effective population sizes could tolerate slightly deleterious accumulation of DNA, resulting in the expansion in genome size, accumulation of repeats, and gene duplications (95, 96). In addition, genetic drift is also associated with an increase in the rate of non-lethal mutations, which tends to be biased towards adenine or thymine mutations such as cytosine deamination or guanine oxidation (97).

This study also highlights the extensive lignocellulolytic machinery and robust plant biomass degradation capability of strain C1A, observations which are consistent with prior studies identifying multiple cellulolytic and hemicellulolytic genes from anaerobic fungal strains (5, 14, 15, 68-90)**,** and documenting the capability of such strains to various plant substrates (98-101). Further, this study clearly demonstrates that the GH machinery in the C1A genome is markedly different from that of aerobic lignocellulolytic fungi. Such differences appear to be driven by physiological considerations, variations in the employed biomass degradation strategy, and habitat distinction. The recent demonstration of an O_2-dependent mode of metabolism for

GH61 enzymes could explain the ubiquity of this family in aerobic fungal genomes and its absence in the C1A genome (102). The utilization of a cellulosomal strategy for plant biomass degradation by strain C1A, compared to the free extracellular enzyme strategy of aerobic fungi could explain the identification of a large number of GH genes with dockerin domain in the C1A genome. Finally, the rumen habitat of the Neocallimastigomycota, and the widespread gene acquisition of bacterial GH genes in the C1A genome could explain the occurrence of GH genes belonging to families rarely encountered in aerobic fungi e.g. GH8 and GH48.

Interestingly, while gene acquisition from prevalent rumen bacterial lineages plays an important role in shaping the C1A lignocellulolytic machinery, a fraction of C1A GH genes were associated with bacterial lineages that are not regarded as integral members of the bovine rumen microbiota e.g. phyla Actinobacteria, Thermotoga, Deinococcus, and Chloroflexi. This intriguing observation could possibly be explained by the occasional identification of some of these phyla as minor components in the bovine rumen (103). Further, the extensively studied bovine rumen should not be regarded as the only possible habitat for anaerobic fungal gene acquisition, since anaerobic fungi have a wide distribution in the rumen, hindgut, and feces of multiple ruminant and non-ruminant herbivorous (2). Finally, it is important to note that evolution of anaerobic fungi from an Opisthokonta ancestor has preceded the evolution of their metazoan herbivores hosts (104). As such, anaerobic fungi could have acquired such genes prior to their association with the reptilian or mammalian alimentary tracts.

Transcriptional studies indicated that a large number of polymer-degrading GH genes are constitutively expressed in cellobiose-grown cultures. However, C1A cellulose-grown cultures exhibited significant increase in the overall transcription levels of specific cellulase (GH5) and cellobiohydrolase (GH9 and GH48) GH families, suggesting a prominent role for these three

families in cellulose metabolism. The increase in overall levels of transcripts belonging to a specific GH family was mainly attributed to the upregulation of a fraction of its genes (Fig 2.5c). For example, while the overall transcriptional level of GH48 cellobiohydrolases increased 8-fold in cellulose-grown cultures, only 5 out of 12 genes were upregulated in cellulose-grown cultures, while 2 genes were not significantly impacted, and 4 were downregulated. Factors influencing this observed selective regulation remains to be elucidated.

Finally, our results suggest that the lignocellulolytic capabilities of strain C1A could be exploited outside the rumen for the production of biofuels from plant biomass. The most promising approach for lignocellulosic biofuel production involves consolidated bioprocessing, which combines the saccharification of lignocellulose and the fermentation of the resulting sugars in a single step, and is carried out by a single microorganism or microbial consortia (105). Here, we show that strain C1A simultaneously couples the saccharification of the cellulosic and hemicellulosic fractions of plants to the fermentation of the resulting hexose and pentose sugars. Further, the invasive nature and filamentous growth pattern of these anaerobic fungi allows plant biomass degradation to proceed without pretreatment, and the process was significantly enhanced using mild pretreatments (Fig 2.6). To our knowledge, the extent of lignocellulosic biomass degradation by strain C1A has not been reported for a single microorganism in the absence of saccharification enzymes. Anaerobic fungi thus represent extremely promising microorganisms for exploitation in direct lignocellulolytic schemes. As part of its fermentative metabolism, strain C1A is capable of producing ethanol as a minor end product during pyruvate metabolism. Indeed, 1 copy of alcohol dehydrogenase has been identified, and C1A can tolerate up to 3% ethanol (data not shown). However, given its relatively low ethanol productivity and relatively low ethanol tolerance, efforts towards improving alcohol production and tolerance via

physiological and genetic manipulations are needed to improve ethanol productivity in this remarkable plant biomass-degrading anaerobic fungal strain.

Figure 2-1. Unique features in the *Orpinomyces* sp. strain C1A genome. (A) The C1A genome has the lowest G+C content in all fungal genomes described thus far. Averages and ranges were computed from the publicly available genomes of Ascomycota (n=90), Basidiomycota (n=16), basal fungal lineages (n=6), and Microsporidia (n=7). (B) The C1A genome has large intergenic non-coding regions compared to publicly available fungal genomes. List of 110 genomes for comparison is available in Table S5. (C) The C1A genome has the highest recorded abundances of simple sequence repeats within the Mycota, with the majority of repeats in intergenic regions and introns. (D) The homopolymeric (A/T) mononucleotide repeats in the C1A genome were not only abundant, but also reached lengths of up to 151bp in intergenic regions. Color-coding: genome (black); Intergenic region (red); Introns (green); cDNA (purple).

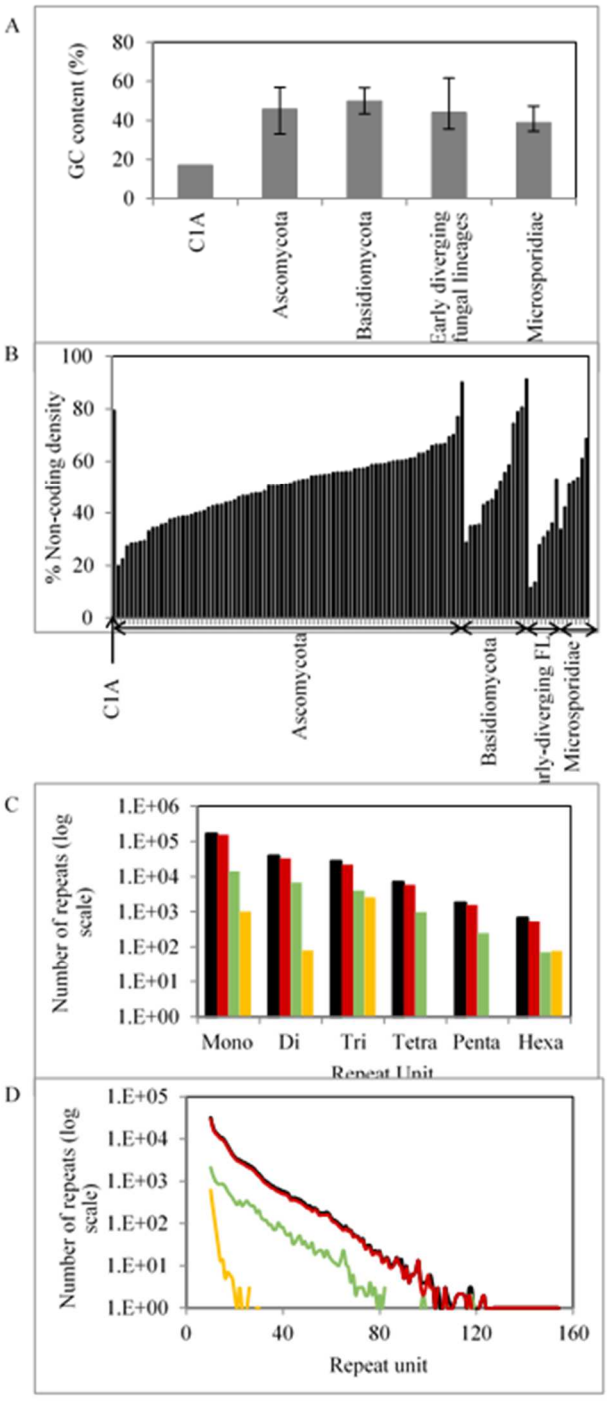

A
GC content (%)
80
60
40
20
0
C1A
Ascomycota
Basidiomycota
Early diverging fungal lineages
Microsporidiae
B
% Non-coding density
100
80
60
40
20
0
C1A
Ascomycota
Basidiomycota
Microsporidiae
C
Number of repeats (log scale)
1.E+06
1.E+05
1.E+04
1.E+03
1.E+02
1.E+01
1.E+00
Mono
Di
Tri
Tetra
Penta
Hexa
Repeat Unit
D
Number of repeats (log scale)
1.E+05
1.E+04
1.E+03
1.E+02
1.E+01
1.E+00
0
40
80
120
160
Repeat unit

Figure 2-1

Figure 2-2. (A) Gene duplication in C1A genome compared to other fungal genomes. Color-coding: C1A (blue); *Laccaria bicolor* (red); *Magnaporthe grisae* (green); *Saccharomyces cerivisiae* (purple); *Candida albicans* (yellow); *Encephalitozoon cuniculi* (orange). (B) Identification of C1A genes with at least one ortholog within: (-□-) Dikarya (n=116), or (-□-) Mycota (Dikarya + basal fungi n=4) genomes at different e-value cutoffs. (i) Panther classification (53) and putative phylogenetic affiliation based on blast first hit of C1A genes not encountered in Mycota (-□- non-fungal eukaryotes, -□- prokaryotes), and (ii) genes encountered only in basal fungi.

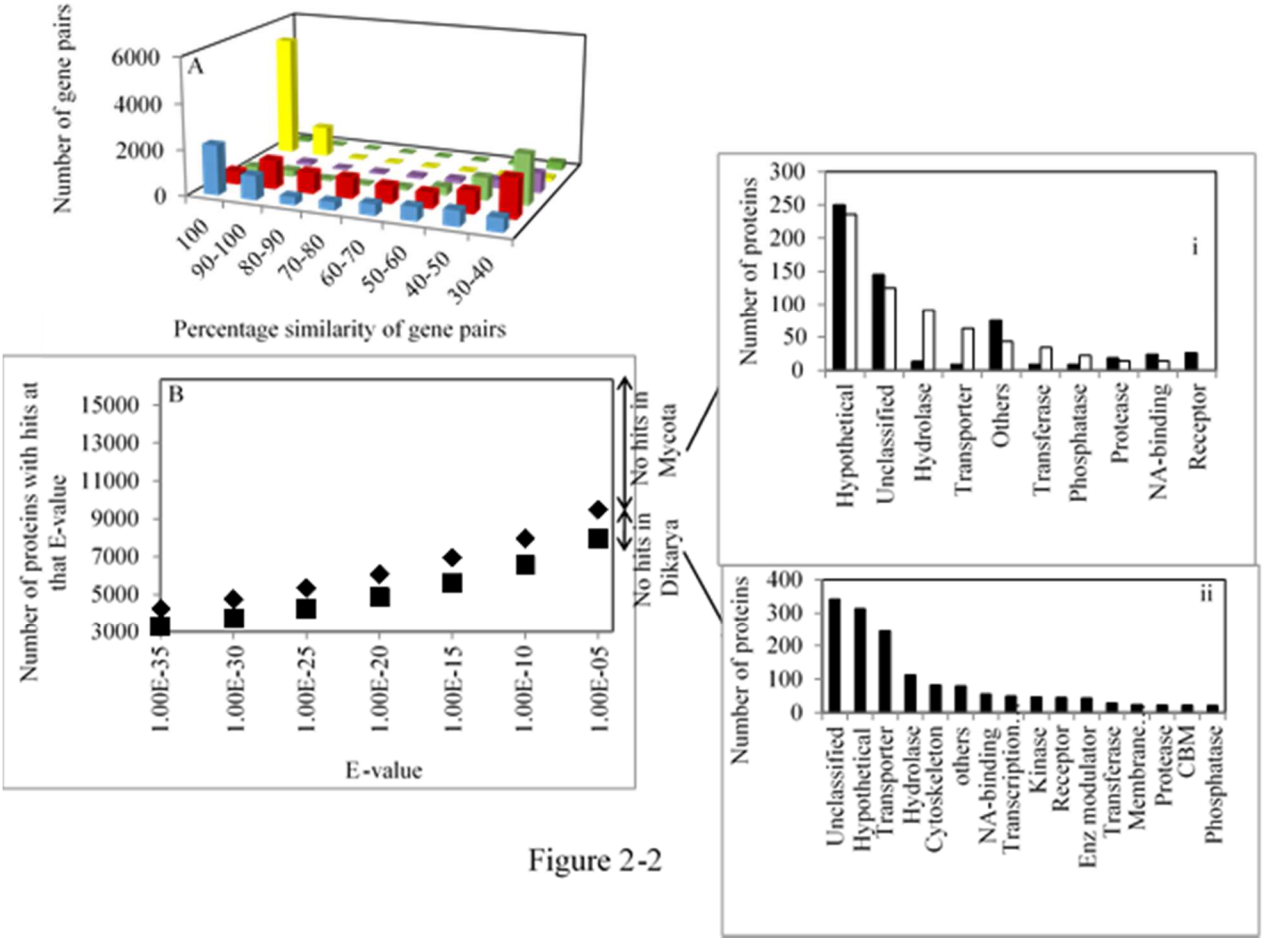

A
Number of gene pairs
6000
4000
2000
0
100
90-100
80-90
70-80
60-70
50-60
40-50
30-40
Percentage similarity of gene pairs
B
Number of proteins with hits at that E-value
15000
13000
11000
9000
7000
5000
3000
1.00E-35
1.00E-30
1.00E-25
1.00E-20
1.00E-15
1.00E-10
1.00E-05
E-value
No hits in Mycota
No hits in Dikarya
i
Number of proteins
300
250
200
150
100
50
0
Hypothetical
Unclassified
Hydrolase
Transporter
Others
Transferase
Phosphatase
Protease
NA-binding
Receptor
ii
Number of proteins
400
300
200
100
0
Unclassified
Hypothetical
Transporter
Hydrolase
Cytoskeleton
others
NA-binding
Transcription..
Kinase
Receptor
Enz modulator
Transferase
Membrane..
Protease
CBM
Phosphatase

Figure 2-2

Figure 2-3. Reconstruction of C1A hydrogenosome from genomic data. The double blue lines depict the hydrogenosomal outer and inner membrane. Various functional proteins groupings are color-coded: Outer mitochondrial membrane translocase complex components (TOM) are shown in green, Outer membrane sorting and assembly complex components (SAM) are shown in purple, Inner membrane complex components (TIM) are shown in orange, Intermembrane space import and assembly proteins (MIA) are shown in blue, Intermembrane space small TIMs are shown in yellow, mitochondrial peptidases (inner membrane peptidase (IMP), mitochondrial processing peptidase (MPP), mitochondrial intermediate peptidase (MIP), and the mitochondrial signal peptidase PCP1) are shown in red, mitochondrial distribution and morphology (MDM) proteins are shown in white, chaperones and co-chaperones are shown in brown, Fe-S assembly proteins are shown in black, and membrane transporters are shown in grey (VDAC: voltage-dependent anion channel, MCF: mitochondrial carrier family). Enzymes involved in pyruvate metabolism, substrate level phosphorylation and redox carriers regeneration are numbered: 1. Pyruvate kinase, 2. Phosphoenolpyruvate (PEP) carboxykinase, 3. Malate dehydrogenase, 4. Fumarase, 5. Fumarate reductase, 6. Citrate synthase, 7. Aconitase, 8. Isocitrate dehydrogenase, 9. Pyruvate formate lyase, 10. D-lactate dehydrogenase, 11. Acetaldehyde dehydrogenase, 12. Alcohol dehydrogenase, 13. Acetyl-CoA hydrolase (acetyl-CoA:succinyl transferase), 14. Succinyl-CoA synthase, 15. ATP synthase. SDH: succinate dehydrogenase, H_2ase: hydrogenase, I: complex I NADH dehydrogenase, Q: quinone.

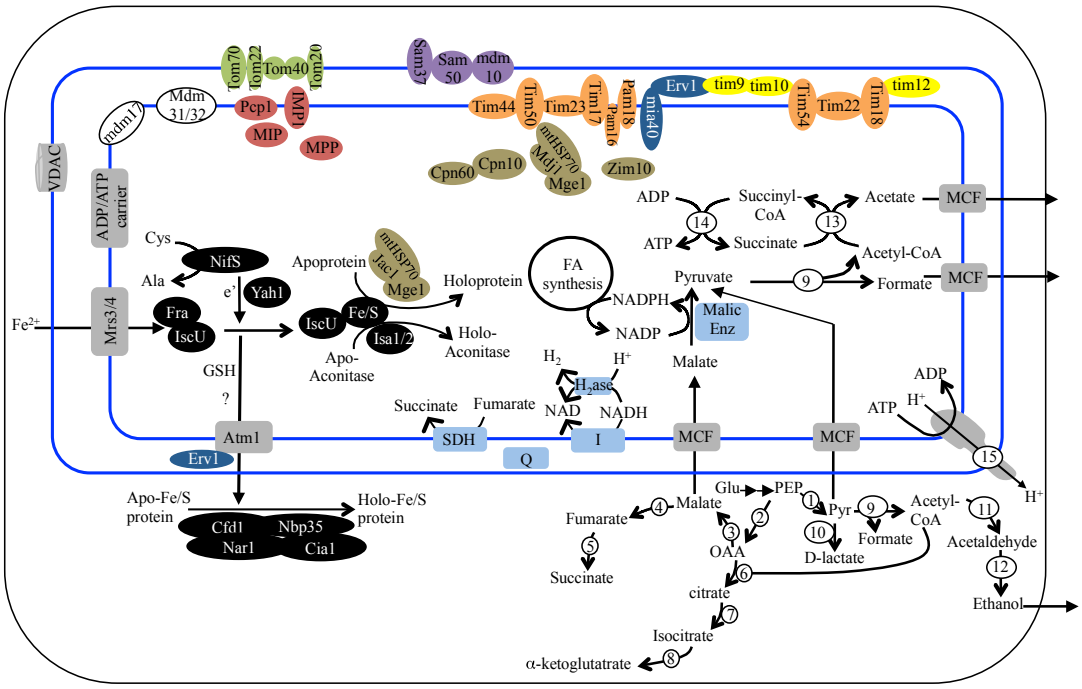
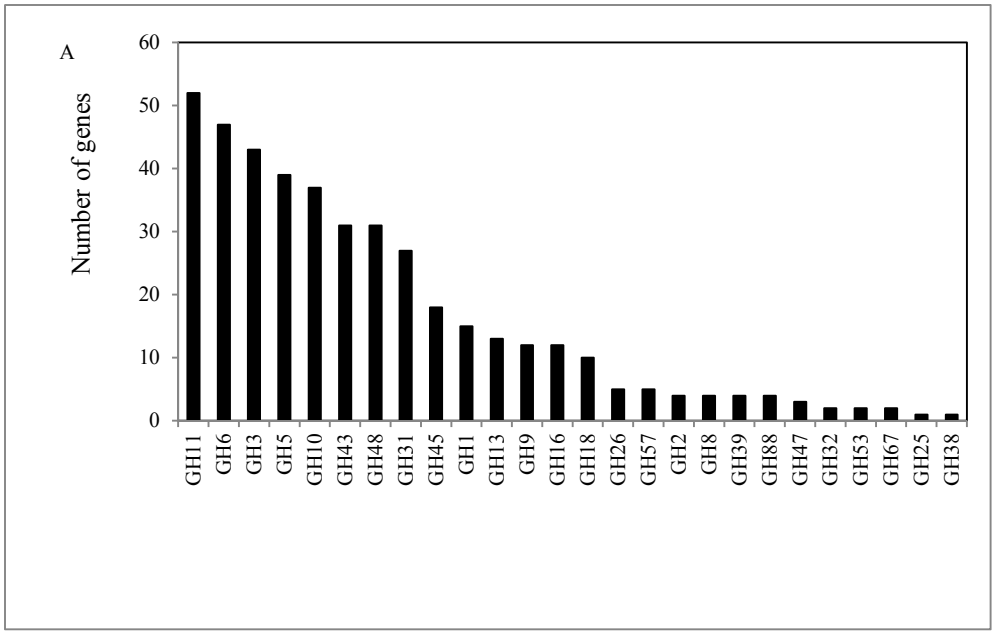

Tom70
Tom22
Tom40
Tom20
Sam37
Sam 50
mdm 10
Erv1
tim9
tim10
tim12
Tim54
Tim22
Tim18
Tim44
Tim50
Tim23
Tim17
Pam18
Pam16
tim40
mdm12
Mdm 31/32
Pcp1
Nde1
MIP
MPP
mtHSP70
Mdj1
Mge1
Cpn60
Cpn10
Zim10
VDAC
ADP/ATP carrier
Mrs3/4
Fe2+
Cys
NifS
Ala
e'
Yah1
Fra
IscU
GSH
?
Atm1
Erv1
Apoprotein
Iac1
Mge1
Holoprotein
IscU
Fe/S
Isa1/2
Apo-
Aconitase
Holo-
Aconitase
FA
synthesis
NADPH
NADP
Malic
Enz
Pyruvate
Malate
ADP
Succinyl-
CoA
Acetate
MCF
ATP
Succinate
14
13
Acetyl-CoA
9
Formate
MCF
H2
H+
H2ase
NAD
NADH
Succinate
Fumarate
SDH
Q
I
MCF
MCF
ADP
ATP
H+
15
H+
Apo-Fe/S
protein
Holo-Fe/S
protein
Cfd1
Nbp35
Nar1
Cia1
Glu
PEP
1
Pyr
9
Acetyl-
CoA
11
Acetaldehyde
12
Ethanol
4
Malate
2
3
OAA
10
D-lactate
Fumarate
5
Succinate
6
citrate
7
Isocitrate
8
α-ketoglutatrate

Figure 2-4. Glycoside hydrolase (GH) families in the C1A genome. (A) Number of C1A genes belonging to different GH families. (B) Principal-coordinate analysis biplot of the distribution of GH families in the C1A genome, compared to those in selected 19 other fungal and bacterial genomes. Genomes are represented by stars and GH families are represented by arrows. The arrow directions follow the maximal abundance, and their lengths are proportional to the maximal rate of change between genomes. Am: *Allomyces macrogynus*, At: *Anaerocellum thermophilum* DSM 6725, Ao: *Aspergillus Oryzae*, Bd: *Batrachochytrium dendrobatidis*, Co: *Caldicellulosiruptor obsidiansis*, Cp: *Clostridium phytofermentans* ISDg, Ct: *Clostridium thermocellum* ATCC 27405, Fs: *Fibrobacter succinogenes* subsp. *succinogenes* S85, Mg: *Magnaporthe grisea*, Mc: *Mucor circinelloides*, Mt: *Myceliophthora thermophila*, Nc: *Neurospora crassa*, Pa: *Podospora anserina*, Pp: *Postia placenta*, Ro: *Rhizopus oryzae*, Ra: *Ruminococcusus* 7, Sp: *Spizellomyces punctatus*, Tr: *Trichoderma reesei*. *Spizellomyces punctatus*, Tr: *Trichoderma reesei*.


A
Number of genes
60
50
40
30
20
10
0
GH11
GH6
GH3
GH5
GH10
GH43
GH48
GH31
GH45
GH1
GH13
GH9
GH16
GH18
GH26
GH57
GH2
GH8
GH39
GH88
GH47
GH32
GH53
GH67
GH25
GH38

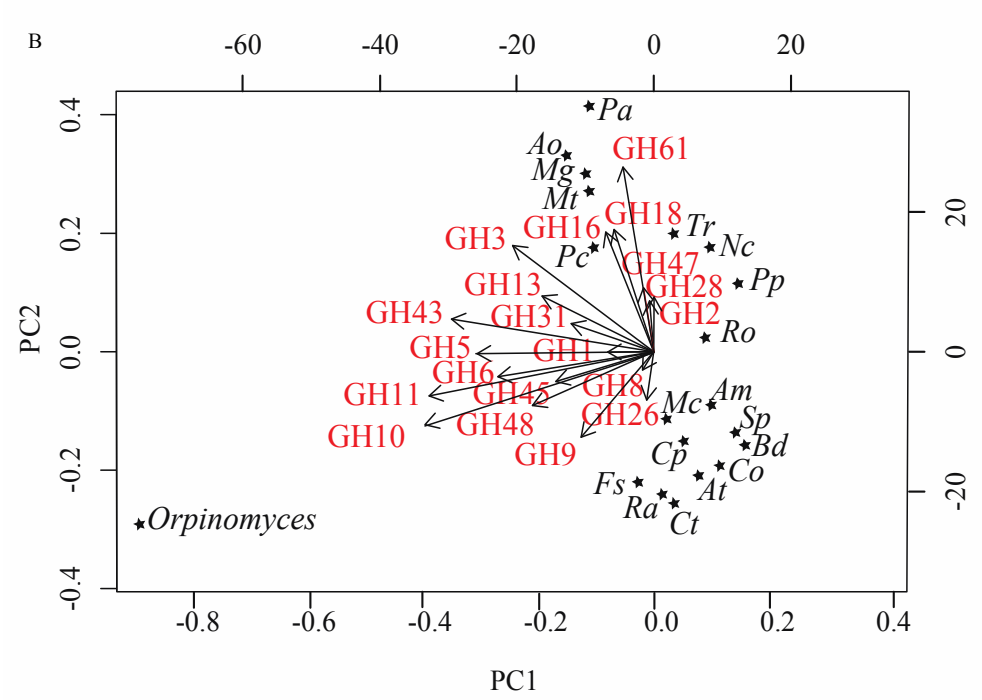

B
-60
-40
-20
0
20
0.4
0.2
0.0
-0.2
-0.4
PC2
PC1
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
20
0
-20
Pa
Ao
Mg
Mt
GH61
GH18
GH16
GH3
Tr
Nc
Pc
GH47
GH13
GH28
Pp
GH43
GH31
GH2
Ro
GH5
GH1
GH6
GH11
GH45
GH8
Am
Mc
GH26
Sp
GH10
GH48
Bd
GH9
Cp
Co
Fs
At
Ra
Ct
Orpinomyces

Figure 2-5. (A-B) Transcription levels of various GH families genes involved in cellulose degradation in cellulose (blue)- versus cellobiose (red)-grown cultures. Transcription levels are expressed as absolute transcripts per million (TPM) in (A), and as normalized TPM relative to a suite of glycolytic genes in (B). (C) Differential GH genes expression by strain C1A grown on cellulose (Y-axis) and cellobiose (X-axis) expressed as $\log_2$TPM. Only genes with TPM $\geq$10 in at least one growth condition were used to construct the graph. The 2 diagonal lines represent boundaries between genes upregulated (above the upper line), downregulated (below the lower line), or not significantly changed (in between the 2 lines) in cellulose- versus cellobiose-grown cultures. Color-coding: cellulases and cellobiohydrolases (blue); β-glucosidases (red); other polymer-degrading GHs (green); other oligomer-degrading GHs (grey).

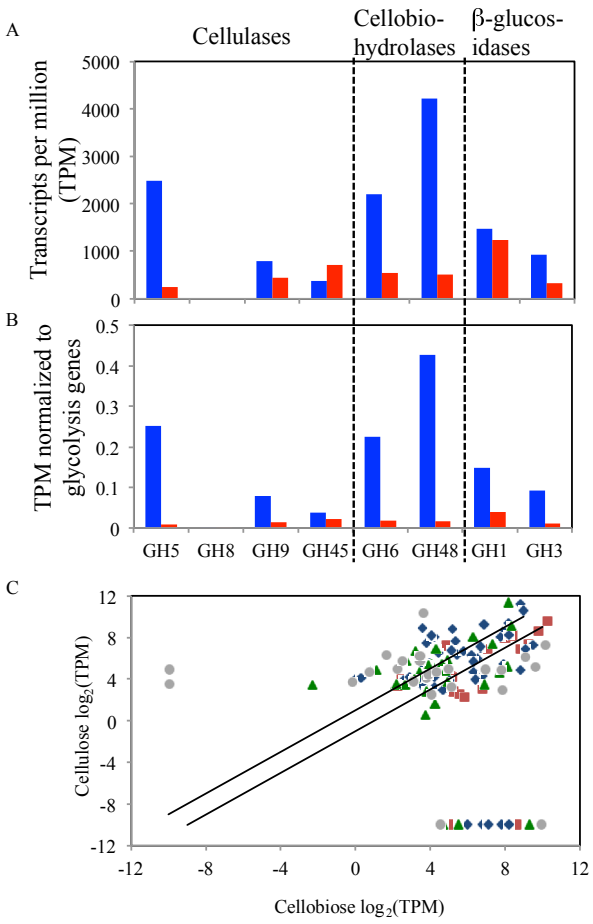

A
Cellulases
Cellobio-
hydrolases
β-glucos-
idases
Transcripts per million
(TPM)
5000
4000
3000
2000
1000
0
B
TPM normalized to
glycolysis genes
0.5
0.4
0.3
0.2
0.1
0
GH5
GH8
GH9
GH45
GH6
GH48
GH1
GH3
C
Cellulose log2(TPM)
12
8
4
0
-4
-8
-12
-12
-8
-4
0
4
8
12
Cellobiose log2(TPM)

Figure 2-6. Lignocellulolytic capabilities of strain C1A. Grams of dry weight (◆), cellulose (■), hemicellulose (▲), and lignin (●) lost in microcosms that contained untreated (A), sodium hydroxide-treated (B), acid-treated (C), and hydrothermolysis-treated (D) switchgrass. The percentages of dry weight (■), cellulose (■), and hemicellulose (□) lost in microcosms with different types of untreated and sodium hydroxide treated plant materials are shown in (E).

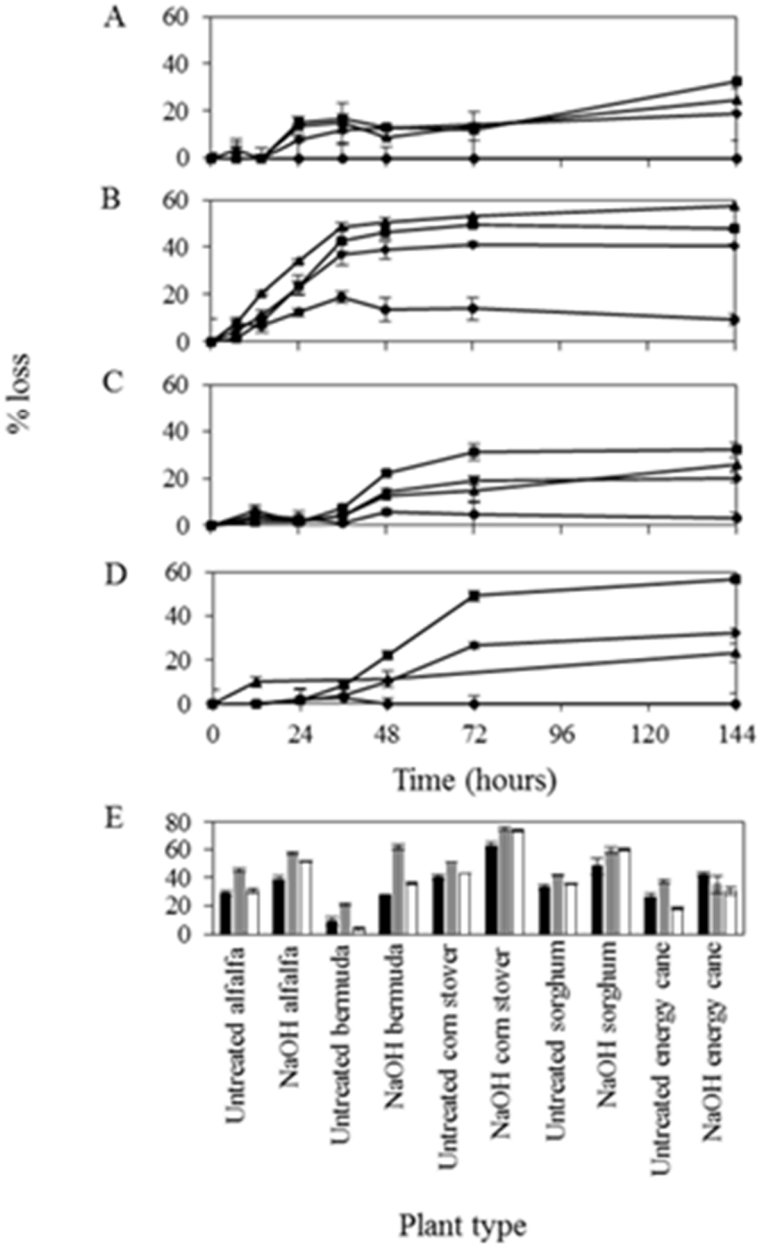

A
B
C
D
E
% loss
Time (hours)
Plant type
Untreated alfalfa
NaOH alfalfa
Untreated bermuda
NaOH bermuda
Untreated corn stover
NaOH corn stover
Untreated sorghum
NaOH sorghum
Untreated energy cane
NaOH energy cane

Table 2-1- Salient differences identified between strain C1A genome, basal fungal genomes, Dikarya Fungal genomes, and unicellular Opisthokont genomes.

Category	Gene ID	Product name	Present in basal fungi				Present in Dikarya fungi	Present in Opisthokonts	
			Allomyces macrogynus ATCC 38327	*Batrachochytrium dendrobatidis* JEL423	*Spizellomyces punctatus* DAOM BR117	*Rhizopus oryzae* RA 99-880		*Monosiga brevicollis*	*Capsaspora owczarzaki*
1-Intramembrane proteases.									
A. Gamma secretase complex	2510864531	Aph-1 protein	yes	yes	yes	yes	no	yes	yes
	Orpinomyces_14795	Nicastrin	no	yes	yes	yes	no	yes	yes
	Orpinomyces_14978	Nicastrin	no	yes	yes	yes	no	yes	yes
	Orpinomyces_3406	Nicastrin	no	no	no	no	no	no	no
	2510858953	Presenilin enhancer-2	no	no	no	no	no	no	yes
	Orpinomyces_8327	Presenilin	yes	yes	yes	yes	no	yes	yes
	Orpinomyces_13802	Presenilin	yes	yes	yes	yes	no	yes	yes
B. Site 2 peptidase	Orpinomyces_8676	Peptidase family M50A	no	no	yes	no	no	no	no
2-Post-translational protein modification	Orpinomyces_8992	GT family 10 (fucosyltransferase)	no	no	no	yes	no[a]	no	no
	Orpinomyces_15723	GT family 10 (fucosyltransferase)	no	no	no	yes	no[a]	no	no
3-Focal adhesion proteins									
A. Adaptor proteins	2510867334	paxillin	no	no	no	yes	no	yes	yes
	Orpinomyces_16019	Talin	no	no	no	no	no	yes	yes
	Orpinomyces_11547	Band4.1/ Talin	no	yes	yes	no	no	yes	yes
	Orpinomyces_13613	Vinculin family	no	no	no	no	no	no	no
	Orpinomyces_13888	Vinculin family	no	no	yes	no	no	yes	no
	Orpinomyces_8561	Vinculin family	no	yes	yes	no	no	yes	yes
B. IPP complex components	Orpinomyces_13334	Integrin-linked kinase	no	no	yes	no	no	no	yes
	Orpinomyces_16310	Integrin-linked kinase	no	yes	yes	no	no	no	yes
	2510867610	α-parvin	no	no	yes	no	no	no	yes
	Orpinomyces_13214	PINCH-1	yes	yes	yes	yes	no	yes	yes
C. Downstream-acting elements	Orpinomyces_3201	ROCK	no	yes	yes	no	no	yes	no
4. Axonemal proteins	Several (43) proteins were identified in the C1A genome and are thought to be involved in axoneme structure and function.		Some[b]	Some[b]	Some[b]	Some[b]	no	Some[c]	Some[c]
5. Protease inhibitors									
A. Serine proteases (serpins) family I4	Orpinomyces_14234	Protease inhibitor I4	no	no	no	no	no	no	no
	Orpinomyces_14684	Protease inhibitor I4	no	no	no	no	no	no	no
	Orpinomyces_6138	Protease inhibitor I4 with Dockerin domain	no	no	no	no	no	no	no
	Orpinomyces_2311	Protease inhibitor I4 with Dockerin domain	no	no	no	no	no	no	no
B. Apoptosis inhibitors (BIR domain) cysteinyl protease inhibitors family I32	Orpinomyces_4452	Apoptosis inhibitor and related BIR domain proteins	no	no	yes	yes	Yes[d]	no	no
C. Cysteinyl protease inhibitor family I42	2510865883	Chagasin	yes	no	no	no	no	no	no
	2510871764	Chagasin	no	no	no	no	no	no	no
	2510864005	Chagasin	no	no	no	no	no	no	no

[a]In all Dikarya genomes currently available in Genbank database, only one gene annotated as "hypothetical protein" in *Puccinia graminis* (EFP84060) with low sequence similarity to plant fucosyltransferases has been identified.

[b]Homologues for some of the C1A axonemal proteins were found in the Basal fungal genomes. None of the BBSome proteins were detected in any of the fungal genomes.

[c]Homologues for axonemal proteins were detected before in Opisthokonts representatives.

[d]Homologues are found only in *Coccidioides immitis* RS, *Coccidioides posadasii* C735 delta SOWgp, *Coprinopsis cinerea* okayama7#130, *Fusarium oxysporum* lycopersici FGSC 4286, *Gibberella zeae* PH-1, *Nosema ceranae* BRL01, *Phanerochaete chrysosporium* RP-78, *Saccharomyces cerevisiae* AWRI1631, *Saccharomyces cerevisiae* YJM789, *Saitoella complicata* NRRL Y-17804, *Scheffersomyces stipitis* CBS 6054, *Ustilago maydis* 521

References

1. **Orpin CG**. 1975. Studies on the rumen flagellate *Neocallimastix frontalis*. J. Gen. Microbiol. **91**: 249-262.

2. **Orpin CG**. 1994. Anaerobic fungi: Taxonomy, biology, and distribution in nature, p. 1-45. *In* Mountfort D O and Orpin C G (ed), Anaerobic fungi: biology, ecology, and function. Marcel Dekker, Inc.

3. **Ma L-J, Ibrahim AS, Skory C, Grabherr MG, Burger G, Butler M, Elias M, Idnurm A, Lang BF, Sone T, Abe A, Calvo SE, Corrochano LM, Engels R, Fu J, Hansberg W, Kim J-M, Kodira CD, Koehrsen MJ, Liu B, Miranda-Saavedra D, O'Leary S, Ortiz-Castellanos L, Poulter R, Rodriguez-Romero J, Ruiz-Herrera J, Shen Y-Q, Zeng Q, Galagan J, Birren BW, Cuomo CA, and Wickes BL**. 2009. Genomic analysis of the basal lineage fungus *Rhizopus oryzae* reveals a whole-genome duplication. PLoS Genet. **5**: e1000549.

4. **Joneson S, Stajich JE, Shiu S-H, and Rosenblum EB**. 2011. Genomic transition to pathogenicity in Chytrid Fungi. PLoS Pathog. **7**: e1002338.

5. **Ljungdahl LG**. 2008. The cellulase/hemicellulase system of the anaerobic fungus *Orpinomyces* PC-2 and aspects of its use. Ann. N. Y. Acad. Sci. **1125**: 308-321.

6. **Mentel M and Martin W**. 2008. Energy metabolism among eukaryotic anaerobes in light of Proterozoic ocean chemistry. Phil. Trans. R. Soc. B. **363**: 2717-2729.

7. **Schneider RE, Brown MT, Shiflett AM, Dyall SD, Hayes RD, Xie Y, Loo JA, and Johnson PJ**. 2011. The *Trichomonas vaginalis* hydrogenosome proteome is highly

reduced relative to mitochondria, yet complex compared with mitosomes. Int. J. Parasitol. **41**: 1421-1434.

8. **Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, Yu R-Y, van der Giezen M, Tielens AGM, and Martin WF**. 2012. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. Microbiol. Mol. Biol. Rev. **76**: 444-495.

9. **Tachezy J and Doležal P**. 2007. Iron-Sulfur proteins and iron-sulfur cluster assembly in organisms with hydrogenosomes and mitosomes., p. 105-134. *In* Martin W F and Müller M (ed), Origin of Mitochondria and Hydrogenosomes. Springer-Verlag, Berlin Heidelberg.

10. **Lowe SE, Griffith GG, Milne A, Theodorou MK, and Trinci APJ**. 1987. The life cycle and growth kinetics of an anaerobic rumen fungus. J. Gen. Microbiol. **133**: 1815-1827.

11. **Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PE, Eriksson OE, Huhndorf S, James T, Kirk PM, L¸cking R, Thorsten LH, Lutzoni F, Matheny PB, McLaughlin DJ, Powell MJ, Redhead S, Schoch C, Spatafora JW, Stalpers JA, Vilgalys R, Aime AM, A. Aptroot A, R. Bauer R, D. Begerow D, G. Benny GL, L. A. Castlebury LA, Crous PC, Dai Y-C, Gams W, Geiser D, Griffith GW, Gueidan C, Hawksworth DL, Hestmark G, Hosaka K, Humber RA, Hyde KD, Ironside JE, Kıljalg U, Kurtzman CP, Larsson K-H, Lichtwardt R, Longcore J, Miadlikowska J, Miller A, Moncalvo J-M, Mozley-Standridge S, Oberwinkler F, E. Parmasto E, V. Reeb V, Rogers JD, Roux C, Ryvarden L, Sampaio J, Sch AF, J. Sugiyama, Thorn RG, Tibell L, Untereiner WA, Walker C, Wang Z, Weir A, Weiss M, White M,**

Winka K, Yao Y-J, and Zhang N. 2007. A higher-level phylogenetic classification of the Fungi. Mycol. Res. **111**: 509-547.

12. **Lowe SE, Theodorou M, and Trinci A**. 1987 Growth and fermentation of an anaerobic rumen fungus on various carbon sources and effect of temperature on development. Appl. Environ. Microbiol. . **53**: 1210-1215.

13. **Raghothama S, Eberhardt RY, Simpson P, Wigelsworth D, White P, Hazlewood GP, Nagy T, Gilbert HJ, and Williamson MP**. 2001. Characterization of a cellulome dockerin domain from the anaerobic fungus *Piromyces equi.* Nature Struct. Biol. **8**: 775-778.

14. **Williams AG and Orpin CG**. 1987. Glycoside hydrolase enzymes present in the zoospore and vegetative growth stages of the rumen fungi *Neocallimastix patriciarum, Piromonas communis*, and an unidentified isolate, grown on a range of carbohydrates. Can. J. Microbiol. **33**: 427-434.

15. **Williams AG and Orpin CG**. 1987. Polysaccharide-degrading enzymes formed by three species of anaerobic rumen fungi grown on a range of carbohydrate substrates. Can. J. Microbiol. **33**: 418-426.

16. **Theodorou MK, Brookman J, and Trinci A**. 2005. Anaerobic fungi, p. 55-66. *In* Makkar H P and McSweeney C S (ed), Methods in gut microbial ecology for ruminants. Springer, Dordrech, the Netherlands.

17. **Marvin-Sikkema F, Richardson A, Stewart C, Gottschal J, and Prins R**. 1990. Influence of hydrogen-consuming bacteria on cellulose degradation by anaerobic fungi. Appl. Environ. Microbiol. **56**: 3793-3797.

18. **Bryant M**. 1972. Commentary on the Hungate technique for culture of anaerobic bacteria. Am. J. Clin. Nutr. **25**: 1324-1328.

19. **Balch WE and Wolfe R**. 1976. New approach to the cultivation of methanogenic bacteria: 2-mercaptoethanesulfonic acid (HS-CoM)-dependent growth of *Methanobacterium ruminantium* in a pressureized atmosphere. Appl. Environ. Microbiol. . **32**: 781-791.

20. **Brownlee AG**. 1989. Remarkably AT-rich genomic DNA from the anaerobic fungus *Neocallimastix.* Nucleic Acids Res. **17**: 1327-1335.

21. **Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Cheetham RK, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Catenazzi MCE, Chang S, R RNC, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DR, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fajardo KVF, Furey WS, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT,**

Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Jones TAH, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ng BL, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Pinkard DC, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Rodriguez AC, Roe PM, Rogers J, Bacigalupo MCR, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Sohna JES, Spence EJ, Stevens K, Sutton N, L.Szajkowski, L.Tregidgo C, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, .Oaks FL, Lundberg PL, Klenerman D, Durbin R and Smith AJ. 2008 Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456**: 53-59

22. **Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson B, Chaudhuri B, F.Christians, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong K, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu**

D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, and Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. Science. **323** 133-138.

23. **Zerbino DR and Birney E**. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. . **18** 821-829.

24. **Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, and Phillippy AM**. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. **30**: 693-700.

25. **Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan ML, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, and Venter JC**. 2000. A whole-genome assembly of *Drosophila*. Science **287**: 2196-2204. .

26. **Miller JR, Delcher AL, Kore S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, and Sutton G**. 2008. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics **24**: 2818-2824.

27. **Magoc T and Salzberg SL**. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics **27** 2957-2963.

28. **Ewing B, Hillier L, Wendl MC, and Green P**. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8**: 175-185.

29. **Parra G, Bradnam K, and Korf I**. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. **23**: 1061-1067.

30. **Wang Z, Gerstein M, and Snyder M**. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. . **10** 57-63.

31. **Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Palma Fd, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, and Regev A**. 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29**: 644-652.

32. **Li B and Dewey CN**. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. **12**: 323.

33. **Majoros WH, Pertea M, and Salzberg SL**. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. **20**: 2878-2879.

34. **Stanke M, Schöffmann O, Morgenstern B, and Waack S**. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics **7**: 6.

35. **Haas BJ, Delcher AL, Mount SM, Wortman JR, Jr RKS, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, and White O**. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. . **31**: 5654-5666.

36. **Wu TD and Watanabe CK**. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. **21** 1859-1875.

37. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL**. 2009 BLAST+: architecture and applications. BMC Bioinformatics **10**: 421.

38. **Eddy SR**. 2011. Accelerated Profile HMM Searches. PLoS Comput. Biol. **10**: e1002195.

39. **Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, and Finn RD**. 2012. The Pfam protein families database. Nucleic Acids Res. **40** D290-301. .

40. **Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, Mackie RI, Pennacchio LA, Tringe SG, Visel A, Woyke T, Wang Z, and Rubin EM**. 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science. **331**: 463-467.

41. **Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, and Henrissat B**. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. **37**: D233-238.

42. **Petersen TN, Brunak S, Heijne Gv, and Nielsen H**. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods. **8**: 785-786.

43. **Krogh A, Larsson B, Heijne Gv, and Sonnhammer EL**. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. **305**: 567-580.

44. **Jurka J, Bao W, Kojima K, and Kapitonov VV**. 2001. Repetitive elements: Bioinformatic identification, classification and analysis. eLS. **DOI**: 10.1002/9780470015902.a0005270.pub2.

45. **Dhillon B and Goodwin SB**. 2011. Identification and annotation of repetitive sequences in fungal genomes, p. 33-50. *In* Xu J-R and Bluhm B H (ed), Fungal genomics, vol. 722. Humana Press.

46. **Author**. 2006-2010. Phobos 3.3.11 of Work

47. **Floudas D, Binder M, R.Riley, Barry K, Blanchette RA, Henrissat B, Martínez AT, Otillar R, Spatafora JW, Yadav JS, Aerts A, Benoit I, Boyd A, Carlson A, Copeland A, Coutinho PM, Vries RPd, Ferreira P, Findley K, Foster B, J.Gaskell, Glotzer D, Górecki P, Heitman J, Hesse C, Hori C, Igarashi K, Jurgens JA, Kallen N, Kersten P, Kohler A, Kües U, Kumar TKA, Kuo A, LaButti K, Larrondo LF, Lindquist E, Ling A, Lombard V, Lucas S, Lundell T, Martin R, McLaughlin DJ, Morgenstern I, Morin E, Murat C, Nagy LG, Nolan M, Ohm RA, Patyshakuliyeva A, Rokas A, Ruiz-Dueñas FJ, Sabat G, Salamov A, Samejima M, Schmutz J, Slot JC, John FS, Stenlid J, Sun H, Sun S, Syed K, Tsang A, Wiebenga A, Young D, Pisabarro A, Eastwood DC, Martin F, Cullen D, Grigoriev IV, and Hibbett DS**. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science. **336**: 1715-1719.

48. **Price AL, Jones NC, and Pevzner PA**. 2005. De novo identification of repeat families in large genomes. Bioinformatics. **21**: i351-i358.

49. Smit AFA, Hubley R, and Green P, *RepeatMasker. Current Version: open-3.3.0*. 2003.

50. **Xu Z and Wang H**. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35**: W265-W268.

51. **Schattner P, Brooks AN, and Lowe TM**. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. **33**: W686-W689.

52. **Claros MG and Vincens P**. 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur. J. Biochem. **241**: 779-786.

53. **Thomas PD, Campbell MJ, Kejariwal A, Huaiyu M, Karlak B, Daverman R, Diemer K, Muruganujan A, and Narechania A**. 2003. PANTHER: A library of protein families and subfamilies indexed by function. Genome Res. **13**: 2129-2141.

54. **Suryawati L, Wilkins WR, Bellmer DD, Huhnke RL, Maness NO, and Banat IM**. 2008. Simultaneous saccharification and fermentation of Kanlow switchgrass pretreated by hydrothermolysis using *Kluyveromyces marxianus* IMB4. Biotechnol. Bioeng. **101**: 894-902.

55. **Xu J, Cheng JJ, Sharma-Shivappa RR, and Burns JC**. 2010. Sodium hydroxide pretreatment of switchgrass for ethanol production. Energy Fuels. **24**: 2113-2119.

56. **Torget R, Werdene P, Himmel M, and Grohmann K**. 1990. Dilute acid pretreatment of short rotation woody and herbaceous crops. Appl. Biochem. Biotechnol. **24**: 115-126.

57. **Sluiter A, Hames B, Ruiz R, Scarlata C, Templeton D, and Crocker D**. 2008 Determination of structural carbohydrates and lignin in biomass. Technical Report NREL/TP-510-48825.

58. **Sluiter A and Sluiter J**. 2010. Summative Mass Closure. Technical Report NREL/TP-510-42618.

59. **Faga B WMaIB**. 2010. Ethanol production through simultaneous saccharification and fermentation of switchgrass using *Saccharomyces cerevisiae* D5A and thermotolerant *Kluyveromyces marxianus* IMB strains. Bioresource Technol. **101**: 2273-2279.

60. **Karaoglu H, Lee CMY, and Meyer W**. 2004. Survey of simple sequence repeats in completed fungal genomes. Mol. Biol. Evol. **22**: 639-649.

61. **Sebé-Pedrós A, Roger AJ, Lang FB, King N, and Ruiz-Trillo I**. 2010. Ancient origin of the integrin-mediated adhesion and signaling machinery. Proc. Natl. Acad. Sci. **107**: 10142-10147.

62. **Steenbakkers PJM, Irving JA, Harhangi HR, Swinkels WJC, Akhmanova A, Dijkerman R, Jetten MSM, van der Drift C, Whisstock JC, and Op den Camp HJM**. 2008. A serpin in the cellulosome of the anaerobic fungus *Piromyces* sp. strain E2. Mycol. Res. **112**: 999-1006.

63. **Rada P, Doležal P, Jedelský PL, Bursac D, Perry AJ, Šedinová M, Smíšková K, Novotný M, Beltrán NC, Hrdý I, Lithgow T, and Tachezy J**. 2011. The core components of organelle biogenesis and membrane transport in the hydrogenosomes of *Trichomonas vaginalis.* PLoS ONE. **6**: e24428.

64. **Boxma B, Voncken F, Jannink S, Van Alen T, Akhmanova A, Van Weelden SWH, Van Hellemond JJ, Ricard G, Huynen M, Tielens AGM, and Hackstein JHP**. 2004. The anaerobic chytridiomycete fungus *Piromyces* sp. E2 produces ethanol via pyruvate:formate lyase and an alcohol dehydrogenase E. Mol. Microbiol. **51**: 1389-1399.

65. **Yarlett N, Orpin CG, Munn EA, Yarlett NC, and Greenwood CA**. 1986. Hydrogenosomes in the rumen fungus *Neocallimastix patriciarum.* Biochem. J. **236**: 729-739.

66. **Hackstein JHP, Baker SE, van Hellemond JJ, and Tielens AGM**. 2008. Hydrogenosomes of anaerobic chytrids: an alternative way to adapt to anaerobic environments, p. 147-162. *In* Tachezy J (ed), Hydrogenosomes and mitosomes: mitochondria of anaerobic eukaryotes. Springer-Verlag, Berlin Heidelberg.

67. **Hackstein J, Tjaden J, Koopman WJH, and Huynen M**. 2007. Hydrogenosomes (and related organelles, either) are not the same., p. 135-160. *In* Martin W F and Müller M (ed), Origin of mitochondria and hydrogenosomes. Springer Verlag, New York, NY.

68. **Chen H, Hopper S, Li X-L, Ljungdahl L, and Cerniglia C**. 2006. 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.

69. **Chen H, Li X-L, Blum D, Ximenes E, and Ljungdahl L**. 2003. CelF of *Orpinomyces* PC-2 has an intron and encodes a cellulase (CelF) containing a carbohydrate-binding module. Appl. Biochem. Biotechnol. **108**: 775-785.

70. **Chen H, Li XL, and Ljungdahl LG**. 1997. Sequencing of a 1,3-1,4-beta-D-glucanase (lichenase) from the anaerobic fungus *Orpinomyces* strain PC-2: properties of the enzyme expressed in *Escherichia coli* and evidence that the gene has a bacterial origin. J. Bacteriol. **179**: 6028-6034.

71. **Chen H-L, Chen Y-C, Lu M-Y, Chang J-J, Wang H-T, Ke H-M, Wang T-Y, Ruan S-K, Wang T-Y, Hung K-Y, Cho H-Y, Lin W-T, Shih M-C, and Li W-H**. 2012. A highly efficient beta-glucosidase from the buffalo rumen fungus *Neocallimastix patriciarum* W5. Biotechnol. Biofuels. **5**: 24.

72. **Durand R, Rascle C, and Fèvre M**. 1996. Molecular characterization of xyn3, a member of the endoxylanase multigene family of the rumen anaerobic fungus *Neocallimastix frontalis*. Curr. Genet. **30**: 531-540.

73. **Eberhardt RY, Gilbert HJ, and Hazlewood GP**. 2000. Primary sequence and enzymic properties of two modular endoglucanases, Cel5A and Cel45A, from the anaerobic fungus *Piromyces eq*ui. Microbiology. **146**: 1999-2008.

74. **Fanutti C, Ponyi T, Black GW, Hazlewood GP, and Gilbert HJ**. 1995. The conserved noncatalytic 40-residue sequence in cellulases and hemicellulases from anaerobic fungi functions as a protein docking domain. J. Biol. Chem. **270**: 29314-29322.

75. **Fujino Y, Ogata K, Nagamine T, and Ushida K**. 1998. Cloning, sequencing, and expression of an endoglucanase gene from the rumen anaerobic fungus *Neocallimastix frontalis* MCH3. Biosci. Biotechnol. Biochem. **62**: 1795-1798.

76. **Harhangi HR, Akhmanova A, Steenbakkers PJM, Jetten MSM, van der Drift C, and Op den Camp HJM**. 2003. Genomic DNA analysis of genes encoding (hemi-)cellulolytic enzymes of the anaerobic fungus *Piromyces* sp. E2. Gene. **314**: 73-80.

77. **Harhangi HR, Akhmanova AS, Emmens R, van der Drift C, de Laat WTAM, van Dijken JP, Jetten MSM, Pronk JT, and Op den Camp HJM**. 2003. Xylose metabolism in the anaerobic fungus *Piromyces* sp. strain E2 follows the bacterial pathway. Arch. Microbiol. **180**: 134-141.

78. **Harhangi HR, Freelove ACJ, Ubhayasekera W, van Dinther M, Steenbakkers PJM, Akhmanova A, van der Drift C, Jetten MSM, Mowbray SL, Gilbert HJ, and Op den Camp HJM**. 2003. Cel6A, a major exoglucanase from the cellulosome of the anaerobic fungi *Piromyces* sp. E2 and *Piromyces equi.* Biochim. Biophys. Acta. **1628**: 30-39.

79. **Harhangi HR, Steenbakkers PJM, Akhmanova A, Jetten MSM, van der Drift C, and Op den Camp HJM**. 2002. A highly expressed family 1 β-glucosidase with transglycosylation capacity from the anaerobic fungus *Piromyces* sp. E2. Biochim. Biophys. Acta. **1574**: 293-303.

80. **Li X-L, Ljungdahl L, Ximenes E, Chen H, Felix C, Cotta M, and Dien B**. 2004. Properties of a recombinant β-glucosidase from polycentric anaerobic fungus

Orpinomyces PC-2 and its application for cellulose hydrolysis. Appl. Biochem. Biotechnol. **113**: 233-250.

81. **Liu J-H, Selinger BL, Tsai C-F, and Cheng K-J**. 1999. Characterization of a *Neocallimastix patriciarum* xylanase gene and its product. Can. J. Microbiol. **45**: 970-974.

82. **Madhavan A, Tamalampudi S, Ushida K, Kanai D, Katahira S, Srivastava A, Fukuda H, Bisaria V, and Kondo A**. 2009. Xylose isomerase from polycentric fungus *Orpinomyces* gene sequencing, cloning, and expression in *Saccharomyces cerevisia*e for bioconversion of xylose to ethanol. Appl. Microbiol. Biotechnol. **82**: 1067-1078.

83. **Nicholson MJ, Theodorou MK, and Brookman JL**. 2005. Molecular analysis of the anaerobic rumen fungus *Orpinomyces* – insights into an AT-rich genome. Microbiology. **151**: 121-133.

84. **Pai C-K, Wu Z-Y, Chen M-J, Zeng Y-F, Chen J-W, Duan C-H, Li M-L, and Liu J-R**. 2010. Molecular cloning and characterization of a bifunctional xylanolytic enzyme from *Neocallimastix patriciarum.* Appl. Microbiol. Biotechnol. **85**: 1451-1462.

85. **Qiu X, Selinger B, Yanke LJ, and Cheng KJ**. 2000. Isolation and analysis of two cellulase cDNAs from *Orpinomyces joyonii.* Gene. **245**: 119-126.

86. **Steenbakkers P, Freelove A, Van Cranenbroek B, Sweegers B, Harhangi H, Vogels G, Hazlewood G, Gilbert H, and Op den Camp H**. 2002. The major component of the cellulosomes of anaerobic fungi from the genus *Piromyces* is a family 48 glycoside hydrolase. DNA seq. **13**: 313-320.

87. **Steenbakkers PJM, Harhangi HR, Bosscher MW, van der Hooft MMC, Keltjens JT, van der Drift C, Vogels GD, and op den Camp HJM**. 2003. Beta-glucosidase in

cellulosome of the anaerobic fungus *Piromyces* sp. strain E2 is a family 3 glycoside hydrolase. Biochem. J. **370**: 963-970.

88. **Steenbakkers PJM, Ubhayasekera W, Goossen HJAM, van Lierop EMHM, van der Drift C, Vogels GD, Mowbray SL, and Op den Camp HJM**. 2002. An intron-containing glycoside hydrolase family 9 cellulase gene encodes the dominant 90 kDa component of the cellulosome of the anaerobic fungus *Piromyces* sp. strain E2. Biochem. J. **365**: 193-204.

89. **Xue G-P, Gobius KS, and Orpin CG**. 1992. A novel polysaccharide hydrolase cDNA (celD) from *Neocallimastix patriciarum* encoding three multi-functional catalytic domains with high endoglucanase, cellobiohydrolase and xylanase activities. J. Gen. Microbiol. **138**: 2397-2403.

90. **Zhou L, Xue GP, Orpin CG, Black GW, Gilbert HJ, and Hazlewood GP**. 1994. Intronless celB from the anaerobic fungus *Neocallimastix patriciarum* encodes a modular family A endoglucanase. Biochem. J. **297**: 359-364.

91. **Scheller HV and Ulvskov P**. 2010. Hemicelluloses. Annu. Rev. Plant Biol. **61**: 263-289.

92. **Alvira P, Thomas-Pejo E, Ballesteros M, and Negro MJ**. 2010. Pretreatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: A review. Bioresource Technol. **101**: 4851-4861.

93. **Mosiera N, Wyman C, Dale B, Elander R, Le YY, Holtzapple M, and Ladisch M**. 2005. Features of promising technologies for pretreatment of lignocellulosic biomass. Bioresource Technol. **96**: 673-686.

94. **Waldbauer JR, Newman DK, and Summons RE**. 2011. Microaerobic steroid biosynthesis and the molecular fossil record of Archean life. Proc. Nat. Acad. Sci. **108**: 13409-13414.

95. **Lynch M and Conery JS**. 2003. The origins of genome complexity. Science. **302**: 1401-1404.

96. **Kellar YD and Ochman H**. 2011. Causes and consequences of genome expansion in Fungi. Genome Biol. Evol. **4**: 13-23.

97. **McCutcheon JP and Moran NA**. 2012. Extreme genome reduction in symbiotic bacteria. Nature Rev. Microbiol. **13**: 13-26.

98. **Sijtsma L and Tan B**. 1996. Degradation of perennial ryegrass leaf and stem cell walls by the anaerobic fungus *Neocallimastix* sp. strain CS3b. Appl. Environ. Microbiol. **62**: 1437-1440.

99. **Edwards JE, Kingston-Smith AH, Jimenez HR, Huws SA, Skot KP, Griffith GW, McEwan NR, and Theodorou MK**. 2008. Dynamics of initial colonization of nonconserved perennial ryegrass by anaerobic fungi in the bovine rumen. FEMS Microbiol. Ecol. **66**: 537-545.

100. **McSweeney CS, Dulieu A, Katayama Y, and Lowry JB**. 1994. Solubilization of lignin by the ruminal anaerobic fungus *Neocallimastix patriciarum.* Appl. Environ. Microbiol. **60**: 2985-2989.

101. **Borneman W, Hartley R, Morrison WH, Akin D, and Ljungdahl L**. 1990. Feruloyl and p-coumaroyl esterase from anaerobic fungi in relation to plant cell wall degradation. Appl. Microbiol. Biotechnol. **33**: 345-351.

102. **Quinlan RJ, Sweeneya MD, Leggiob LL, Ottenb H, Poulsenb J-CN, Johansenc KS, Kroghc KBRM, Jørgensenc CI, Tovborgc M, Anthonsenc A, Tryfonad T, Walterc CP, Dupreed P, Xua F, Daviese GJ, and Waltone PH**. 2011. Insights into the oxidative degradation of cellulose by a copper metalloenzyme that exploits biomass components. Proc. Natl. Acad. Sci. USA. **108**: 15079-15084.

103. **Kim M, Morrison M, and Yu Z**. 2011. Status of the phylogenetic diversity census of ruminal microbiomes. FEMS Microbiol. Ecol. **76**: 49-63.

104. **Lücking R, Huhndorf S, Pfister DH, Plata ER, and Lumbsch HT**. 2009. Fungi evolved right on track. Mycologia. **101**: 6.

105. **Xu Q, Singh A, and Himmel ME**. 2009. Perspectives and new directions for the production of bioethanol using consolidated bioprocessing of lignocellulose. Curr. Opin. Biotechnol. **20**: 364–371.

106. **Youssef, NH, Couger MB, Struchtemeyer CG, Liggenstoffer AS, Prade RA, Najar FZ Atiyeh HK, Wilkins MR, and Elshahed MS** 2013. Genome of the anaerobic fungus *Orpinomyces* sp. C1A reveals the unique evolutionary history of a remarkable plant biomass degrader. *Appl. Environ. Microbiol.* **79:**4620-4634.

CHAPTER III

TRANSCRIPTOMIC ANALYSIS OF LIGNOCELLULOSIC BIOMASS DEGRADTION BY THE ANAEROBIC FUNGAL ISOLATE *ORPINOMYCES* SP. STRAIN C1

Abstract

Anaerobic fungi reside in the rumen and alimentary tract of herbivores where they play an important role in the digestion of ingested plant biomass. The anaerobic fungal isolate *Orpinomyces* sp. strain C1A is an efficient biomass degrader, capable of simultaneous saccharification and fermentation of the cellulosic and hemicellulosic fractions in multiple types of lignocellulosic biomass. To understand the mechanistic and regulatory basis of biomass deconstruction in anaerobic fungi, we analyzed the transcriptomic profiles of C1A when grown on four different types of lignocellulosic biomass (alfalfa, energy cane, corn stover, and sorghum), versus a soluble sugar monomer (glucose). A total of 468.2 million read (70.2 GB) were generated and assembled into 27,506 distinct transcripts. CAZyme transcripts identified included 385, 246, and 44 transcripts belonging to 46, 13, and 8 different glycoside hydrolases (GH), carbohydrate esterases (CE), and polysaccharide lyases (PL) families, respectively. Examination of CAZyme transcriptional patterns indicates that strain C1A constitutively transcribes a high baseline level of CAZyme transcripts on glucose, with a broad induction of the majority of transcripts when grown on plant biomass. Further, strain C1A co-transcribes multiple functionally redundant enzymes for cellulose and hemicellulose saccharification that are mechanistically and structurally distinct. Analysis of fungal dockerin domain (FDD)-containing transcripts strongly suggests that anaerobic fungal cellulosomes represent distinct catalytic units capable of independently attacking and converting intact plant fibers to sugar monomers. Collectively, these results demonstrate that strain C1A achieves fast, effective biomass degradation by the simultaneous employment of a wide array of constitutively-transribed cellulosomal-bound and free enzymes with considerable functional overlap. We argue that the utilization of this indiscriminate strategy could be justified by the evolutionary history of

anaerobic fungi, as well as their functional role within their natural habitat in the herbivorous gut.

Introduction

Lignocellulosic biomass is a vast and underutilized resource for the production of biofuels. Compared to current schemes that rely on edible crops, lignocellulosic biomass utilization for sugar and biofuel production offers multiple advantages. It is abundant, renewable, and alleviates the moral stigma of using edible crops for industrial purposes. Further, the utilization of available lignocellulosic biomass overcomes the need for expanding farming acreage, and the subsequent increase in input of chemical fertilizers to the environment [1-3].

One of the most important procedures for the production of lignocellulosic biofuels involves the utilization of enzymes to extract sugar from plant polymers. The extracted sugars are then converted into biofuel using dedicated sugar-fermenting microorganisms [4]. However, the sugar extraction process from lignocellulosic biomass is far more complicated than sugar extraction from cereal grains (mainly corn in the US) [5]. This is due to the fact that the target substrates in lignocellulosic biomass (cellulose and hemicellulose) are structural components of plant cell walls, which are chemically bound to a variety of complex macromolecules (mainly lignin) [6]. Therefore**,** a combination of chemical pretreatments and exogenous enzyme cocktails addition are required for their effective mobilization and deconstruction [7, 8]. Enzymatic treatment of lignocellulosic biomass is a complex endeavor requiring multiple enzymes, a fact that significantly raises the cost of the process.

One alternative that circumvents the need for harsh pretreatments and exogenous enzymes amendments for the extraction of sugar monomers extraction from lignocellulosic biomass is the use of specialized microbial cultures for biomass deconstruction [9-11]. Microbial strains capable of cellulose and/or hemicellulose degradation produce not only cellulolytic and

xylanolytic enzymes targeting the backbone of these polymers, but also multiple accessory enzymes for removing side chains and breaking lignin-hemicellulose bonds [12-14]. Of special interest are lignocellulolytic microbes exhibiting anaerobic fermentative mode of metabolism since a significant fraction of the starting substrates could be recovered as fermentation end product.

The anaerobic gut fungi (Phylum Neocallimastigomycota) are unique in combining the resilience and invasiveness of fungi with the metabolic capabilities of anaerobic fermentative prokaryotes [15]. Anaerobic fungi are inhabitants of the rumen and alimentary tract of herbivores where they play an important role in the metabolism of ingested plant material [16]. It has been established that in such habitats, these organisms play a role akin to their aerobic counterparts in soils and streams. By attaching themselves to plant materials, they colonize and excrete extracellular enzymes that mobilize the structural plant polymers to be available to other microbes. Anaerobic fungi possess a powerful cellulolytic and hemicellulolytic enzymatic machinery [12] that aids in the required fast and efficient degradation of plant material in its relatively short residence time within the herbivorous gut [17]. Such capabilities have been demonstrated through experimental evaluation of anaerobic fungal isolates [18-21], biochemical characterization of anaerobic fungal enzymes [12], and recent genomic analysis of their lignocellulolytic repertoire [22].

We are currently exploring the utility of an anaerobic fungal isolate (*Orpinomyces* sp. strain C1A, henceforth referred to as strain C1A) for use in a consolidated bioprocessing framework for biofuels production. Developing an understanding of the genetic and regulatory mechanisms that enables efficient biomass degradation by strain C1A is central to gauging its

potential as a sugar extraction platform in biofuel production schemes. Our previous efforts have documented the lignocellulosic biomass degrading capabilities of C1A [22, 23], and the expansion of carbohydrate active enzymes (CAZymes) in its genome [22]. However, key questions regarding strain C1A lignocellulolytic capabilities remains unanswered. For example, patterns of differential transcription of various CAZyme families, especially those mediating apparently similar enzymatic activities, when grown on different types of substrates are currently unclear. Similarly, the differential transcriptional patterns and putative contribution to biomass degradation of the large number of CAZyme genes identified in C1A genome has not been investigated in anaerobic fungi. Finally, the transcriptional profiles and differential transcriptional patterns of fungal dockerin-containing (putatively cellulosomal-bound) in anaerobic fungi have yet to be determined in anaerobic fungi.

Here we present a detailed comparative analysis of the transcriptomic profiles of C1A when grown on four different types of lignocellulosic biomass (alfalfa, energy cane, corn stover, and sorghum), versus a soluble sugar monomer (glucose). Our analysis aimed at addressing the patterns of regulation of lignocellulosic gene transcription in C1A, the contribution of various CAZyme gene families to biomass degradation in C1A, and the significance of gene expansion and duplication observed in the C1A genome on its lignocellulolytic capabilities.

Methods

***Orpinomyces* sp. strain C1A.** Strain C1A was isolated from the feces of an Angus steer in our laboratory on a cellobiose-switchgrass medium as described previously [22]. Strain C1A is maintained by biweekly subculture on a cellobiose-rumen fluid medium as described previously [35].

Plant biomass. Samples of mature Sorghum (*Sorghum bicolor)* and mature energy cane (*Saccharum officianarum* var. Ho02) were obtained from Oklahoma State University experimental plots in Stillwater, OK. Dried alfalfa (*Medicago sativa*) was obtained from a local farm and ranch supplier. Samples of corn stover from *Zea mays* were obtained from the Industrial Agricultural Products Center at the University of Nebraska in Lincoln.

Experimental setup. All transcriptomic experiments were conducted in a rumen fluid-free basal media containing ($g.L^{-1}$): 0.5 g yeast extract, 0.47 g sodium butyrate, 2.4 g sodium acetate, 0.8 g sodium propionate, 2 g tryptone, 2 ml hemin solution (5 $g.L^{-1}$ in 1M NaOH), 9.3 ml of fatty acid solution (composition $ml.L^{-1}$: 11.7 ml isobutyric caid, 11.7 ml valeric acid, 11. 7 ml isovaleric acid, and 11.7 ml methylbutyric acid), 150 ml of mineral solution I (3 $g.L^{-1}$ K_2HPO_4), 150 ml of mineral solution II (composition $g.L^{-1}$: 3 g KH_2PO_4, 6 g $(NH_4)_2SO_4$, 6 g NaCL, 0.6 g $MgSO_4.7H_2O$, 0.6 g $CaCl_2.2H_2O$), 10 ml of Balch Vitamin solution (composition $mg.L^{-1}$: 2 mg biotin, 2 mg folic acid, 10 mg pyridoxine-HCl, 5 mg thiamine-HCl, 5 mg riboflavin, 5 mg nicotinic acid, 5 mg DL calcium pantothenate, 0.1 mg vitamin B12, 5 mg PABA, 5 mg lipoic acid), 1 ml Wolin's metal solution (composition $g.L^{-1}$: 0.5 g EDTA, 3 g $MgSO_4.7H_2O$, 0.5 g $MnSO_4.H_2O$, 1 g NaCl, 0.1 g $CaCl_2.2H_2O$, 0.1 g $FeSO_4.7H_2O$, 0.1 g $ZnSO_4.7H_2O$, 0.01 g $CuSO_4.7H_2O$, 0.01 g $AlK(SO_4)$, 0.01 g $Na_2MoO_4.2H_2O$, 0.01 g boric acid, 0.005 g Na_2SeO_4,

0.003 g $NiCl_2.6H_2O$, 0.1 g $CoCl_2.6H_2O$). After the medium was prepared, the pH was adjusted to 6.6. The medium was then dispensed under strictly anaerobic conditions as previously described [36, 37]. After the medium was dispensed, sodium carbonate (6 $g.L^{-1}$) was added and the bottles were stoppered, sealed, and autoclaved at 121°C for 20 minutes. After autoclaving, the bottles were cooled to room temperature. Bottles that were amended with plants materials were moved into an anaerobic glove bag (Coy Laboratory Products Grass Lake, MI), where the appropriate type of plant biomass (10 $g.L^{-1}$) was added. The bottles were then stoppered, sealed, and removed from the glove bag, and the headspace was replaced by repeated vacuuming and repressurization with 100% CO_2 (insert Balch reference). Bottles that contained glucose were amended with 3.75 $g.L^{-1}$ from an anaerobic, sterile stock solution. All experiments that were conducted with plant biomass and glucose were performed in duplicate. The inoculum source for these experiments consisted of strain C1A that was grown in a rumen fluid-free cellobiose media (same composition as above with the addition of 10 $g.L^{-1}$ cellobiose as the carbon source) until late log/early stationary phase. The inoculum was then centrifuged and resuspended in 20 mls of basal media with no carbon source. The experiment was started by adding this slurry of basal medium and fungal biomass (approximately 48 mg) into the appropriate bottles described above.

RNA extraction and sequencing. RNA extraction was conducted on late-log phase cultures after 48 hours of inoculation. Fungal biomass was harvested by vacuum filtration and ground into fine particles with a pestle under liquid nitrogen as previously described [35]. Total cellular RNA was extracted from ground fungal biomass using Epicentre MasterPure Yeast RNA Purification kit (Epicentre, Madison, WI, USA), stored in the provided RNase-free TE buffer, and quantified using Qubit fluorometer (Life Technologies, Carlsbad, CA, USA).

RNA-Seq [38] was conducted using the HiSeq 2000 platform with 125x2 paired-end read chemistry at the University of Georgia Genomics Facility (Athens, GA, USA). Biological replicate sequencing libraries for all conditions (glucose, corn stover, sorghum energy cane and alfalfa) were created with poly-A tailed mRNA enrichment using the standard Illumina TruSeq mRNA RNA-Seq protocol (http://www.utsouthwestern.edu/labs/next-generation-sequencing-core/assets/truseq-stranded-mrna-sample-prep-guide.pdf). The sequencing libraries had an average insert size of approximately of ~300 bp. Illumina recommend quality filter methodology was used for base calling.

Transcriptome Assembly and RNA-Seq Quantification. To represent all biological iso-forms present in the various growth conditions, the generated Illumina sequencing RNA-Seq [38] reads were assembled [39] using the *de novo* transcriptomic assembly program Trinity [40] using previously established protocols [41]. All settings for Inchworm, Chrysalis, and Butterfly steps were implemented according to the recommended protocol for fungal genomes, with the exception of the absence of the "–jaccard_clip" flag due to the low gene density of anaerobic fungal genomes. The assembly process was conducted on the Oklahoma State University High Performance Computing Cluster using a dual Intel Xeon E5-2620 "Sandy Bridge" hex core 2.0 GHz CPU node with 256GB of RAM (https://hpcc.okstate.edu/content/cowboy-overview). Quantitative levels for all assembled transcripts were generated by mapping all generated sequencing reads to the assembled transcripts using the short read alignment mapping program Bowtie2 [42]. The quantitative program RSEM [43] was used to calculate all quantitative values in Fragments Per Kilobase of transcript per Million mapped reads (FPKM). To assess variability between biological replicates, the coefficient of determination R^2 was calculated between biological replicate pairs using RSEM-generated FPKM values. All FPKM values were

normalized to the library size using the R package DESeq [44]. The obtained p-values were used to assess the significance of transcripts' up and down regulation as shown in Tables 2-4 footnotes. All normalized FPKM values shown are averages of two biological replicates. Total normalized FPKM of different GH families when *Orpinomyces* C1A was grown on the different substrates were used in a principal component analysis (PCA) using the R statistical package Labdsv [45] and results were visualized in a biplot.

Transcripts Functional Annotation and CAZyme Identification. Transcripts annotation of all genes was conducted using a combination of homology comparison to public databases, protein domain identification, and peptide secretion signal prediction. Predicted protein sequences from the assembled transcripts were generated using the Transdecoder software portion in the Trinity package [40]. Transcripts that were present in at least one condition with an FPKM ≥ 1 and contained a predicted peptide coding regions were used in subsequent analysis. Predicted peptides were compared to public databases to identify the phylogeny using NCBI Blast C++ [46], where an e-value of e^{-5} or less was used as a cutoff for Blast classification. Signal peptide prediction was conducted using signalP 4.0 [47] using the recommend settings and eukaroytic training set. Protein domain identification [48] was achieved using the hmmscan portion of the HMMER software package [49]. An e-value of e^{-4} was used as a cutoff for significance for domain assignment. All predicted peptide sequences were profiled against the PFAM 27.0 database [48] for general functional domain assignment. To specifically identify peptide sequences that are putative Carbohydrate Active Enzymes (CAZymes), all sequences were profiled against the Database for automated Carbohydrate-active enzyme Annotation (dbCAN) [50]. Sequences identified were further classified through manual curation and structural comparisons. Putative cellulosomal localization of transcripts was identified by the presence of

the CBM_10 (Dockerin) domain that has previously been established as the enzyme attachment component to cellulosome in anaerobic fungi [51].

Differential transcriptional patterns between different conditions were conducted by comparing Log_2 [$FPKM_{biomass}$/ $FPKM_{glucose}$] values. For inter-conditions comparisons, a threshold of log_2 ratio >1, and log_2 ratio <-1 (corresponding to two-fold over-, or under-expression, respectively) was used to designate a specific transcript as significantly over-, or under-expressed, respectively. To study the effect of plant biomass on the cellulosomal composition, we utilized likelihood-ratio-Chi-squared test to examine the significant difference between the relative abundances of various protein categories in the glucose-grown versus plant biomass-grown cultures.

Sequence availability and accession numbers. Raw sequencing reads from each condition and the assembled transcript sequences will be available at GenBank under the accession number SRX1030108 and in MGRAST under the accession number 4667732.3.

Results

RNA-seq output summary. A total of 468,159,494 (70.2 GB) quality-filtered reads were used for transcriptome assembly and quantitative RNA-seq analysis (Table 3.1). The number of reads generated for each growth condition ranged from 58.61 million (8.7 GB) in alfalfa-grown cultures to 141.24 million (21.19 GB) in sorghum-grown cultures (Table 3.1). This level corresponds to 88.73X -201.77X genomic coverage, and 426.73X-1115.07X predicted cDNA coverage. The generated assembly had an N50 of 1,319 bp. A total of 27,506 distinct transcripts with predicted peptides were identified in the assembly.

Strain C1A CAZymes and potential lignocellulolytic capabilities. A total of 385, 246, and 44 distinct transcripts belonging to 44, 13, and 8 different GH, CE, and PL families, respectively, were identified in at least one condition, with the majority being transcribed under all five-growth conditions examined. Collectively the CAZyme transcripts identified demonstrate the capability of strain C1A to degrade cellulose (putative endoglucanases of GH5, GH9, GH45, GH48, GH124; cellobiohydrolases of GH6 and GH48; β-glucosidases of GH1 and GH3); major types of hemicellulose including arabinoxylans/ glucuronoarabinoxylans (putative xylanases of GH10, and GH11; β-xylosidases and α-L-arabinofuranosidase of GH39 and GH43; β-galacotosidases of GH2; α-glucuronidase of GH67 and GH115), glucomannans/galactoglucomannans (putative mannanases and mannosidases of GH26; and β-galactosidases of GH2), mixed glucans (putative β-(1-3, 1-4) endoglucanase of GH16; and β-glucosidases of GH1 and GH3), and xyloglucans (putative xyloglucanases of GH67 and GH74; and α-fucosidase of GH95). In addition to cellulose and hemicellulose, transcripts indicative of the capacity to degrade laminarin (putative 1,3-β-D-endoglucanase of GH55 and GH64; and β-

glucosidases of GH1 and GH3), starch (putative α-amylase of GH13and GH119 and α-amylase/ amylopullulanase of GH57), pectin (putative polygalacturonase of family GH28; endo-β-1,4-galactanase of family GH53; α-L-rhamnosidase of GH78; unsaturated rhamnogalacturonyl hydrolase of GH105; pectate lyases of PL3, PL9, PL10; pectin lyase of PL1; rhamnogalacturonan lyase of PL4 and PL11; and oligogalacturonate lyase of PL22), chitin (putative chitinase of GH18 and GH51), and polygalactosamine (putative endo-α-1,4-polygalactosaminidase of GH114) were also identified

Transcriptional patterns of CAZymes in strain C1A at the family and transcript levels. We analyzed the transcriptional patterns of CAZymes in strain C1A at the family and transcript levels. When grown on glucose, strain C1A constitutively transcribes a relatively high baseline level of CAZyme (GHs, CEs, but not PLs) transcripts that include a wide range of cellulolytic, hemicellulolytic, amylolytic, and accessory enzymes. Indeed, many of the CAZymes families were transcribed at levels comparable to, or even exceeding, those of key glycolytic enzymes such as pyruvate kinase (normalized FPKM of 115.2), Fructose-bisphosphate aldolase (normalized FPKM value 1563.5), and even in few cases (e.g. GH45 endogluconase, GH48 cellobiohydrolase, and GH119 α-amylase), glyceraldehyde-3-P dehydrogenase (normalized FPKM of 2970.4). Growth on lignocellulosic biomass was associated with few distinct changes in transcriptional levels of several GH families (Figure 1, table 3.2). In total, 6 GH families (GH1, GH6, GH18, GH57, GH109, and GH114) were significantly (p-value <0.01) up regulated, while one (GH119) was significantly (p-value <0.01) down regulated across all four lignocellulosic biomass growth conditions. In addition, few families (GH9, GH25, GH55, GH67, and GH124) showed increased (higher normalized FPKM values), or decreased (lower normalized FPKM values, GH4, GH8, GH28, GH30, GH37, GH45, GH97, and GH115)

transcriptional levels in all examined growth conditions, although this change was statistically significant (p-value <0.01) only in some, but not all, growth conditions examined.

Within highly transcribed GH families putatively involved in cellulose degradation (GH1, GH3, GH5, GH6, GH9, GH45, and GH48, defined using a normalized FPKM cutoff value > pyruvate kinase, the glycolytic gene with lowest transcriptional level under all growth conditions, only one putative cellobiohydrolase (GH6) and one putative β-glucosidase (GH1) families were significantly upregulated in all plant biomass conditions compared to glucose. one putative endoglucanase family (GH9) showed higher transcriptional levels on all plant biomass conditions, although this upregulataion was significant (p-value <0.01) only in three (energy cane, corn stover, and sorghum) out of four examined growth conditions. One the other hand GH48 cellobiohydrolases were significantly downregulated in alfalfa and sorghum grown cultures compared to glucose (Figure 3.1).

While few, yet distinct, differential regulation patterns where observed in cellulolytic GH families, no clear family wide up or downregulation patterns were observed in xylanolytic families. Transcriptional levels of the GH10 putative xylanases, GH39 and GH43 putative xylosidases did not show any statistically significant difference when comparing all four plant biomass conditions, compared to glucose (Figure 3.1). Within GH11 xylanases, only significant upregulation was observed only in sorghum grown cultures compared to glucose (Figure 3.1). Collectively, these results suggest that strain C1A constitutively transcribes high level of lignocellulosic enzyme transcripts, even in the absence of lignocellulosic substrates, with growth on lignocellulosic biomass with the substrate elicits few distinct changes in transcriptional patterns of specific GH families (Figure 3.1). This overall pattern of transcriptional change, or

lack thereof, is quite distinct from the scheme utilized by aerobic lingocellulolytic fungi (e.g. *Aspergillus niger* and *Trichoderma reesi* [24, 25]), where growth on lignocellulosic biomass causes a drastic induction of cellulolytic and lignocellulolytic enzymes from low, almost undetectable transcriptional levels on glucose. However, this pattern is broadly similar to transcriptomic response observed in anaerobic lignocellulolytic bacteria (e.g. *Clostridium phytofermentans, C. cellulolyticum, C. thermocellum* [26-28]), which grow and express their CAZymes on glucose as well as lignocellulolytic biomass.

On a single transcript level, 39 (energy cane) to 48 (alfalfa) GH transcripts were significantly ($P<0.01$) upregulated in biomass-grown versus glucose-grown cultures, while a broadly comparable numbers of transcripts (53 sorghum - 66 corn stover) were significantly downregulated. The majority of transcripts (192 in corn stover and energy cane, and 210 in alfalfa and sorghum), however, did not show a significant change in transcription levels ($P>0.1$). A similar pattern was also observed for CE and PL families as well.

We also correlated transcriptional levels of various GH families with the composition (cellulose and hemicellulose content) of plant materials examined as growth substrate in this study. Transcriptional levels of some cellulolytic CAZyme families e.g. GH5, GH6, GH9, GH48, and GH124 were positively correlated (Pearson correlation coefficients of 0.42, 0.81, 0.71, 0.58, and 0.62) with the substrates' cellulose content (i.e. overall normalized FPKM of the family was higher in plants with higher cellulose content). However, no such correlation was observed for GH8 or GH45 (Pearson correlation coefficients of 0.06, -0.36, respectively). On the other hand, no clear correlation was observed between transcriptional levels of xylanase CAZyme families (GH10 and GH11) and hemicellulose content (Pearson correlation coefficient of -0.32 and -0.19,

respectively). GH39 xylosidase showed a positive correlation with hemicellulose content (Pearson correlation coefficient of 0.60), while GH43 xylosidase showed a strong negative correlation with hemicellulose content (Pearson correlation coefficient of -0.93).

Strain C1A employs multiple functionally redundant but structurally and mechanistically distinct processes for biomass degradation. To examine the relative contribution of various CAZyme families to biomass degradation under different growth conditions, we quantified the relative transcriptional levels of families putatively mediating the deconstruction of various plant polymers as a fraction of an overall specific activity. Our results (Figure 3.2) demonstrate that strain C1A co-transcribes multiple functionally redundant enzymes (i.e. mediating the exact same chemical reaction and targeting the same substrate) that are, nevertheless, mechanistically and structurally distinct. While the identification of many of these genes in anaerobic fungi has been previously documented [22, 29], their differential transcriptional patterns and relative contribution to biomass degradation under various growth conditions have not been previously studied. For example, transcripts of putative endoglucanases belonging to five distinct families were identified, three of which (the $(\alpha/\beta)_8$ TIM barrel retaining GH5, the $(\alpha/\alpha)_6$ barrel inverting GH9, and the β barrel inverting GH45) represented >15% of overall endoglucanases under all growth conditions (Figure 3.2A). A similar high level of co-transcription of the inverting α/β barrel GH6 putative cellobiohydrolase acting on the non-reducing end of cellulose molecules and the retaining $(\alpha/\beta)_8$ TIM barrel putative cellobiohydrolase acting on the reducing end of the cellulose molecule were observed (Figure 3.2B). Finally, a high co-transcriptional level of GH1 and GH3 putative β-glucosidases were also observed (Figure 3.2C). Within putative xylanolytic enzymes, a similar phenomenon is observed between the retaining $(\alpha/\beta)_8$ TIM barrel GH10 putative xylanase and the retaining β-jelly roll GH11 (Figure 3.2D), and the same dynamic was

observed putative xylosidases (GH39 and GH43) mediating depolymerization of xylooligomers (Figure 3.2E).

Interestingly, distinct shifts in the relative transcripts abundances of GH families as a fraction of an overall specific activity were frequently observed (Figure 3.2). Within glucose-grown cultures, the majority of putative endoglucanases belonged to GH45 (65% of putative endoglucanases normalized FPKM in glucose grown cultures). However, when grown on plant biomass, the relative abundance of GH45 decreased, with a concomitant increase in the relative abundance of GH9 putative endoglucanases (Figure 3.3A). Similarly, growth on plant biomass was invariably associated with an increase in the relative contribution of GH6 and a reciprocal decrease in the relative contribution of GH48 to the overall cellobiohydrolase activity (Figure 3.3B).

A limited number of lignocellulolytic transcripts are highly transcribed under all growth conditions. Within a single CAZyme gene family, often a large number of distinct transcripts were identified, and this was especially true for families with a high overall transcriptional activity. Indeed, a broad positive correlation between the total FPKM level of a specific GH family and the number of transcripts identified belonging to this family was observed. To further zoom in on the putative variations in the contribution of specific transcripts belonging to a certain GH family to biomass degradation, we examined the transcriptional levels of all individual transcripts within key GH families. Out of the large number of transcripts identified in each family, a fairly limited (1-6) number of transcripts were dominant (i.e. represent >10% of the total normalized family FPKM under at least one growth condition) in all instances (Figure 3.4). Transcriptional patterns of dominant transcripts under different growth conditions varied

across different CAZyme families. In some families (e.g. GH6, GH13, and GH39) a single transcript represented the majority (>60%) of all FPKM levels across all growth conditions. In other instances, few (2-3) transcripts consistently represented the majority of family transcripts, with their relative abundance patterns remaining fairly stable across various growth conditions (e.g. GH18, GH43, and GH57). Within the remaining families, a significant shift in the relative transcriptional level, and hence putative contribution, was observed between different growth conditions. For example, specific transcripts in GH5 (m.22928), GH13 (m.23494), GH43 (m.5510), GH45 (m.23474), and GH48 (m.19942) appear to be highly transcribed in glucose-grown cultures, but their relative importance diminishes in lignocellulosic biomass-grown cultures. Conversely, some transcripts appear to be prominent and differentially upregulated in lignocellulosic biomass-grown cultures, while their contribution to the overall activity dwindles in glucose-grown cultures (e.g. m.17949, and m.17964 in GH9, m.20865 in GH10, m.21149 in GH11, and m.23473 in GH45). Collectively, the results demonstrate that while some families show differential transcriptional patterns in response to growth conditions, a few stable "core" of transcripts, especially within highly transcribed CAZyme families in strain C1A, appears to be consistently predominant.

Fungal dockerin domain (FDD)-containing transcripts. Anaerobic fungi produce cellulosomes, surface-attached structures where multiple enzymes act synergistically towards the degradation of lignocellulosic biomass. As previously described, cellulosome-bound genes in anaerobic fungi usually harbor a fungal dockerin domain (FDD) that is similar in structure to carbohydrate-binding module family 10 (CBM10) [12]. By determining FDD occurrence in all transcripts, a total of 278, 283, 292, 288, and 291 were putatively identified as "cellulosomal-bound transcripts" in glucose-, alfalfa-, energy cane-, corn stover-, and sorghum-grown C1A

cultures, respectively, with the absolute majority of transcripts identified in all examined growth conditions. Cellulosomal-bound transcripts were affiliated with 4 broad major categories (Figure 3.5): biomass-degrading CAZymes and accessory enzymes; hypothetical and conserved hypothetical proteins; proteases, phosphohydrolases and protease inhibitors (serpins); and the enigmatic CotH family protein transcripts previously observed in fungal and bacterial cellulosomes and previously implicated as a structural component of the cellulosome [30]. Analysis of the transcriptional patterns of FDD transcripts under different growth conditions indicated that the relative contribution of the four major categories described above to the overall cellulosomal composition did not vary significantly when C1A was grown on glucose versus plant biomass (likelihood ratio $\chi 2=59.88$, p-value=0.055).

Examination of FDD CAZyme and accessory transcripts suggests the involvement of the cellulosome in all stages of cellulose (putative endoglucanases, cellobiohydrolases, and β-glucosidases), arabinoglucoxylan (putative xylanases, xylosidases, arabinofuranosidases, acetylxylan esterase, and feruoyl esterases), xyloglucan (xyloglucanases), and glucomannan (putative mannanases/mannosidases) degradation. Within a specific GH family, the relative contributions of FDD transcripts to the overall family transcriptional level varied. Based on number of transcripts and transcriptional activity, FDD transcripts represent the absolute majority of transcriptional activity in GH48 putative cellobiohyrolases, the majority in GH5 putative endoglucanases, roughly half the transcriptional activity in GH9 putative endoglucanases, GH10 putative xylanases, and GH43 putative β-xylosidases, and a small fraction of the transcriptional activities of GH11 putative xylanases and GH45 putative endoglucanases. Interestingly, overall expression of GH and accessory enzymes transcripts were significantly down regulated in three (alfalalfa, energy cane, and sorghum) growth conditions,

mainly due to the significant downregulation of GH48, a major component of the cellulosome, under these growth conditions. Other notable contributions of the putatively cellulosomal-bound, FDD harboring transcripts to biomass degradation include the prevalence of carbohydrate esterases (3.5-5.7% of overall FDD transcripts, depending on the growth condition), and feruloyl esterases (up to 3.8% of overall FDD transcripts) within all FDD-harboring transcripts (Figure 3.5), suggesting an important role of the cellulosome in the mobilization and debranching of hemicellulose backbones. In addition to CAZyme families responsible for cell wall decomposition, an important accessory transcript belonging to the swollenin/expansin enzyme family was identified as cellulosomal-bound. This enzyme family enables plant cell lengthening through non catalytic disruption of hydrogen bonds in plant cell walls [31]. Homologs of this enzyme family have also been shown to enhance cell wall decomposition when utilized by microorganisms [32]. Out of the five swollenin/expansin transcripts identified, four contained an FDD, and they represented 89-97% of total normalized FPKM activity, depending on the growth condition, of total swollenin transcripts identified in C1A transcriptome. Although swollenin and GH45 are structurally related [33], the predominantly cellulosomal transcriptional pattern of the non-enzymatic swollenin is in contrast to that observed mostly free extracellular patterns for GH45 transcripts. The predominance of this non-catalytic homolog in the cellulosome emphasizes their important role in cell wall weakening as an additional mechanism to enhance plant biomass degradation efficiency by cellulosomal catalytic enzymes.

Discussion

In this study, we analyzed transcriptional patterns in strain C1A when grown on plant biomass as well as soluble (glucose) substrates. Collectively, our results suggest that strain C1A constitutively transcribes a wide array of FDD containing (i.e. putatively cellulosomal-bound) and free extracellular lignocellulolytic enzymes under all examined conditions. The results also highlight the simultaneous involvement of multiple functionally redundant CAZymes in plant biomass degradation, arguably as a tool to improve the speed and extent of biomass degradation by anaerobic fungi within its natural habitat (the herbivorous gut). Finally, the results provide an in-depth evaluation of the contribution of free versus FDD-containing (i.e. putatively cellulosomal-bound) enzymes in biomass degradation in strain C1A.

Our results demonstrate that strain C1A constitutively transcribes a wide array of transcripts encoding lignocellulolytic enzymes. Microorganisms growing on lignocellulosic biomass invariably spend a large fraction of their carbon and energy reserves on the synthesis and export of lignocellulolytic enzymes (CAZymes). Therefore, regulation of the biosynthesis of such enzymes is key for optimal ecological fitness and resource allocation. Within model lignocellulolytic aerobic fungi, e.g. *Aspergillus niger* and *Trichoderma reesi*, growth on lignocellulosic biomass causes a drastic induction of cellulolytic and lignocellulolytic enzymes from almost undetectable transcriptional levels on glucose-grown cultures, to ≈ 12-20% of the overall mRNA [24, 25]. This induction pattern is associated with a drastic change in the relative composition of the CAZyme transcriptome from a glucoamylase-dominated profile when grown on glucose or other soluble substrate, to an endoglucanase-, cellobiohydrolase-, xylanase-, arabinofuranosidase-, acetyl-xylan esterase-, and polysaccharide monooxygenase-dominated

profile when grown on lignocellulosic biomass [24, 25]. On the other hand, multiple anaerobic prokaryotes (e.g. *Clostridium cellulolyticum, C. phytofermentans, and C. thermocellum*) possess constitutively expressed CAZymes and a high overall transcriptional levels of lingocellulolytic enzymes is observed in glucose-grown cultures [26-28]. Indeed, it is postulated that glucose sensing appears to act as a priming mechanism that stimulates a wide range of CAZymes biosynthesis [26-28]. Our results suggest that anaerobic fungi employ a model similar to anaerobic bacteria as opposed to aerobic fungi. This conclusion is in accordance with our understanding of the ecological niche and life cycle of anaerobic fungi within its restricted habitat in the herbivorous gut. In such an environment, the life cycle of anaerobic fungi alternates between metabolically dormant spores, and hyphae germinating from spores when ingested plant biomass is encountered in the gut. Fungal germination and growth is hence invariably linked to the availability of ingested plant biomass. Therefore, spore germination, hyphal growth, and production of lignocellulolytic enzymes in anaerobic fungi are tightly linked, and it is inconceivable to envision a situation in which anaerobic fungi grow solely on a soluble substrate within their natural habitat. Therefore we argue that, due to their ecological niche, their role as initial colonizers of plant biomass, and their sole dependence on plant biomass as a substrate within their natural habitat, the need for development of sophisticated mechanisms for regulating the expression of CAZyme genes is non-existent in anaerobic fungi. This is drastically different from what is encountered by aerobic lignocellulolytic fungi in their natural environments, where gradients in environmental conditions (temperature, pH, moisture), substrate availability (by season) and type (plant biomass versus sugars), and the relatively large residence time and degradation rates necessitate development of regulatory processes for enzymatic biosynthesis. Nevertheless, despite this constitutive pattern of CAZyme genes

transcription in anaerobic fungi it appears that growth on plant biomass triggers a distinct response in CAZyme GH families and individual transcripts (Figure 3.2, Figure 3.3). The rationale behind these family and transcript level shifts, observed mainly within GH families and transcripts involved mainly in various aspects of cellulose degradation remains unclear.

Another interesting characteristic in lignocellulosic biomass degradation by strain C1A is the simultaneous engagement of a large number of functionally redundant enzymes in the degradation of a single polymer (e.g. cellulose or arabinoxylan). We argue that this strategy is employed by C1A to increase the efficacy and speed of the degradation process, and hence maximize the extent of plant biomass degraded within its relatively short residence time in the herbivorous gut. Further, the complementary nature of this strategy is further accentuated by variations in the location of the enzymes (cellulosomal versus free extracellular), the nature of the substrate targeted (chain length and side chains preferences), and the target position (e.g. reducing versus non reducing end) within the substrate (Figure 6). Transcripts encoding most enzymatic activities required for the degradation of cellulose and hemicellulose are well represented in both putatively cellulosomal and non-cellulosomal fractions, allowing for the simultaneous degradation of these polymers at two distinct locations (Figure 3.6). Strain C1A simultaneously transcribes high levels of GH10 and GH11 family transcripts. GH10 enzymes are known to have broader substrate specificity, with the capability to attack xylan backbones with a high degree of substitutions and smaller xylo-oligosaccharides [34]. Therefore, such a pattern of high co-transcription allows for the instant and sustained breakdown of xylan backbone polymer regardless of their length and progress in side chain removal by accessory enzymes. Finally, the co-transcription of GH6 and GH48 cellobiohydrolases by C1A allows for the simultaneous

targeting of both reducing and reducing ends of cellulose and cellooligosaccharides in plant biomass to improve speed and efficiency of cellulose degradation.

Third, our results highlight the importance of anaerobic fungal cellulosomes for biomass degradation. While broad upregulation in FDD transcripts were observed in plant-biomass grown versus glucose-grown cultures, no drastic changes in membership (presence/absence) of specific transcripts, or composition (relative levels of specific transcripts) were observed (Figure 3.5). The results suggest that cellulosomal structure does not vary considerably depending on the growth substrate, as previously suggested. Further, FDD transcripts identified strongly suggest that cellulosomes represent distinct catalytic units capable of independently attacking and converting intact plant fibers to sugar monomers. A large number of highly transcribed transcripts are involved in the initial disruption of plant fiber architecture through non-catalytic hydrolysis of hydrogen bonds (swollenin), mobilization of target plant polymers (feruloyl esterases), side chains removal (acetyl xylan esterase, polysaccharide deacetylase), and plant polymers degradation to sugar monomers (endoglucanases, cellobiohydrolases, β-glucosidases; xylanases and xylosidases).

In conclusion, our work demonstrates that strain C1A constitutively transcribes a wide array of lignocellulolytic enzymes under different growth conditions. Although many of these enzymes are functionally redundant, differences in location (cellulosomal versus free extracellular), substrate preference (polymer length and substitution patterns), and target position within the substrate (e.g. reducing versus non reducing end) allow for fast efficient utilization of target substrates in the relatively short time frame of availability within the herbivorous gut. The utilization of this indiscriminate strategy as an ecological and evolutionary necessity, as well as

capability of anaerobic fungi to utilize a broad range of plant biomass including lignocellulosic biomass substrates, render anaerobic fungi appealing, yet understudied, candidates for utilization in biomass conversion to sugars and biofuels.

Figure 3-1. Principal component analysis (PCA) of normalized GH families transcription levels. Normalized FPKM values of GH families under different growth conditions were used as input. Stars depict growth substrates and arrows represent GH families. Growth substrates with similar transcriptional profiles are closer together in the ordination plot than substrates with different transcriptional profile. The direction of the GH families arrows in the biplot are indicative of the respective maximal transcription, while the lengths of the arrows are indicative of the differential transcription.

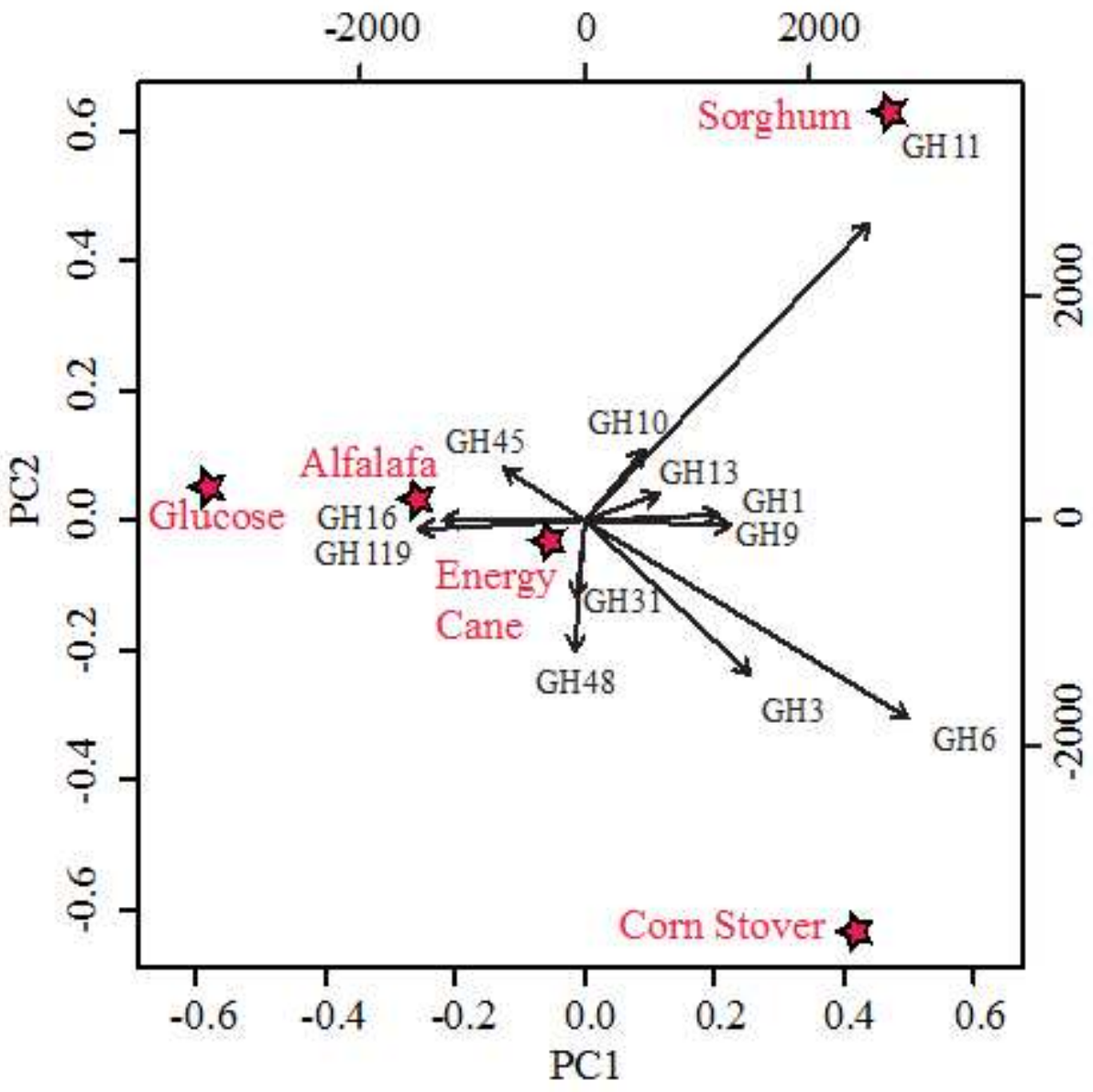

-2000
0
2000
Sorghum
GH11
GH10
GH45
Alfalafa
GH13
GH1
Glucose
GH16
GH9
GH119
Energy
Cane
GH31
GH48
GH3
GH6
Corn Stover
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
PC2
2000
0
-2000
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
PC1

Figure 3-2. Relative contribution of various GH families putatively mediating key enzymatic activities required for cellulose and xylan degradation under different growth conditions.

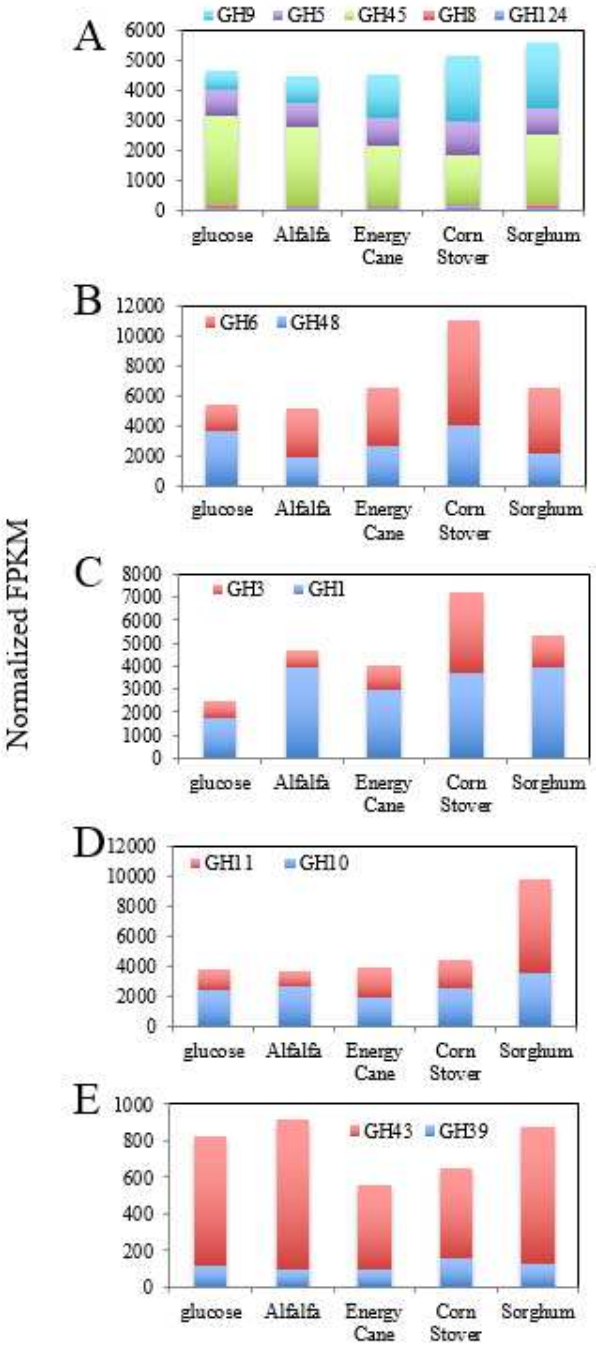

A
GH9
GH5
GH45
GH8
GH124
6000
5000
4000
3000
2000
1000
0
glucose
Alfalfa
Energy Cane
Corn Stover
Sorghum
B
GH6
GH48
12000
10000
8000
6000
4000
2000
0
C
GH3
GH1
8000
7000
6000
5000
4000
3000
2000
1000
0
D
GH11
GH10
12000
10000
8000
6000
4000
2000
0
E
GH43
GH39
1000
800
600
400
200
0
Normalized FPKM

Figure 3-3. Relative contribution of dominant transcripts within various GH families under different growth conditions. Only families with overall transcriptional level under all growth conditions above 1% that of a suite of glycolytic genes (pyruvate kinase, glyceraldehyde-3-phosphate dehydrogenase, and fructose-1,6-bisphosphate aldolase) were studied. Within these, genes were selected that represented 10% or more of the overall moralized FPKM under any growth condition. “Others” denotes all additional transcripts that never exceeded >10% of overall moralized FPKM under any growth condition.

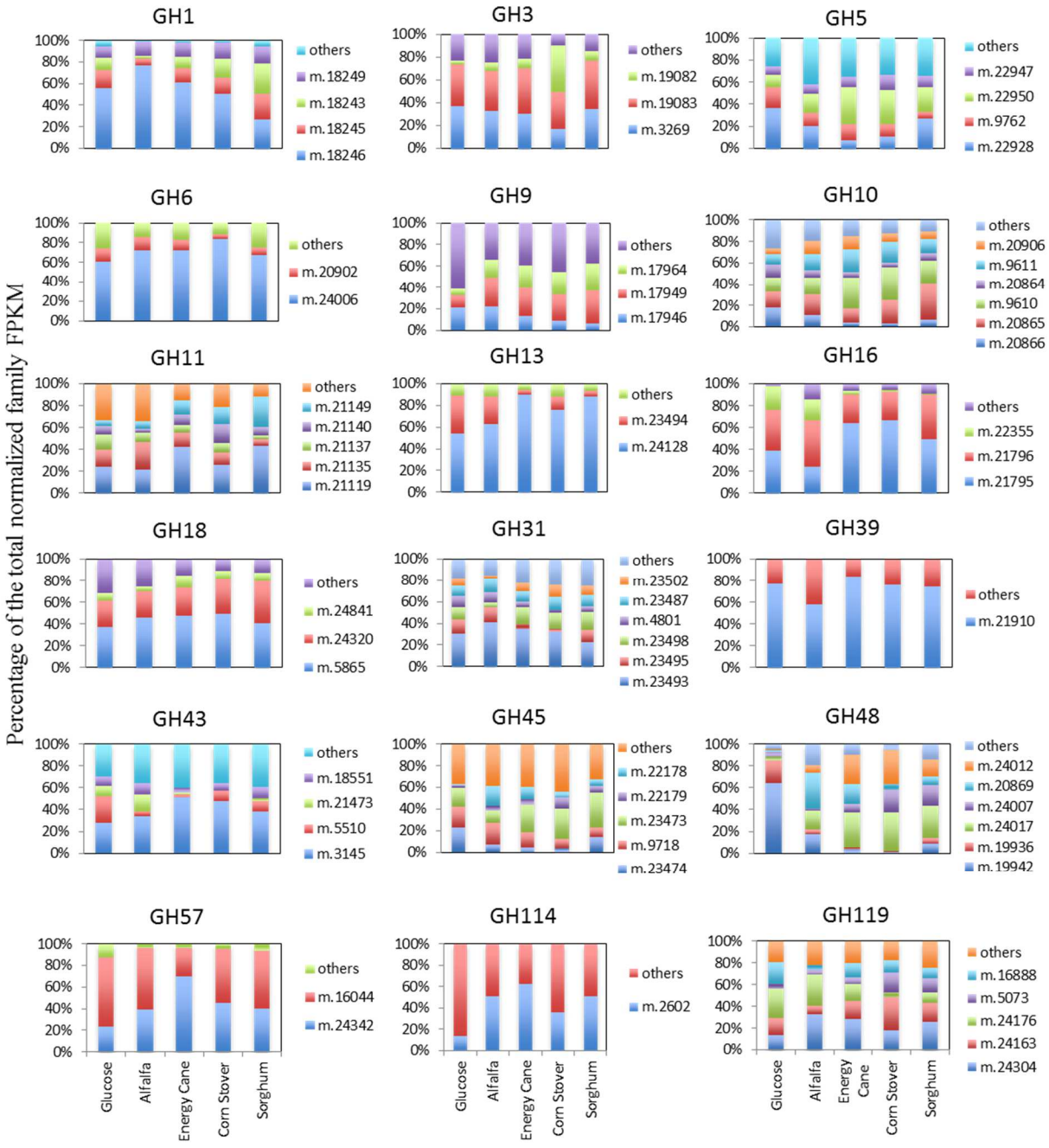

Percentage of the total normalized family FPKM
GH1
others
m.18249
m.18243
m.18245
m.18246
GH3
others
m.19082
m.19083
m.3269
GH5
others
m.22947
m.22950
m.9762
m.22928
GH6
others
m.20902
m.24006
GH9
others
m.17964
m.17949
m.17946
GH10
others
m.20906
m.9611
m.20864
m.9610
m.20865
m.20866
GH11
others
m.21149
m.21140
m.21137
m.21135
m.21119
GH13
others
m.23494
m.24128
GH16
others
m.22355
m.21796
m.21795
GH18
others
m.24841
m.24320
m.5865
GH31
others
m.23502
m.23487
m.4801
m.23498
m.23495
m.23493
GH39
others
m.21910
GH43
others
m.18551
m.21473
m.5510
m.3145
GH45
others
m.22178
m.22179
m.23473
m.9718
m.23474
GH48
others
m.24012
m.20869
m.24007
m.24017
m.19936
m.19942
GH57
others
m.16044
m.24342
GH114
others
m.2602
GH119
others
m.16888
m.5073
m.24176
m.24163
m.24304
100%
80%
60%
40%
20%
0%
Glucose
Alfalfa
Energy Cane
Corn Stover
Sorghum

Figure 3-4. FDD-containing, putatively cellulosomal transcripts. Each square depicts transcriptional patterns under a specific growth condition as shown above the squares. The size of each square, and sections within, is proportional to the transcriptional level (normalized FPKM values). The sections are color coded by their predicted activity as follows: Green, GH families; Dark blue, swollenin/ expansins accessory enzymes; Pink, acetylxylan esterases, carboxyl esterases (CE), and feruloyl esterases (FE); Black, hypothetical proteins; Purple, conserved hypothetical proteins; Brown, protease inhibitors (serpins); Dark brown, serine and threonine protein phosphatases (P-ases), alkyl transferases (AT), proteases, and rhamnogalacturonases (RG); Light blue, dual activity enzymes including polysaccharide deacetylases (PD)/GH5 proteins, feruloyl esterases (FE)/GH45 proteins, and carboxyl esterases (CE)/GH9 proteins.

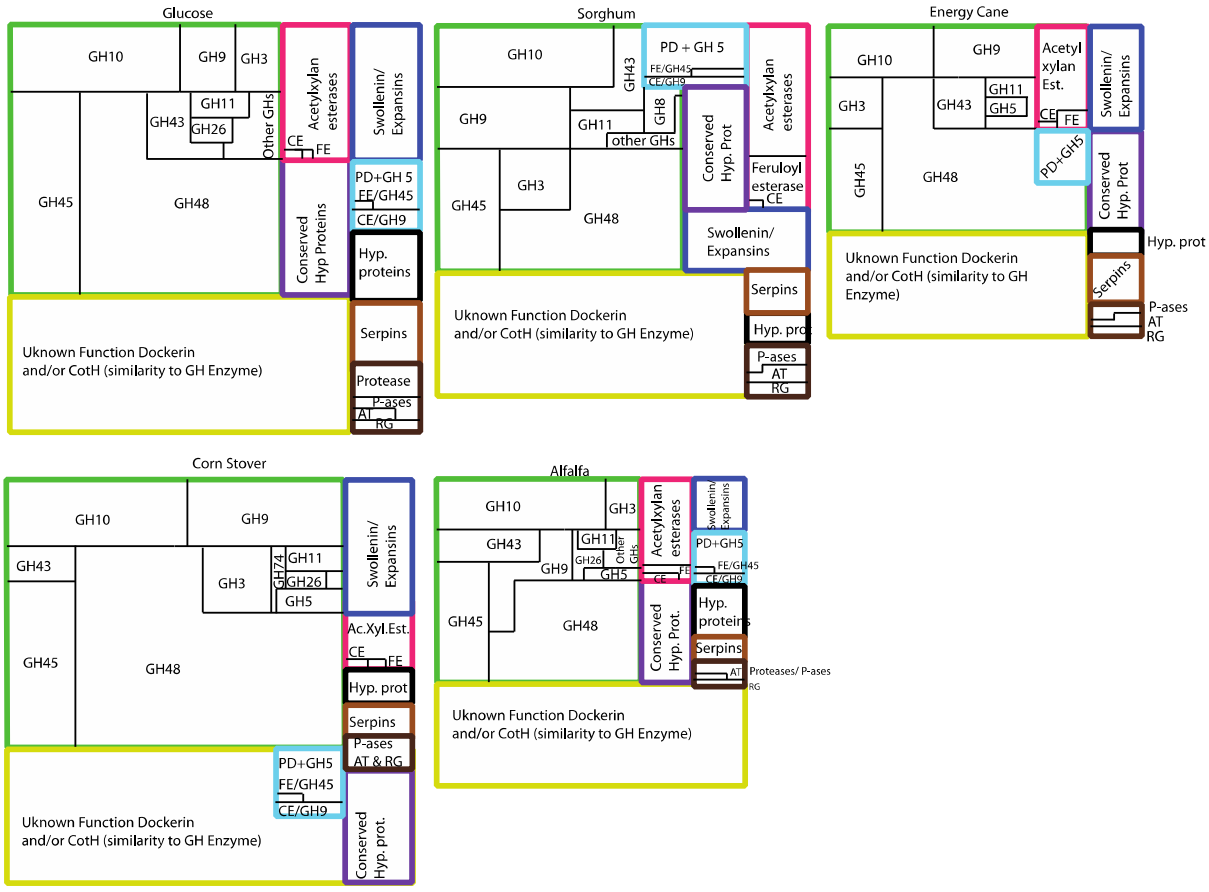

Glucose
GH10
GH9
GH3
GH11
GH43
GH26
Other GHs
GH45
GH48
Acetylxylan esterases
CE
FE
Swollenin/ Expansins
PD+GH 5
FE/GH45
CE/GH9
Conserved Hyp Proteins
Hyp. proteins
Serpins
Protease
P-ases
AT
RG
Uknown Function Dockerin and/or CotH (similarity to GH Enzyme)
Sorghum
GH10
GH43
PD + GH 5
FE/GH45
CE/GH9
GH9
GH11
GH8
other GHs
Acetylxylan esterases
GH45
GH3
GH48
Conserved Hyp. Prot
Feruloyl esterase
CE
Swollenin/ Expansins
Serpins
Hyp. prot
P-ases
AT
RG
Uknown Function Dockerin and/or CotH (similarity to GH Enzyme)
Energy Cane
GH10
GH9
Acetyl xylan Est.
GH3
GH11
GH43
GH5
CE
FE
Swollenin/ Expansins
GH45
GH48
PD+GH5
Conserved Hyp. Prot
Hyp. prot
Serpins
P-ases
AT
RG
Uknown Function Dockerin and/or CotH (similarity to GH Enzyme)
Corn Stover
GH10
GH9
Swollenin/ Expansins
GH43
GH74
GH11
GH3
GH26
GH5
Ac.Xyl.Est.
CE
FE
GH45
GH48
Hyp. prot
Serpins
P-ases
AT & RG
PD+GH5
FE/GH45
CE/GH9
Uknown Function Dockerin and/or CotH (similarity to GH Enzyme)
Conserved Hyp. prot.
Alfalfa
GH10
GH3
Acetylxylan esterases
Swollenin/ Expansins
GH43
GH11
Other GHs
PD+GH5
GH9
GH26
GH5
FE
FE/GH45
CE
CE/GH9
GH45
GH48
Conserved Hyp. Prot.
Hyp. proteins
Serpins
AT
Proteases/ P-ases
RG
Uknown Function Dockerin and/or CotH (similarity to GH Enzyme)

Table 3.1. **General statistics of RNA-seq output.**

Condition	Total Reads	Total Bases	Genome Coverage[1]	cDNA Coverage[2]	Assembled Transcript Coverage[3]	R^2 Value[3]
Glucose	81,468,482	12,220,272,300	121.59	590.27	349.15	0.89
Alfalfa	58,612,544	8,791,881,600	87.48	424.67	251.20	0.99
Energy Cane	93,381,914	14,007,287,100	139.38	676.58	400.21	0.99
Corn stover	100,842,114	15,126,317,100	150.51	730.63	432.18	0.99
Sorghum	141,241,616	21,186,242,400	210.81	1,023.34	605.32	0.99
Total Reads/Coverage	468,159,494	70,223,924,100	698.75	3,391.97	2,038.05	0.99

[1]Genome coverage based on an estimated 100.5 Mb genome size [18].
[2]cDNA coverage is based on a 20.76% genome coding density [18].
[3]Assembled Transcript Coverage is based on the total assembled transcript size (35.0 MB).

References

1. **Sanderson K. Lignocellulose: A chewy problem.** Nature. 2011;474:S12-4.

2. **Schubert C. Can biofuels finally take center stage?** Nature Biotechnol. 2006;24:777-84.

3. **National Research Council. Renewable Fuel Standard: Potential Economic and Environmental Effects of U.S. Biofuel Policy.** Washington, DC 2011.

4**. Alvira P, Tomas-Pejo E, Ballesteros M, Negro MJ. Pretreatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: A review.** Bioresource Technol. 2010;101: 4851-61.

5. **Nichols NN, Monceaux DA, Dien BS, Bothast RJ. Production of ethanol from corn and sugarcane. In: Wall JD, Harwood CS, Demain A, editors.** Bioenergy. Washington, D. C: ASM Press; 2008. p. 3-15.

6. **Scheller HV, Ulvskov P. Hemicelluloses.** Annual Rev Plant Biol. 2010;61:263-89.

7. **Balan V, Bals B, Chundawat SP, Marshall D, Dale BE. Lignocellulosic biomass pretreatment using AFEX**. Methods Mol Biol. 2009;58:61-77.

8. **da Costa Lopesa AM, Joãoa KG, Rubika DF, Bogel-Łukasikc E, Duartea LC, Andreausb J et al. Pre-treatment of lignocellulosic biomass using ionic liquids: Wheat straw fractionation**. Bioresource Technol. 2013;142:198-208.

9. **Minty JJ, Singer ME, Scholz SA, Bae C-H, Ahn J-H, Foster CE et al. Design and characterization of synthetic fungal-bacterial consortia for direct production of isobutanol from cellulosic biomass.** Proc Natl Acad Sci USA. 2013;110:14592-7.

10. **Peralta-Yahya PP, Zhang F, del Cardayre SB, Keasling JD. Microbial engineering for the production of advanced biofuels.** Nature. 2012;488:320-8.

11**. Hu ZH, Liu SY, Yue ZB, Yan LF, Yang MT, Yu HQ. Microscale analysis of in vitro anaerobic degradation of lignocellulosic wastes by rumen microorganisms.** Environ Sci Technol. 2008;42:276-81.

12. **Ljungdahl LG. The cellulase/hemicellulase system of the anaerobic fungus *Orpinomyces* PC-2 and aspects of its use.** Ann N Y Acad Sci. 2008;1125:308-21.

13. **Chung D, Cha M, Guss AM, Westpheling J. Direct conversion of plant biomass to ethanol by engineered *Caldicellulosiruptor bescii*.** Proc Natl Acad Sci USA. 2014;111:8931-6.

14. **Akinosho H, Yee K, Close D, Ragauskas A. The emergence of *Clostridium thermocellum* as a high utility candidate for consolidated bioprocessing applications.** Front Chem. 2014;2:66.

15. **Orpin CG. Anaerobic fungi: Taxonomy, biology, and distribution in nature.** Anaerobic fungi: biology, ecology, and function. 1994.

16. **Gruninger RJ, Puniyab AK, Callaghanc TM, Edwardsc JE, Youssef N, Dagare SS et al. Anaerobic Fungi (Phylum Neocallimastigomycota): Advances in understanding of their taxonomy, life cycle, ecology, role, and biotechnological potential**. FEMS Microbiol Ecol. 2014;90:1-17.

17. **Trinci APG, Davies DR, Gull K, Lawrence MI, Nielsen BB, Rickers A et al.** Anaerobic fungi in herbivorous animals. Mycol Res. 1994;98:129-52.

18. **Grenet E, Barry P. Colonization of thick walled plant tisues by anaerobic fungi.** Anim Fee Sci Technol. 1988;19:25-31.

19. **Joblin KN, Matsui H, Naylor GE, Ushida K. Degradation of Fresh Ryegrass by methanogenic co-cultures of ruminal fungi in the presence or abscence of *Fibrobacter succinogenes*.** Curr Microbiol. 2002;45:46-53.

20**. Joblin KN, Naylor GE. Fermentation of woods by rumen anaerobic fungi.** FEMS Microbiol Lett. 1989;65:119-22.

21**. Theodorou MK, Longland AC, Dhanoa MS, Lowe SE, Trinci APJ. Growth of *Neocallimastix* sp. strain R1 on Italian Ryegrass Hay: removal of neutral sugars from plant cell walls.** Appl Environ Microbiol. 1989;55:1363-7.

22**. Youssef NH, Couger MB, Struchtemeyer CG, Liggenstoffer AS, Prade RA, Najar FZ et al. Genome of the anaerobic fungus *Orpinomyces* sp. C1A reveals the unique evolutionary history of a remarkable plant biomass degrader.** Appl Environ Microbiol 2013;79:4620-34.

23**. Liggenstoffer AS, Youssed NH, Wilkins MR, Elshahed MS. Evaluating the utility of hydrothermolysis pretreatment approaches in enhancing lignocellulosic biomass degradation by the anaerobic fungus *Orpinomyces* sp. strain C1A.** J Microbiol Meth. 2014;104:43-48.

24. **Ries L, Pullan ST, Delmas S, Malla S, Blythe MJ, Archer DB. Genome-wide transcriptional response of *Trichoderma reesei* to lignocellulose using RNA sequencing and comparison with *Aspergillus niger*.** BMC Genomics. 2013;14:541.

25. **Delmas S, Pullan ST, Gaddipati S, Kokolski M, Malla S, Blythe MJ et al. Uncovering the genome-wide transcriptional responses of the filamentous fungus *Aspergillus niger* to lignocellulose using RNA sequencing.** PLoS Genet. 2012;8:e1002875.

26. **Boutard M, Cerisy T, Nogue P-Y, Alberti A, Weissenbac J, Salanoubat M et al. Functional diversity of carbohydrate-active enzymes enabling a bacterium to ferment plant biomass.** PLoS Genet. 2014;10:e1004773.

27. **Wilson CM, Rodriguez M, Johnson CM, Martin SL, Chu TM, Wolfinger RD et al. Global transcriptome analysis of *Clostridium thermocellum* ATCC 27405 during growth on dilute acid pretreated Populus and switchgrass.** Biotechnol Biofuels. 2013;6:179.

28. **Xu C, Huang R, Teng L, Wang D, Hemme CL, Borovok I et al. Structure and regulation of the cellulose degradome in *Clostridium cellulolyticum*.** Biotechnol Biofuels. 2013;6:73.

29. **Wang T-Y, Chen H-L, Lu M-YJ, Chen Y-C, Sung H-M, Mao C-T et al. Functional characterization of cellulases identified from the cow rumen fungus *Neocallimastix patriciarum* W5 by transcriptomic and secretomic analyses.** Biotechnol Biofuels. 2011;4:24.

30. **Zverlov VV, Velikodvorskaya GA, Schwarz WH. Two new cellulosome components encoded downstream of celI in the genome of *Clostridium thermocellum*: the non-processive endoglucanase CelN and the possibly structural protein CseP.** Microbiology. 2003;149:515-24.

31**. McQueen-Mason S, Cosgrove DJ. Disruption of hydrogen bonding between plant cell wall polymers by proteins that induce wall extension.** Proc Natl Acad Sci USA. 1994;91(14):6574-8.

32**. Wang Y, Tang R, Tao J, Gao G, Wang X, Mu Y et al. Quantitative investigation of non-hydrolytic disruptive activity on crystalline cellulose and application to recombinant swollenin.** Appl Microbiol Biotechnol. 2011;91:1353-63.

33**. Cosgrove DJ. Growth of the plant cell wall.** Nat Rev Mol Cell Biol. 2005;6:850-61.

34. **Pollet A, Delcour JA, Courtin CM. Structural determinants of the substrate specificities of xylanases from different glycoside hydrolase families.** Crit Rev Biotechnol. 2010;30:176-91.

35. **Struchtemeyer CG, Couger MB, Ranganathan A, Liggenstoffer AS, Youssef NH, Elshahed MS. Survival of the anaerobic fungus *Orpinomyces* sp. strain C1A after prolonged air exposure.** Sci Reports. 2014;4:6892.

36. **Balch WE, Wolfe R. New approach to the cultivation of methanogenic bacteria: 2-mercaptoethanesulfonic acid (HS-CoM)-dependent growth of *Methanobacterium ruminantium* in a pressureized atmosphere.** Appl Environ Microbiol 1976;32:781-91.

37. **Bryant M. Commentary on the Hungate technique for culture of anaerobic bacteria**. Am J Clin Nutr. 1972.;25:1324-8.

38. **Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics.** Nat Rev Genet. 2009;10:57-63.

39. **Martin JA, Wang Z. Next-generation transcriptome assembly.** Nat Rev Genet. 2011;12:671-82.

40. **Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome.** Nat Biotechnol. 2011;29:644-52.

41. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J et al**. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.** Nat Protoc. 2013;8:1494-512.

42**. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2.** Nat Methods. 2012;9:357-9.

43. **Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** BMC Bioinformatics. 2011;12:323.

44. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K et al. BLAST+: architecture and applications.** BMC Bioinformatics. 2009;10:421.

45**. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions.** Nat Methods. 2011;8:785-6.

46. **Eddy SR. Accelerated Profile HMM Searches.** PLoS Comput Biol. 2011;7:e1002195.

47. **Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure.** BMC Bioinformatics. 2010;11:431.

48. **Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation.** Nucleic Acids Res. 2012;40:W445-51.

49**. Raghothama S, Eberhardt RY, Simpson P, Wigelsworth D, White P, Hazlewood GP et al. Characterization of a cellulosome dockerin domain from the anaerobic fungus *Piromyces equi*.** Nat Struct Biol. 2001;8:775-8.