Diversity within the genus *Thermoanaerobacter* and its potential implications in lignocellulosic biofuel production through consolidated bioprocessing

by

Tobin James Verbeke

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Department of Microbiology

University of Manitoba

Winnipeg, Manitoba

Canada

Abstract

A major obstacle to achieving commercially viable lignocellulosic biofuels through consolidated bioprocessing (CBP) is the lack of "industry-ready" microorganisms. Ideally, a CBP-relevant organism would achieve efficient and complete hydrolysis of lignocellulose, simultaneous utilization of the diverse hydrolysis products and high yields of the desired biofuel. To date, no single microbe has been identified that can perform all of these processes at industrially significant levels.

As such, thermophilic decaying woodchip compost was investigated as a source of novel lignocellulolytic, biofuel producing bacteria. From a single sample, a collection of physiologically diverse strains were isolated, which displayed differences in substrate utilization and biofuel production capabilities. Molecular characterization of these isolates, and development of a genome relatedness prediction model based on the chaperonin-60 universal target sequence, identified these isolates as strains of *Thermoanaerobacter thermohydrosulfuricus*. Application of this model to other *Thermoanaerobacter* spp. further identified that these isolates belong to a divergent and lesser characterized lineage within the genus.

Based on this, the CBP-potential of a single isolate, *T. thermohydrosulfuricus* WC1, was selected for further investigation through metabolic, genomic and proteomic analyses. Its ability to grow on polymeric xylan, potentially catalyzed by an endoxylanase found in only a few *Thermoanaerobacter* strains, distinguishes *T. thermohydrosulfuricus* WC1 from many other strains within the genus. The simultaneous consumption of two important lignocellulose constituent saccharides, cellobiose and

xylose was also observed and represents a desirable phenotype in CBP-relevant organisms.

However, at elevated sugar concentrations, *T. thermohydrosulfuricus* WC1 produces principally lactate, rather than the desired biofuel ethanol, as the major end-product. Proteomic analysis identified that all likely end-product forming proteins were expressed at high levels suggesting that the end-product distribution patterns in *T. thermohydrosulfuricus* WC1 are likely controlled via metabolite-based regulation or are constrained by metabolic bottlenecks.

The xylanolytic and simultaneous substrate utilization capabilities of *T. thermohydrosulfuricus* WC1 identify it as a strain of interest for CBP. However, for its development into an "industry-ready" strain as a co-culture with a cellulolytic microorganism, improved biofuel producing capabilities are needed. The practical implications of CBP-relevant phenotypes in *T. thermohydrosulfuricus* WC1 in relation to other *Thermoanaerobacter* spp. will be discussed.

Acknowledgements

With deepest gratitude, I would like to first thank my advisor, Richard Sparling for granting me the opportunity to do this research and challenging me throughout. Your insights and abilities as a researcher have provided an example of who I aspire to become. Thanks also to David Levin, whose commitment to this work has been nothing short of admirable and whose insights have been incredibly valuable. I also extend my sincere appreciation to Vladimir Yurkov and Teresa De Kievit, whose advice and unfaltering encouragement has provided me strength throughout this degree.

Thank you to Tim Dumonceaux, who saw potential where I was oblivious. Thanks to Vic Spicer and Oleg Krokhin who guided me through many struggles I encountered; even when those struggles defied all logic. Thank you to Gideon Wolfaardt and Martina Hausner whose generosity and support have been limitless. To Tom Rydzak, John Schellenberg, Carlo Carere and Scott Wushke; thank you for everything our numerous discussions/debates have taught me. Finally, thank you to all who have encouraged and supported me throughout, particularly Nazim Cicek, John Wilkins, Justin Zhang, Peter McQueen, Brian Fristensky, Marcel Taillefer, Umesh Ramachandran, Nathan Wrana, Bruce Ford, Matthew Links, Janet Hill, Andrea Wilkinson, Alexandru Dumitrache, Dmitry Shamshurin, Georg Hausner and Ivan Oresnik.

This work would not have been possible without support provided by a Natural Sciences and Engineering Research Council grant (STPGP 365076), the University of Manitoba, the Manitoba Rural Adaptation Council (MRAC) – Advancing Canadian Agriculture and Agri-Food (ACAAF) program (309009) and by a Genome Canada grant titled “Microbial Genomics for Biofuels and Co-Products from Biorefining Processes.”

Dedication

To my wife, Lindsay,

None of this matters without you.

And to my parents, Dan and Karen, and my sister, Aynsley,

Encouragement. Confidence. Effort. Inspiration. Love.

Enough said.

Table of contents

List of tables

List of figures

List of copyrighted material for which permission was obtained

Material	Source	Page(s) in thesis
Chapter 1.3.2 - 1.3.3	Omic approaches for designing biofuel producing co-cultures for enhanced microbial conversion of lignocellulosic substrates. © 2013, David B. Levin, University of Manitoba	16-32
Fig. 1.2	BMC Microbiology. 2012. **12**: 295.	36
Chapter 2	FEMS Microbiology Ecology. 2011. **78**: 473-487.	49-82
Chapter 3	Systematic and Applied Microbiology. 2011. **34**: 171-179.	83-107

List of abbreviations used

2D-HPLC-MS/MS	two dimensional high performance liquid chromatography tandem mass spectrometry
AA	auxiliary activities; in reference to types of CAZymes
ABC	ATP-binding cassette
ack/ACK	acetate kinase
adh/ADH	alcohol dehydrogenase
adhE/AdhE	bifunctional acetaldehyde:alcohol dehydrogenase
AHH	acid hydrolyzed hemicellulose
aldH/AldH	acetaldehyde dehydrogenase
ANI	average nucleotide identity
ANIb	average nucleotide identity by BLAST
ANIm	average nucleotide identity by MuMmer[1]
ATCC	American Type Culture Collection
atk/ATK	acetate thiokinase
BLAST	basic local alignment search tool
BOX	BOX-repetitive element[2]
CAZy(me)	carbohydrate active enzyme
CBM	carbohydrate binding module
CBP	consolidated bioprocessing

[1] For a description of MuMmer, see p. 88.
[2] The term "BOX" is a name for a type of repetitive element and is not an acronym. See Martin *et al*., (1992).

ccpA	catabolite control protein A
CCR	carbon catabolite repression
CDS	coding sequence(s)
CE	carbohydrate esterase
CoA	coenzyme-A
COG	clusters of orthologous groups
cpn60 UT	chaperonin-60 universal target
cre	catabolite responsive elements
crH	catabolite repression HPr-like
DAP	dihydroxyacetone phosphate
DAPI	4’,6-diamidino-2-phenylindole
DDBJ	DNA Databank of Japan
DSM(Z)	Deutsche Sammlung von Mikroorganismen (und Zellkulturen)
EC	enzyme commission
ECH	energy conserving hydrogenase
EMBL	European Molecular Biology Laboratory
EMP	Embden-Meyerhoff-Parnas pathway; synonymous with glycolysis
FASP	filter aided sample preparation
FBP	fructose-1,6-bisphosphate
FISH	fluorescent *in situ* hybridization
GC	gas chromatography

gDNA	genomic DNA
gen	generation
GH	glycoside hydrolase
GlcNAc	N-acetyl-glucosamine
GPU	graphics processing unit
GT	glycosyltransferase
H_2ase	hydrogenase
HGT	horizontal gene transfer
HOV	higher order variable; e.g. COG, KO, EC, METACYC
HPLC	high performance liquid chromatography
HPr	histidine containing protein
HPrK/P	HPr kinase/phosphatase
IMG	Integrated Microbial Genomes
IMG-ER	Integrated Microbial Genomes - Expert Review
IVS	intervening sequence(s)
JGI	Joint Genome Institute
KO	KEGG Orthology groups
LC	liquid chromatography
ldh/LDH	lactate dehydrogenase
mbh/MBH	membrane bound hydrogenase
mbx/MBX	membrane bound hydrogenase-like enzyme
mdh/MDH	malate dehydrogenase

MGF	Mascot Generic Format
MLST	multi-locus sequence typing
MS	mass spectrometry
MS/MS	tandem mass spectrometry
NCBI	National Center for Biotechnology Information
nfo/NFO	NADH:ferredoxin oxidoreductase; synonymous with Rnf
NHH	neutral hydrolyzed hemicellulose
NOI	niche overlap index/indices
NREL	National Renewable Energy Laboratory
nTIC	normalized total ion current
O/R	ratio of oxidized end-products to reduced end-products formed
OTU	operational taxonomic unit
P-	phospho- or phosphate; e.g. glucose-6-P is glucose-6-phosphate
PAS	Per-ARNT-Sim domain[3]
PBS	phosphate buffered saline
pdh/PDH	pyruvate dehydrogenase
PEP	phosphoenolpyruvate
Pfam	protein family
pfk/PFK	phosphofructokinase

[3] Domain named after the three proteins in which it was first described. See Taylor & Zhulin, (1999).

pfl/PFL	pyruvate formate lyase
PL	polysaccharide lyase
por/POR	pyruvate: ferredoxin oxidoreductase; synonymous with PFOR or PFO
ppdk/PPDK	pyruvate phosphate dikinase
PPi	pyrophosphate
PRD	PTS-regulatory domain
pta/PTA	phosphotransacetylase
PTS	phosphotransferase system (transport)
Rnf	*Rhodobacter* nitrogen fixing; synonymous with NFO
rrn	ribosomal RNA gene sequence
SI	sequence identity
SLH	S-layer homology binding domain
SNP	single nucleotide polymorphism
SWATH-MS	SWATH-mass spectrometry[4]
TCDB	Transporter Classification Database
TIC	total ion current
TIGRFAM	The Institute for Genomic Research Family

[4] For a description of SWATH-MS, see p. 158.

List of genus abbreviations used

A.	*Acetivibrio*
Ac.	*Acetobacterium*
B.	*Bacillus*
C.	*Clostridium*
Ca.	*Caldicellulosiruptor*
Cal.	*Caldanaerobacter*
E.	*Escherichia*
Et.	*Ethanoligenens*
G.	*Geobacillus*
L.	*Lactobacillus*
M.	*Moorella*
P.	*Pyrococcus*
S.	*Saccharomyces*
T.	*Thermoanaerobacter*[5]
Th.	*Thermoanaerobacterium*
The.	*Thermotoga*
Ther.	*Thermococcus*

[5] To assist with taxonomic assignments of strains within this genus see Appendix A.8.

Chapter 1. Literature review[6]

1.1 Introduction

The need for renewable energy, whether driven by environmental, economic, or sociopolitical concerns, has been well established (Rosegrant *et al*., 2008; Mussato *et al*., 2010; Naik *et al*., 2010; Olson *et al*., 2012). In particular, as oil consumption comprises the largest fraction of the world's energy supply and >60% of the oil consumed (as of 2010) is by the transportation sector (IEA, 2012), finding alternative fuels suitable for this sector is important in reducing the world's reliance on finite petroleum reserves.

Producing liquid biofuels such as ethanol, to use as a blended mixture with gasoline, is a promising approach to reduce petroleum consumption and is currently being commercially produced from sugar or starch based feedstocks (termed "first-generation biofuels") (Naik *et al*., 2010; Li *et al*., 2010a; Goldemberg, 2007). However, increasing ethanol production using first-generation technologies has been intensely debated as concerns over using land for dedicated energy crops, rather than food production, have arisen in recent years (Naik *et al*., 2010; Graham-Rowe, 2011). Further, low reductions in greenhouse gas emissions and increased water consumption associated with expanding agricultural lands for biofuel driven crops, have called into question the environmental benefits of using first-generation technologies (DeLucchi, 2010; Havlík *et al*., 2011). Alternatively, ethanol or other biofuels such as hydrogen, butanol or methane, produced from lignocellulosic biomass (termed "second-generation biofuels") are thought

[6] Modified versions of sections 1.3.2 and 1.3.3, as well as the corresponding Tables, have been submitted for publication. All versions of these sections were authored by Tobin J. Verbeke. Reproduced with permission: **©** 2013, David B. Levin, University of Manitoba. Found in: **Levin DB, Verbeke TJ, Munir R, Islam R, Ramachandran U, Lal S, Schellenberg J, Sparling R**. Omic approaches for designing biofuel producing co-cultures for enhanced microbial conversion of lignocellulosic substrates.

to avoid many of these issues. A specific driving force behind second-generation biofuels is that low-value biomass, derived from cellulosic waste streams from the forestry and agricultural industries, or non-food crops grown on marginal lands, can be used as feedstocks for fuel producing microbes.

While different strategies to produce second-generation biofuels exist (Hamelinck *et al*., 2005; Margeot *et al*., 2009), one of the most economically promising approaches is consolidated bioprocessing (CBP) (Lynd *et al*., 2005). In CBP, the production of lignocellulose hydrolyzing enzymes, and the subsequent conversion of the hydrolysis products to biofuels is catalyzed by microorganisms in a single reactor (Lynd *et al*., 2005). This strategy is considered advantageous over others in terms of process simplification and cost reduction. With CBP, the requirements for expensive pre-treatment strategies, or the use of costly enzymatic cocktails for saccharification of the biomass, can potentially be minimized or eliminated.

Achieving industrially significant ethanol production through CBP is currently limited by the general recalcitrance of lignocellulose to hydrolysis and by the low biofuel yields produced by the catalyzing microbes (Olson *et al*., 2012). The identification of new microorganisms, or the improvement of previously characterized ones, capable of increased hydrolysis and conversion efficiencies, is needed to make lignocellulosic ethanol production a commercially viable alternative fuel.

1.2 Microorganisms for second generation biofuels through CBP

The selection of microorganisms for second generation biofuel production has typically followed three distinct strategies, which include: i) the use of naturally

occurring microbial consortia capable of biomass hydrolysis and biofuel production; ii) the genetic engineering and development of established platform microorganisms; or iii) the identification, characterization and development of novel microorganisms into industry-relevant microbes (as monocultures or in designer co-cultures).

1.2.1 Naturally occurring consortia

The enrichment of natural microbial communities is one approach for achieving simultaneous biomass hydrolysis and biofuel production. While these communities have often been used for bioprospecting studies targeted at the identification of novel strains or enzymes (Krause *et al*., 2003; Hess *et al*., 2011; Amore *et al*., 2013), communities arising from diverse environments including the guts of herbivorous insects (Shi *et al*., 2011) or ruminants (Ho *et al.*, 2011; Weimer, 2011), soil, manure and/or compost (Haruta *et al*., 2002; Wongwilaiwalin *et al*., 2010) or cellulose containing waste streams (Liu *et al*., 2003; Fan *et al*., 2006) have all been investigated for their lignocellulose hydrolyzing/biofuel producing potential. For industrial purposes, adopting a non-aseptic approach, as can be achieved when using natural communities, is advantageous in terms of feedstock preparation. However, in using this strategy, the continued selective enrichment and stable existence of an efficient biofuel producing community can be a significant challenge (Shong *et al*., 2012) given the dynamic nature of native microbial communities. This is particularly true in bioprocessing systems, where selective pressures may favour strains capable of rapid growth rather than biofuel production. While many culturing parameters (e.g. pH, temperature, residence time, etc.) can be kept consistent to reduce dynamic progression of the communities, unavoidable perturbations,

such as variations in feedstock composition, can significantly affect community composition (Wang *et al*., 2009; Allgaier *et al*., 2010; Reddy *et al*., 2011), making consistent biofuel production from natural consortia difficult.

The production of non-methane biofuels in natural consortia is also typically low (Zuroff & Curtis, 2012). This may be partly attributable to the fact that, in natural communities, fuel molecules and other end-products are consumed by other consortium members making the targeted production of certain biofuels challenging. For example, a biofuel such as hydrogen may be readily consumed in the presence of methanogens. While engineering and chemical approaches exist to select against such fuel-consuming microbes (Valdez-Vazquez *et al*., 2005), doing so requires additional costs ultimately affecting the commercial viability of the process if methane is not the desired fuel.

Finally, the continued improvement of biofuel producing cultures is limited when using natural consortia. The undefined and dynamic nature of these communities makes identifying and understanding physiological factors limiting biofuel production challenging. This ultimately reduces the potential for targeted strategies designed to alleviate such limitations.

1.2.2 Established platform organisms

The continued development of current platform microorganisms, such as *Saccharomyces cerevisiae* and *Escherichia coli*, for second-generation biofuels is an alternative approach to using natural consortia. In some respects, this approach is advantageous in that a wealth of accumulated knowledge regarding strain physiology, as well as established genetic engineering tools, already exist for these strains (Fischer *et al*.,

2008). The challenge in this approach is that many desirable CBP-relevant phenotypes are not innately present in organisms like *E. coli* or *S. cerevisiae* and have to ultimately be introduced through genetic engineering. Expressing non-native pathways or phenotypes into these hosts may also unintentionally be detrimental to their native physiology (Toivari *et al*., 2001).

For example, *S. cerevisiae* is currently used for first generation biofuels due to its ethanologenic capabilities. However, its inability to ferment pentose sugars (Dellomonaco *et al*., 2010), a major component of hemicellulose, limits its suitability for second-generation biofuel production. Many attempts at engineering a xylose-utilization pathway into *S. cerevisiae* have occurred (reviewed in van Zyl *et al*., 2007; Nevoigt *et al*., 2008). One such approach involved the heterologous expression of a xylose reductase and xylitol dehydrogenase from a xylose fermenting yeast, *Pichia stipitis*, in *S. cerevisiae* (Toivari *et al*., 2001). Expression of these enzymes would allow for the conversion of xylose to xylulose, which is natively consumed by *S. cerevisiae*. The heterologous expression of the xylose reductase and xylitol dehydrogenase, in combination with an increased expression of the native *S. cerevisiae* xylulose kinase, conferred xylose utilization capabilities into the strain (Toivari *et al*., 2001).

However, when grown under anaerobic conditions, much of the consumed xylose was excreted as xylitol (representing unfermented/unconverted xylose) and sustained growth was not achieved. The limited growth was due to imbalances in cellular redox levels involving NADH accumulation and NADPH depletion (Toivari *et al*., 2001; Dellomonaco *et al*., 2010). Strategies that modified co-factor dependence, such as mutating the xylose reductase (Jeppsson *et al*., 2006) or ammonia assimilation pathway

(Grotkjaer *et al*., 2005) to be NADH- rather than NADPH-dependent were necessary to help restore redox balances. While both strategies reduced xylitol excretion, neither approach fully eliminated excretion and significant amounts of carbon were still lost.

Conversely to *S. cerevisiae*, *E. coli* has the capability to ferment xylose, but can also produce significant amounts of organic acids (lactic, formic, succinic, acetic) in addition to ethanol (Clomburg & Gonzalez, 2010). The production of non-biofuel end-products, which ultimately reduce ethanol yields, is partly attributable to a lack of sufficient reducing equivalents suitable for ethanol production (Clomburg & Gonzalez, 2010). In native *E. coli,* the generation of a single pyruvate molecule through glycolysis yields only one reducing equivalent in the form of NAD(P)H. Further, under anaerobic conditions, pyruvate is principally consumed via a pyruvate formate lyase (PFL) enzyme as the pyruvate dehydrogenase (PDH) complex is repressed (Clark, 1989; Sauer *et al*., 1999). The PFL-catalyzed reaction consumes pyruvate and produces formate and acetyl-CoA, but yields no additional reducing equivalents. As the conversion of acetyl-CoA to acetaldehyde, and then to ethanol, requires two NAD(P)H equivalents in native *E. coli*, homo-ethanol production is not possible.

In separate experiments, Ingram *et al*., (1987) and Ohta *et al*., (1991) were able to circumvent redox imbalances by engineering *E. coli* to express the pyruvate decarboxylase from *Zymomonas mobilis*. This permitted the direct conversion of pyruvate to acetaldehyde, without the consumption of a reducing equivalent (NAD(P)H). In doing this, only the single NAD(P)H produced in pyruvate formation is necessary for ethanol production. In both studies, ethanol became the principal end-product formed from the fermentation of glucose.

Platform organisms, such as *S. cerevisiae* and *E. coli*, have undergone extensive genetic engineering in an attempt to develop them into biofuel producing microbes (for more extensive reviews see van Zyl *et al.*, 2007; Nevoigt, 2008; Cloumburg & Gonzalez, 2010). Apart from the previously mentioned limitations in wild type strains, both organisms similarly have no natural capacities for lignocellulose hydrolysis, and, particularly in *S. cerevisiae*, are also limited in their usage of oligosaccharides derived from lignocellulose hydrolysis (Fischer *et al*., 2008). Use of these organisms in CBP systems would therefore additionally require expensive pre-treatment strategies or enzymatic cocktails designed to hydrolyze lignocellulose into soluble sugars, co-culturing with hydrolytic strains, or additional engineering efforts to impart hydrolytic capabilities into the strains themselves. As engineering cellulolytic and/or hemicellulolytic ability into these strains is also being extensively investigated (van Zyl *et al*., 2007; Shin *et al*., 2010; Olson *et al*., 2012; Saha & Cotta, 2012), this may represent a promising approach. Given that, in many cases, engineering these established platform organisms has been successful, the limitations in their development may solely be the time and effort required to incorporate all desired phenotypes into a single strain without detrimentally influencing the wild type physiology.

1.2.3 Novel platform organisms

1.2.3.1 Bioprospecting for novel strains

Bioprospecting for novel enzymes and/or microorganisms has been a promising third approach for CBP-platform development. This strategy, which seeks to exploit the innate potential found in many diverse microorganisms, has made significant

advancements in recent years. Novel enzyme discovery, specifically for enzymes involved with biomass hydrolysis, has particularly benefitted from the development of meta-analyses technologies such as metagenomics (Warnecke *et al*., 2007; Hess *et al*., 2011) and/or metatranscriptomics (Tartar *et al*., 2009; Takasaki *et al*., 2013). However, bioprospecting solely for novel enzymes has limitations in terms of its implementation into CBP systems. First, while newly discovered enzymes may be cloned into host organisms and expressed, genetic engineering of many CBP-relevant strains has proven to be challenging (Chapter 1.3.5). Secondly, while these enzymes can be cloned into genetically amenable hosts and purified, the use of enzyme cocktails is an undesirable cost in CBP systems, which seek to make use of a strain's *in situ* hydrolytic capabilities.

A second approach to bioprospecting involves the identification of novel microorganisms. In many cases, lignocellulose degrading environments, such as those used for investigating natural consortia (Chapter 1.2.1), serve as valuable resources for the isolation of new strains. Efforts to identify potential platform strains have remained consistent since the U.S. oil crises of the 1970s. Further, technological developments in genomics and gene expression profiling have also recently prompted researchers to revisit strains originally isolated decades ago (Table 1.1). While the advantages of using strains with innate CBP-relevant phenotypes are obvious, one of the major limitations is that the physiology of most of these innately capable strains is also not well understood, even with the increasing availability of genomic and gene expression profiling datasets available for many of the organisms. This therefore requires significant effort to characterize CBP-relevant phenotypes in non-established platform microorganisms as a means of evaluating their potential for industrial biofuel production.

Table 1.1. List of select biofuel relevant microorganisms identified through bioprospecting since the late 1970s.

Strain	Isolation reference	Bioproject ID[a]	Genome released[b]
Genus: *Acetivibrio*			
A. cellulolyticus CD2	Patel *et al*., 1980	45843	2012
Genus: *Caldanaerobacter*			
Cal. subterraneus subsp. *tengcongensis* MB4	Xue *et al*., 2001	249	2006
Genus: *Caldicellulosiruptor*			
Ca. bescii DSM 6725	Svetlichnyi *et al*., 1990	29407	2009
Ca. hydrothermalis 108	Miroshnichenko *et al*., 2008	40831	2011
Ca. kristjanssonii I77R1B	Bredholt *et al*., 1999	41727	2011
Ca. kronotskyensis 2002	Miroshnichenko *et al*., 2008	52409	2011
Ca. lactoaceticus DSM 9545	Miadenovska *et al*., 1995	40219	2012
Ca. obsidiansis OB47	Hamilton-Brehm *et al*., 2010	40355	2011
Ca. owensensis OL	Huang *et al*., 1998	40833	2011
Ca. saccharolyticus DSM 8903	Rainey *et al*., 1994	13466	2007
Genus: *Clostridium*			
C. carboxidivorans P7	Liou *et al*., 2005	29495	2010
C. cellulolyticum H10	Petitdemange *et al*., 1984	17419	2009
C. cellulovorans 743B	Sleat *et al*., 1984	37295	2011
C. intestinale URNW	Ramachandran *et al*., 2011	NP[c]	NP
C. papyrosolvens DSM 2782	Madden *et al*., 1982	33857	2009
C. phytofermentans ISDg	Warnick *et al*., 2002	16184	2008
C. ragsdalei P11	Huhnke *et al*., 2010	210333	2012
C. stercorarium DSM 8532	Madden, 1983	195569	2013
C. termitidis CT112	Hethener *et al*., 1992	196409	2013
C. thermocellum DSM 4150	Freier *et al*., 1988	28257	2008
C. thermopalmarium DSM 5974	Lawson *et al*., 1991	NP	NP
C. thermopapyrolyticum UBA 305	Mendez *et al*., 1991	NP	NP

Table 1.1 cont.

Strain	Isolation reference	Bioproject ID	Genome released
Genus: *Ethanoligenens*			
Et. harbinense YUAN-3	Xing *et al*., 2006	39729	2011
Genus: *Geobacillus*			
G. debilis DSM 16016	Banat *et al*., 2004	201207	2012
G. thermoglucosidasius NCIMB 11955	Suzuki *et al*., 1983	40781	2011
Genus: *Thermoanaerobacter*			
T. ethanolicus JW200	Wiegel & Ljungdahl, 1981	51151	2012
T. italicus Ab9	Kozianowski *et al*., 1997	33157	2010
T. mathranii susbp. *mathranii* A3	Larsen *et al*., 1997	33329	2010
T. pseudethanolicus 39E	Zeikus *et al*., 1980	13901	2008
Thermoanaerobacter sp. X514	Roh *et al*., 2002	16394	2008
T. thermohydrosulfuricus DSM 567	Wiegel *et al*., 1979	NP	NP
T. wiegelii Rt8.B1	Cook *et al*., 1996	42251	2011
Genus: *Thermoanaerobacterium*			
Th. saccharolyticum B6A-R1	Lee *et al*., 1993	NP	NP
Th. xyloanolyticum LX-11	Lee *et al*., 1993	50295	2011
Genus: *Thermotoga*			
The. maritima MSB8	Huber *et al*., 1986	111	2006, 2013
The. neapolitana DSM 4359	Jannasch *et al*., 1988	21023	2009
The. petrophila RKU-1	Takahata *et al*., 2001	17089	2007
Genus: *Pyrococcus*			
P. furiosus DSM 3638	Fiala *et al*., 1986	287	2006
Genus: *Thermococcus*			
Ther. kodakarensis	Morikawa *et al*., 1994	13213	2006

[a]GenBank Genome Bioproject ID. Multiple genome accession numbers found for strains which encode plasmids. Entire genome content (chromosome + plasmids) accessible via Bioproject ID.

[b]Year of genome release as listed in the Integrated Microbial Genomes (IMG) database (Markowitz *et al*., 2012).

[c]NP = genome not publicly released or not sequenced.

1.2.3.2 Continued development of novel organisms

Strains receiving significant attention for continued development are found in diverse phylogenetic lineages (Fig. 1.1). However organisms in the phylum *Firmicutes*, and more specifically the orders *Clostridiales* and *Thermoanaerobacteriales*, have garnered the most attention for industrial development (Lynd *et al*., 2008). This largely stems from the realization that multiple strains of *Clostridia*: i) possess native lignocellulosic hydrolysis machinery; ii) can ferment a broad range of substrates; and iii) have diverse pathways suitable for the production of numerous valuable metabolites (Tracy *et al*., 2012). However, no single strain has yet been identified that is capable of performing all CBP-relevant processes at industrially significant levels.

One approach to address these limitations is to construct designer co-cultures. Such designer consortia strive to mimic synergies found in naturally occurring communities, but do so using well defined and characterized constituent members. Co-cultures have been developed using bacteria-yeast combinations (Zuroff *et al*., 2013), mesophilic bacteria (Wang *et al*., 2008; Jiao *et al*., 2012), or most commonly, using thermophilic bacteria (Ng *et al*., 1981; Liu *et al*., 2008; He *et al*., 2011; Xu & Tschirner, 2011; Kridelbaugh *et al*., 2013; Svetlitchnyi *et al*., 2013).

Often, the design of co-cultures is to couple the hydrolytic capabilities of one strain, with the biofuel production capacities of another. This approach has been adopted numerous times involving co-cultures of *Clostridium thermocellum* with strains of the genera *Thermoanaerobacter* (Ng *et al*., 1981; Le Ruyet *et al*., 1984; He *et al*., 2011; Hemme *et al*., 2011) or *Thermoanaerobacterium* (Saddler & Chan, 1984; Liu *et al*., 2008; Argyros *et al*., 2011). Consistent amongst these co-cultures are reports of

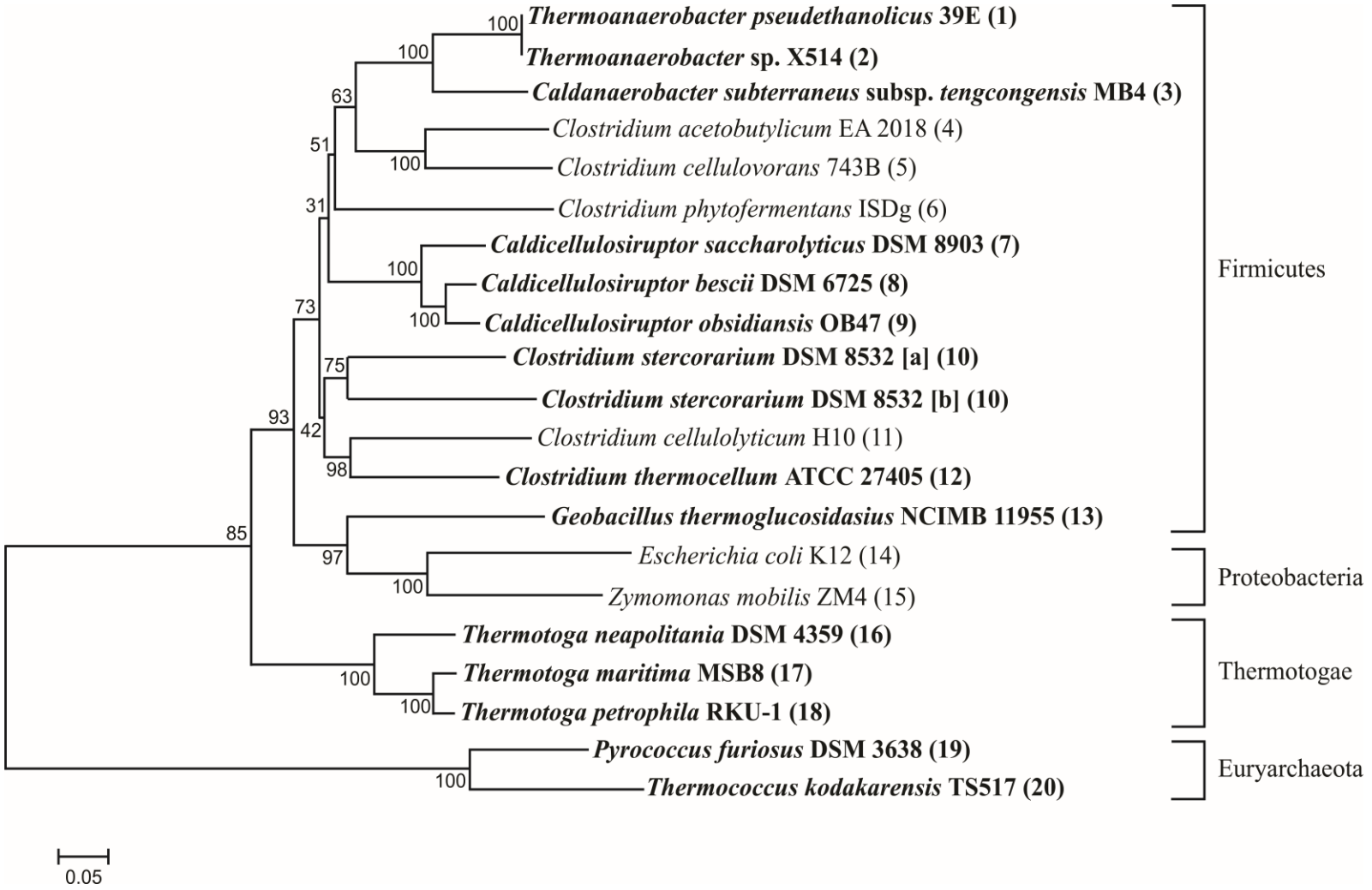


Fig. 1.1. Phylogenetic distribution of microorganisms being investigated as potential platform organisms for the production of second-generation biofuels through CBP. Tree construction is as described (Chapter 3.3.3) using the full length *cpn60* (Type I

& Type II chaperonin) nucleotide sequences accessed via the chaperonin database (cpnDB) (Hill *et al*., 2004). Analyses limited to strains within the Bacterial and Archaeal domains, with published *cpn60* data and also to strains with transcriptomic and/or proteomic currently available. **Bold** lettering identifies thermophilic strains. Bootstrapping values shown beside their respective nodes. Letters in square brackets represent two distinct *cpn60* sequences for *C. stercorarium* DSM 8532. Numbers in round brackets correspond to select literature reporting gene expression data for each strain and are as follows: **1** = Hemme *et al*., 2011; **2** = Hemme *et al*., 2011; Lin *et al*., 2011; **3** = Meng *et al*., 2009; Chen *et al*., 2013; **4** = Hu *et al*., 2011; **5** = Morisaka *et al*., 2012; **6** = Tolonen *et al*., 2011; **7** = VanFossen *et al*., 2009; VanFossen *et al*., 2011; **8** = Dam *et al*., 2011; Lochner *et al*., 2011a; **9** = Lochner *et al*., 2011a, Lochner *et al*., 2011b; **10** = Schellenberg *et al*., submitted; **11** = Blouzard *et al*., 2010; **12** = Raman *et al*., 2011; Rydzak *et al*., 2012; **13** = Loftie-Eaton *et al*., 2013; **14** = Hu & Wood, 2010; **15** = He *et al*., 2012; **16** = Frock *et al*., 2012; **17** = Shockley *et al*., 2005; Frock *et al*., 2012; **18** = Frock *et al*., 2012; **19** = Wong *et al*., 2013; **20** = Čuboňováa *et al*., 2012.

improved biomass hydrolysis and/or overall biofuel production in comparison to monocultures. Additionally, metabolite sharing between these co-cultures has been reported as the *de novo* biosynthetic capabilities of one strain can complement nutritional auxotrophies present in the co-culture partner (Mori, 1990; Mori, 1995). This aspect has the added benefit of potential simplification of fermentation broths used leading to cost reduction.

Most co-culture studies to date have largely been empirical leading to a limited potential for culture development. To avoid these limitations, co-culture design ultimately requires a detailed knowledge of its constituent members' physiologies. As such, multiple physiological criteria (Chapter 1.3) are important to evaluate when selecting one, or multiple, strains for development of CBP-relevant cultures. In recent years, "omic"-based technologies (genomics, transcriptomics, proteomics, etc.) have helped shed insight into many of these criteria and have led to a rapid advancement in our understanding of key physiological processes for many strains of interest. The continued implementation of "omics"-based approaches will remain valuable in the development of newly isolated strains or established candidate organisms for development into CBP platform microbes.

1.3 Criteria to evaluate in the selection or development of novel platform strains

1.3.1 Advantages of thermophilic microorganisms

Of the organisms garnering the most attention for CBP (Fig. 1.1), most of the attention has focused on thermophilic microorganisms due to advantages directly associated with thermophily (Lynd *et al*., 2008; Taylor *et al*., 2009). Firstly, thermophilic

growth conditions may help reduce or minimize the magnitude of reiterative cycling between upstream biomass pre-treatment strategies and the creation of conditions amenable for bacterial growth. Many physical and/or chemical pre-treatment strategies used for lignocellulosic saccharification including, but not limited to, steam explosion (Mosier *et al*., 2005; Datar *et al*., 2007), liquid hot water treatment (Wyman *et al*., 2005), various acid or alkali pre-treaments (Mosier *et al*., 2005; Agbor *et al*., 2011), hydrogen peroxide (Silverstein *et al*., 2007) or wet oxidation (Schmidt & Thomsen, 1998, Martín *et al*., 2007), all involve heating the lignocellulosic biomass. Solutions containing the resulting carbohydrates and carbohydrate polymers suitable for fermentation ultimately have to then be cooled prior to utilization by the microorganisms. Minimizing the magnitude of temperature variation between processes can help reduce energy input costs as well as lower transition time periods.

Secondly, thermostable enzymes capable of lignocellulose hydrolysis, which may be exogenously added during enzymatic pre-treatment, or synthesized *de novo* by the fermentative microorganism in a CBP platform, are inherently more robust than their mesophilic counterparts (Yeoman *et al*., 2010). This increased stability and/or longer half-life may allow for extended activity and increased tolerance to harsh conditions associated with industrial lignocellulose hydrolysis (Turner *et al*., 2007). Additionally, the enzymatic activity of thermostable enzymes is typically higher at elevated temperatures potentially leading to improved hydrolysis of the biomass (Viikari *et al*., 2007).

Growth under thermophilic conditions also facilitates downstream product recovery due to the increased volatility of fuels, such as ethanol, at elevated temperatures.

Reduced energy inputs are required to distill liquid biofuels from fermentation broths used for thermophilic organisms. Additionally, the reduced solubility of gases at elevated temperatures not only facilitates H_2 recovery, but also helps to maintain reactors in deoxygenated states suitable for anaerobic fermentation.

Dependent on the reactor design, the reduced solubility of gases at higher temperature can also favor the maintenance of a low H_2 partial pressure. Combining a low partial pressure, along with the improved thermodynamic efficiencies associated with elevated temperature (Verhaart *et al*., 2010), may improve the formation of biofuels such as H_2 to reach theoretical limits (4 mol H_2 per mol hexose) and improve overall yields (Soboh *et al*., 2004; Carere *et al*., 2012).

Finally, systems operating at high temperatures inherently select against pathogenic microorganisms. Current commercial systems involving yeast based ethanol production address bacterial contamination through the use of antibiotics or other additives (eg. ammonia, urea, hydrogen peroxide) (Skinner *et al*., 2004; Bischoff *et al*., 2007; Beckner *et al*., 2011), which not only adds to overall process costs, but also creates a hazardous waste stream that must be addressed prior to disposal. While elevated temperatures do not ensure culture sterility, the industrial implications of avoiding potential pathogenic contamination are significant.

1.3.2 Lignocellulose hydrolysis potential

The recalcitrant nature of lignocellulose and its resistance to enzymatic hydrolysis are often viewed as one of the major limitations to overcome in improving lignocellulosic biofuel production (Chundawat *et al.*, 2011, Das *et al*., 2012). Selecting strains with a

high maximal *in situ* hydrolytic potential is therefore important in developing efficient CBP platforms. In recent years, investigations into the hydrolytic potential of thermophilic microorganisms have increasingly employed "omic"-based strategies (Raman *et al*., 2009; Lochner *et al.*, 2011a, b; VanFossen *et al*., 2011) to elucidate mechanisms of biomass deconstruction and identify potential limitations.

Lignocellulose hydrolyzing enzymes belong to a broader class of proteins known as Carbohydrate Active enZymes (CAZymes), whose functions are involved in diverse biological roles including, but not limited to, biomass hydrolysis, energy production, cell wall or peptidoglycan synthesis and protein glycosylation (Cantarel *et al*., 2009). While current CAZyme classes include proteins that act as glycoside hydrolases (GHs), glycosyltransferases (GTs), polysaccharide lyases (PLs), carbohydrate esterases (CEs) or those with auxiliary activities (AAs) (Levasseur *et al*., 2013), most "omic" investigations targetteting lignocellulose hydrolysis have focused on GH expression. While GHs, and CAZymes in general, represent a significant portion of protein encoding genes within any given genome (Coutinho *et al*., 2003) the number and diversity of extracellular glycoside hydrolases capable of lignocellulose hydrolysis can be a distinguishing factor between biofuel-producing *Firmicutes* (Table 1.2).

Some strains, such as *Clostridium stercorarium* DSM 8532 or most notably, *Clostridium thermocellum* ATCC 27405, have highly investigated diverse and extensive suites of GHs (Lynd *et al*., 2002; Zverlov *et al*., 2005; Zverlov & Schwarz, 2008) and are seemingly poised for extensive biomass hydrolysis (Table 1.2). Other strains, such as *Caldanaerobacter subterraneus* subsp. *tengcongensis* MB4 and *Thermoanaerobacter* sp. X514, have no predicted extracellular CAZymes and are unlikely to significantly

Table 1.2. Extracellular[a] glycoside hydrolases related to lignocellulose hydrolysis in the genomes of select thermophilic *Firmicutes*.

	Glycoside hydrolase (GH) class																			
Strain	3	5	8	9	10	11	16	26	27	31	42	43	44	48	51	52	53	54	74	81
Ca. bescii DSM 6725		+		+	+	+								+						
Ca. kristjanssonii I77R1B		+		+	+								+						+	
Ca. obsidiansis OB47		+		+	+		+							+					+	
Ca. saccharolyticus DSM 8903		+		+	+									+					+	
Cal. subterraneus subsp. *tengcongensis* MB4																				
C. thermocellum ATCC 27405		+	+	+	+	+	+		+	+	+		+		+			+		
C. stercorarium DSM 8532	+			+	+	+			+			+		+			+			
G. thermoglucosidasius NCIMB 11955					+											+				
T. mathranii susbp. *mathranii* A3					+							+				+				
Thermoanaerobacter sp. X514																				
Th. saccharolyticum JW/SL-YS485		+			+	+		+								+	+			
Th. thermosaccharolyticum DSM 571		+			+			+								+				
Th. xylanolyticum LX-11					+		+									+				+

[a]Extracellular localization for glycoside hydrolases within each CAZyme class predicted using PSortB 3.0 (Yu *et al*., 2010) and using the final subcellular localization prediction reported for each CAZyme. Individual sequences for each strain were identified by accessing the CAZy database (Cantarel *et al*., 2009).

contribute to the hydrolysis of polymeric substrates. These strains would therefore be reliant on hydrolysis through pre-treatment approaches, enzyme addition or would require co-culturing with a more hydrolytic strain. Within the *Firmicutes*, distinct strategies have also evolved for GH localization and achieving hydrolysis of lignocellulosic polymers. Specifically, some strains freely secrete their hydrolytic enzymes (Lochner *et al*., 2011a; Tolonen *et al*., 2011), while others affix their hydrolytic enzymes to their cell wall surface (Shao *et al*., 1995; Brechtel & Bahl, 1999). Others yet, have evolved extracellular appendage-like structures termed "cellulosomes" that co-localize diversely acting hydrolytic enzymes onto a central scaffoldin protein to coordinate hydrolysis activities (Gold & Martin, 2007; Blouzard *et al*., 2010; Tamaru *et al*., 2011; Morisaka *et al*., 2012).

Global profiling strategies have improved our understanding of the hydrolytic potential for many strains and have helped identify their limitations also. For example, comparative genomic analyses of sequenced *Caldicellulosiruptor* strains has identified that the absence of GH48 domain-containing proteins in *Caldicellulosiruptor hydrothermalis* 108, *Caldicellulosiruptor owensensis* OL and *Caldicellulosiruptor kristjanssonii* I77R1B correlates with the reduced ability of those strains, in comparison to other *Caldicellulosiruptor* spp., to grow on micro-crystalline cellulose (Blumer-Schuette *et al*. 2010). This inability to grow exists despite encoding other cellulose degrading enzymes, which permit growth on the lesser crystalline, carboxymethylcellulose. Therefore, strategies targeted towards improving hydrolysis of crystalline cellulose or reducing the crystallinity of the cellulose feedstocks are important

considerations when using *Ca. hydrothermalis* 108, *Ca. kristjanssonii* I77R1B and *Ca. owensensis* OL for CBP biofuel production.

To date, studies coupling gene expression data with CAZyme content have been limited to a few thermophilic *Firmicutes* (Table 1.3). However, these approaches have helped shed significant insight into the realized hydrolytic potential of these strains. Secretome analysis of *C. thermocellum* identifies that, independent of growth substrate, the expression of both cellulose acting, as well as hemicellulose acting, CAZymes occurs simultaneously (Table 1.3). Similar findings were also observed in cellulose grown cultures of *Caldicellulosiruptor bescii* DSM 6725 and *Caldicellulosiruptor obsidiansis* OB47 in which, only a few CAZyme-related proteins were observed in the secretome (Lochner *et al*., 2011a, b). The expression of diversely acting enzymes in these strains may have evolved as a means of coordinating hydrolysis of distinct lignocellulose polymers (cellulose and hemicellulose) between enzyme complexes. This inference is supported by the study of Blumer-Schuette *et al*., (2010), which found that the Avicel-induced secretomes of *Caldicellusiruptor* spp. capable of degrading micro-crystalline cellulose closely matched the xylose-induced secretomes of the same species. In contrast, the strains with reduced growth on micro-crystalline cellulose, *Ca. hydrothermalis* 108, *Ca. kristjanssonii* I77R1B and *Ca. owensensis* OL, had more pronounced differences in their Avicel- vs. xylose-induced secretomes. It is also interesting to note that the potential coordination of divergently functioning CAZymes observed in *Caldicellulosiruptor* spp. mimics the expression profiles of *C. thermocellum* despite apparently secreting far fewer GHs (Table 1.3).

Table 1.3. Observed[a] GH encoding genes[b] involved in lignocellulose hydrolysis in select thermophilic *Firmicutes* with available gene expression data.

GH	*Ca. bescii* DSM 6725	*Ca. obsidiansis* OB47	*Ca. saccharolyticus* DSM8903	*C. thermocellum* ATCC 27405
5	Athe_0594 [A]		Csac_0137 [G, X, XG, XGO, SWG, PO]	Cthe_0536 [C,C+XY,C+P,C+P+XY,CB,Z]
	Athe_1859 [A, XY]		Csac_0678 [G, X, XG, XGO, SWG, PO]	Cthe_0797 [C+XY,C+P,C+P+XY,SWG]
	Athe_1866 [A, XY]		Csac_1077 [G, X, XG, XGO, SWG, PO]	Cthe_0821 [C,C+XY,C+P,C+P+XY,SWG,Z]
			Csac_1078 [G, X, XG, XGO, SWG, PO]	Cthe_1472 [C,C+XY,C+P+XY,SWG,CB,Z]
				Cthe_2147 [C,C+XY,C+P,C+P+XY,SWG,CB]
				Cthe_2193 [C,C+XY,C+P,Z]
				Cthe_2872 [C,C+XY,C+P,C+P+XY,SWG,Z]
8				Cthe_0269 [C,C+XY]
9	Athe_0183 [A]	COB47_1669 [A, CB]	Csac_1076 [G, X, XG, XGO, SWG, PO]	Cthe_0412 [C,C+XY,C+P,C+P+XY,SWG]
	Athe_0618 [A]	COB47_1673 [A, CB]	Csac_1079 [G, X, XG, XGO, SWG, PO]	Cthe_0413 [C,C+XY,C+P+XY,CB,SWG]
	Athe_1865 [A, XY]			Cthe_0433 [C,C+XY,C+P,C+P+XY,SWG,CB,Z]
	Athe_1867 [A, XY]			Cthe_0578 [C,C+XY,C+P,SWG]
				Cthe_0624 [C,C+XY,CP,C+P+XY,SWG,Z]
				Cthe_0745 [C,C+XY,C+P,C+P+XY,SWG]
				Cthe_0825 [C,C+XY,CP,C+P+XY,CB,SWG]
				Cthe_2760 [C,C+XY,C+P,C+P+XY,SWG,Z]
				Cthe_2761 [C,C+XY,C+P,C+P+XY,SWG,Z]
10	Athe_1857 [A, XY]	COB47_1671 [A, CB]	Csac_0204 [G, X, XG, XGO, SWG, PO]	Cthe_0912 [C,C+XY,C+P,C+P+XY,CB]
			Csac_1078 [G, X, XG, XGO, SWG, PO]	Cthe_1838 [C,C+XY,C+P,C+P+XY,SWG,CB,Z]
			Csac_2405 [G, X, XG, XGO, SWG, PO]	Cthe_1963 [C,C+XY,C+P,C+P+XY,CB,Z]
			Csac_2408 [G, X, XG, XGO, SWG, PO]	Cthe_2590 [C,C+XY,SWG,CB,Z]

Table 1.3 cont.

GH	*Ca. bescii* DSM 6725	*Ca. obsidiansis* OB47	*Ca. saccharolyticus* DSM8903	*C. thermocellum* ATCC 27405
10			Csac_2410 [G, X, XG, XGO, SWG, PO]	
11	Athe_0089 [A]			Cthe_2972 [C,C+XY,C+P,C+P+XY,Z]
27				Cthe_0032 [C,C+XY,C+P,C+P+XY,SWG,CB,Z]
				Cthe_1472 [C,C+XY,C+P+XY,SWG,CB,Z]
				Cthe_2811 [C,C+XY,C+P,C+P+XY,SWG]
31				Cthe_2139 [C+XY]
				Cthe_3012 [C,C+P,C+P+XY,CB]
42				Cthe_0661 [C]
44				Cthe_1271 [C,C+P,C+P+XY,CB,Z]
48	Athe_1867 [A, XY]	COB47_1664 [A]	Csac_1076 [G, X, XG, XGO, SWG, PO]	
		COB47_1673 [A, CB]		
51				Cthe_2089 [C,C+XY,C+P,C+P+XY,SWG]
54				Cthe_1400 [C,C+XY,C+P,C+P+XY,CB,Z]
74			Csac_1085 [G, X, XG, XGO, SWG, PO]	
Ref.	Dam *et al*., 2011; Lochner *et al*., 2011a	Lochner *et al*., 2011a; Lochner *et al*., 2011b	VanFossen *et al*., 2011	Raman *et al*., 2009

[a]Observed through either transcriptomic or proteomic analysis during growth on a specific substrate. Substrates identified as superscripts as follows: A = Avicel; C = cellulose; CB = cellobiose; G = glucose; P = pectin; PO = poplar; S = wheat straw; SWG = switchgrass; X = xylose; XG = xyloglucan; XGO = xylogluco-oligosaccharides; XY = xylan; Z = Z-trim (60% cellulose + 16% hemicellulose). Growth conditions differentiated by use of commas and in some cases, multiple substrates compose a single growth condition (eg. C+XY = cellulose plus xylan).

[b]Only locus tags predicted to be located extracellularly included. Subcellular localization predicted using PSortB 3.0 (Yu *et al*., 2010) as in Table 1. Additional sequences within specific CAZy classes exist, though were not identified in the referenced literature.

The production of a cocktail of divergently functioning enzymes, rather than enzymes tailored to a specific function, may serve as a model for designing efficient CBP microorganisms, particularly in the development of designer co-cultures. This is part of the logic under which a *C. thermocellum* ATCC 27405 - *Ca. bescii* DSM 6725 co-culture was built and tested (Kridelbaugh *et al*., 2013). In that study, co-cultures grown on switchgrass or cellulose plus xylan (a major component of some hemicelluloses) showed improved hydrolysis over *C. thermocellum* ATCC 27405 monocultures, which was directly attributed to the xylanolytic capabilities of *Ca. bescii* DSM 6725. Coupling potential synergistic hydrolysis patterns between divergent organisms based on observed gene expression profiles is a promising strategy to improve hydrolysis.

Underlying differences in expression profiles and hydrolytic potential observed in biofuel-relevant *Firmicutes* are complex regulatory networks; the nature of which, have only begun to be elucidated. Identifying these mechanisms is further complicated by the realization that significant differences exist amongst thermophilic *Firmicutes*. However, understanding and manipulating these networks may have significant implications for the industrial implementation of a strain in mono- or co-culture.

For example, quantitative proteomic analysis of *C. thermocellum* ATCC 27405 by Raman *et al*. (2009) identified that xylanase expression decreased when grown on pretreated switchgrass, which contains xylan, in comparison to growth on pure cellulose (Table 1.3). Thus, xylan hydrolysis does not seem to be connected to the presence of xylan for *C. thermocellum* ATCC 27405 and its subsequent hydrolysis may be a limiting factor when using *C. thermocellum* on raw substrates. In contrast, increased expression of exo- and endo-glucanases by *C. thermocellum* ATCC 27405 was observed when

grown on polymeric cellulose in comparison to the disaccharide cellobiose (Raman *et al*., 2009) suggesting cellulase production is connected with the presence of cellulose and not its hydrolysis products.

Further, the *celC* cellulase operon in *C. thermocellum* is reported to be under the regulatory control of the dissacharide laminaribiose (Newcomb *et al*., 2007), while the study by Raman *et al*. (2009) has suggested that pectin may play a role in *C. thermocellum* xylanase expression. In these cases, regulation does not seem to be connected to the carbohydrate upon which the enzymes produced act.

Contrary to *C. thermocellum* ATCC 27405, secretome exo- and endo-glucanase expression levels were higher in *Ca. obsidiansis* OB47 on cellobiose grown cells in comparison to cellulose grown cells (Lochner *et al*., 2011b). Therefore, at least in the case of *Ca. obsidiansis* OB47, cellulose hydrolysis products seem to induce, or alternatively prevent repression, of cellulase encoding sequences.

The complex nature of lignocellulose requires that divergently functioning enzymes show coordinated activity in hydrolysis of the biomass. Understanding how expression of divergent GHs is coordinated within these, as well as lesser characterized organisms, is an important component in their development as industry-relevant strains. Insights gained from global expression profiling may be invaluable to identify specific limitations and lead to targeted strategies to resolve them.

1.3.3 Fermentation of lignocellulose hydrolysis products

Increasing overall biofuel yield and improving the viability of industrial lignocellulosic biofuel production requires the microbial conversion of both cellulose and

hemicellulose constituent saccharides into fuel products (Galbe *et al*., 2007; Gírio *et al*., 2010). Additionally, as hemicellulose is more readily hydrolyzed into its constituent sugars than cellulose, using a feedstock with higher hemicellulose content may be preferable for producing sugars that can be converted into fuels (Reddy & Yang, 2005). This highlights the importance of diversity in the substrate utilization potential of microorganisms being developed for CBP platforms.

The ability of many thermophilic *Firmicutes* to ferment diverse saccharides, in mono-, di- or polymeric form, is advantageous over traditional yeast-based ethanologenesis strategies (Taylor *et al*., 2008). However, this capacity varies widely amongst these strains (Table 1.4). *C. thermocellum* ATCC 27405 contains an extensive suite of diverse hydrolytic enzymes (Table 1.2), but also has one of the most restrictive substrate utilization profiles. Of the major carbohydrate components in lignocellulose, *C. thermocellum* is restricted to cellulose and its hydrolysis products (with the exception of glucose; Ng & Zeikus, 1982; Zhang & Lynd, 2005), and is unable to catabolise any hemicellulose constituent sugar components (Garcia-Martinez *et al*., 1980).

Conversely, many strains of *Caldicellulosiruptor*, as well as *Clostridium stercorarium* DSM 8532, have been reported to catabolise all of the major carbohydrate components (Table 1.4) potentially allowing for improved biomass conversion efficiencies. Other strains, belonging to either the *Thermoanaerobacter* or *Thermoanaerobacterium* genera, also have diverse potential, but are principally restricted from use in CBP monocultures by their inabilities to hydrolyze cellulose.

The hydrolysis of cellulose and hemi-cellulose fractions in a CBP system will generate a pool of mixed saccharides available for fermentation. Ideally, simultaneous

Table 1.4. Reported substrate utilization capabilities of major lignocellulose constituent saccharides by select biofuel producing thermophilic *Firmicutes*.

	Lignocellulose constituent saccharides								
	Cellulose & hydrolysis products[a]			Xylan & hemicellulose hydrolysis products[b]					
Strain	C	Cb	Gl	Ar	Xn	Xy	Ga	Mn	Reference
Cal. subterraneus subsp. *tengcongensis* MB4	-	+	+	-	-	-	+	+	Xue *et al*., 2001
Ca. bescii DSM 6725	+	+	+	+	+	+	+	+	Yang *et al*., 2010
Ca. kristjansonnii I77R1B	+	+	+	-	+	+	+	+	Bredholt *et al*., 1999
Ca. obsidiansis OB47	+	+	+	+	+	+	+	+	Hamilton-Brehm *et al*., 2010
Ca. saccharolyticus DSM 8903	+	+	+	+	+	+	+	+	Rainey *et al*., 1994; de Vrije *et al*., 2010
C. stercorarium DSM 8532	+	+	+	+	+	+	+	+	Madden, 1983; Adelsberger *et al.*, 2004
C. thermocellum ATCC 27405	+	+	-	-	-	-	-	-	McBee, 1954; Garcia-Martinez *et al*., 1980
G. glucosidasius NCIMB 11955	-	+	+	-	NR	+	-	+	Nazina *et al*., 2001; Cripps *et al*., 2009; Coorevits *et al*., 2012
G. stearothermophilus DSM 22	+	-	+	+	NR	+	-	+	Caccamo *et al*., 2000; Nazina *et al*., 2001; Tai *et al*., 2004; Coorevits *et al*., 2012
T. mathranii subsp. *mathranii* A3	-	+	+	+	+	+	+	+	Larsen *et al*., 1997
T. pseudethanolicus 39E	-	+	+	-	-	+	+	+	Onyenwoke *et al*., 2007
Thermoanaerobacter sp. X514	-	+	+	-	-	+	NR	NR	Roh *et al*., 2002; Hemme *et al*., 2011; Lin *et al*., 2011
Th. saccharolyticum B6A-R1	NR	+	+	NR	+	+	+	+	Lee *et al*., 1993
Th. thermosaccharolyticum DSM 571	-	+	+	+	NR	+	+	+	McClung *et al*., 1935; Hollaus & Sleytr, 1972
Th. xylanolyticum LX-11	-	+	+	+	+	NR	+	+	Lee *et al*., 1993

[a]Cellulose hydrolysis products have been limited to cellobiose and glucose and does not include longer chain length cellodextrins.

[b]Hemicellulose constituent saccharides as described (Saha, 2001; Shallom & Shoham, 2003). Also include glucose as well as glucuronic acids. No reports for the utilization of glucuronic acid by any of the listed strains were identified.

Symbols denote substrate utilization (+), no substrate utilization (-) or sugar-strain combination has not yet been reported (NR). Abbreviations: C = cellulose; Cb = cellobiose; Gl = glucose; Ar = arabinose; Xn = xylan; Xy = xylose; Ga = galactose; Mn = mannose.

conversion of the resultant hydrolysis products into biofuels will occur with no distinct preference for one substrate over another. However, carbon catabolite repression (CCR), whereby the presence of one carbon source exerts a regulatory effect on the expression of genes and gene products associated with the utilization of alternative carbon sources (Brückner & Titgemeyer, 2002), may permit only sequential, and not simultaneous, utilization.

CCR has been reported in many *Firmicutes*, including biofuel-relevant strains such as *C. cellulolyticum* H10 (Abdou *et al*., 2008) and *Thermoanaerobacterium saccharolyticum* M2476 (Tsakraklides *et al*., 2012). In contrast, *Firmicutes* such as *Ca. saccharolyticus* DSM 8903 (van de Werken *et al*., 2008; VanFossen *et al*., 2009) and *Thermoanaerobacter* sp. X514 (Lin *et al*., 2011) have been shown to simultaneously co-utilize lignocellulose relevant saccharides, suggesting a lack of CCR mechanisms in these organisms. In others, such as *Thermoanaerobacter pseudethanolicus* 39E, glucose has been shown to have a repressive effect on maltodextrin utilization (Hyun *et al*., 1985), but does not inhibit xylose utilization or synthesis of xylose related gene products (Erbeznik *et al*., 1998a; Jones *et al*., 2002a). Thus, in these organisms, understanding the specific influences of the CCR regulon is important in understanding carbon flux pathways and potential limitations in substrate usage.

Within biofuel-relevant *Firmicutes*, biochemical confirmation of CCR-relevant proteins has not been widely investigated, and this is further reduced when considering only thermophilic strains. However, current models for CCR in Gram positive organisms have been well established in other members of the phylum and gene orthology amongst

members suggests similar models may exist in biofuel producing strains (Tracy *et al*., 2012).

In *Firmicutes*, mechanisms of CCR include: i) inducer exclusion; ii) global regulatory control; and iii) specific transcriptional control (for more in-depth reviews of these mechanisms see Brückner & Titgemeyer, 2002; Warner & Lolkema, 2003; Deutscher, 2008; Görke & Stülke, 2008). Central to all of these processes is a single protein, HPr (histidine-containing protein) belonging to protein family (pfam) classification pfam00381 (Comas *et al*., 2008). The HPr protein is multi-functional in *Firmicutes* and its function *in vivo* is dependent on its phosphorylation state. When phosphorylated at the His_{15} residue by Enzyme I, P-His-HPr acts as a phosphor-carrier protein involved with phosphotransferase system (PTS) mediated transport. The P-His-HPr transfers its phosphate to a PTS-EIIA protein, which ultimately donates the phosphate to the sugar being transported. Alternatively, when phosphorylated at a conserved Ser_{46} residue, the HPr protein plays multiple roles in CCR (discussed below).

A functionally similar HPr-like protein termed crh (catabolite repression HPr) is also known to exist in some *Firmicutes* and plays a role in CCR similar to P-Ser-HPr (Galinier *et al*., 1997; Schumacher *et al*., 2006). Like HPr, crh also contains a conserved Ser_{46} residue that, when phosphorylated, transforms it into an effector molecule for CCR regulatory proteins. Unlike HPr though, crh contains no His_{15} residue and it is not involved in PTS-mediated transport.

Phosphorylation at the Ser_{46} residue is catalyzed by a bidirectionl HPr kinase/ phosphatase (HPrK/P) (pfam07475) (Comas *et al*., 2008) whose activity is modulated by the allosteric activator fructose-1,6-bisphosphate (FBP). When activated, the HPrK/P

phosphorylates the Ser_{46} residue of HPr or crh through the consumption of an ATP. Thus, when intracellular FBP concentrations, as well as ATP levels, are high (indicative of the rapid metabolism of a preferred carbon source), P-Ser-HPr and/or P-Ser-crh are readily formed and a CCR effect is observed. Upon depletion of the carbon source, intracellular FBP and ATP levels decline and the HPrK/P dephoshorylates P-Ser-HPr alleviating the CCR effect.

P-Ser-HPr is also involved with inducer exclusion as a CCR mechanism. The protein interacts with transport permeases preventing transport of alternative carbon sources, which may serve as inducing molecules for their own catabolism (Ye *et al*., 1995). This mechanism has been demonstrated *in vivo* in *Lactobacillus casei* (Viana *et al*., 2000) whereby the addition of glucose immediately arrests maltose uptake in maltose growing cells. A similar effect was observed in a *L. casei* catabolite control protein A (ccpA) mutant suggesting that the observed CCR was not ccpA-dependent. The CCR effect was not observed when glucose was added to maltose growing cells containing a HPr mutation (Ser-46-Ala) showing that P-Ser-HPr plays a direct role in maltose transport. In *Lactobacillus brevis*, it was additionally noted the P-Ser-HPr bound to inside-out membrane vesicles containing lactose permease protein (Ye *et al*., 1995). Therefore, further evidence was provided that the P-Ser-HPr dependent allosteric regulation of specific transport permeases is involved with inducer exclusion.

In *Firmicutes*, P-Ser-HPr or P-Ser-crh are also effector molecules for the global transcriptional regulator ccpA, a *lac*I/*gal*R-family transcriptional repressor protein (Henkin, 1996). The binding of either P-Ser-HPr (Schumacher *et al*., 2004) or P-Ser-crH (Schumacher *et al*., 2006; Galinier *et al*., 1999) induces a conformational change to ccpA

allowing it to bind consensus catabolite responsive element (*cre*) operator sequences and repress transcriptional activity.

CCR-related specific transcriptional control mechanisms can be mediated through antiterminator proteins, which contain PTS-regulatory domains (PRD) (pfam00874) (Stülke *et al*., 1998). One example is the licT protein in *B. subtilis*. In the absence of β-glucosides, which are transported via PTS-mediated mechanisms in *B. subtilis*, licT is phosphorylated by the β-glucoside-specific EIIB protein of the PTS complex (Görke & Stülke, 2008). The site of phosphorylation is domain-1 of the licT PRD and when phosphorylated, licT is unable to dimerize and is inactive. When β-glucosides are present, domain-1 donates its phosphate back to the EIIB protein. At the same time, P-His-HPr phosphorylates domain-2 of the licT PRD. This allows licT to dimerize and activates its antitermination activity allowing transcription of the *bglPH* operon, whose gene products are involved with β-glucoside consumption (Le Coq *et al*., 1995). While this is a more specific mode of regulation than ccpA-dependent regulation, it can also allow for the preferential use of PTS-sugars (Görke & Stülke, 2008).

Understanding CCR in lignocellulosic biofuel producing microorganisms is an important component in developing strategies towards maximizing carbohydrate conversion efficiencies in these organisms. This is particularly true in designer co-culture strategies whereby the complexities associated with understanding sugar usage regulons in a single organism are further magnified by potential inter-strain interactions that develop in co-cultures. In such scenarios, the hydrolysis products of a single strain may influence the carbon metabolism profile of the co-culture partner differently than the innate hydrolysis capabilities found in the co-culture partner itself. For example, when

grown in monoculture, *C. thermocellum* is known to hydrolyze xylan, but its inability to utilize the resulting products allows for the accumulation of xylo-dextrins in the fermentation medium (Ng *et al*., 1981). The free xylose or xylodextrins are available for use by an appropriate co-culture partner. Their usage though may be dependent on the simultaneous availability of cellulose hydrolysis products (also generated through *C. thermocellum* mediated hydrolysis), which may exert a CCR effect on the co-culture partner preventing xylose utilization. Therefore, understanding sugar usage preferences, and coupling this knowledge with an understanding of the lignocellulose hydrolysis patterns of strains in co-culture, may provide valuable insights into strain selection for co-culture design.

While simple growth studies can be used to identify sugar utilization preferences in microorganisms, genomic and expression profiling facilitates the identification of potential CCR regulons. Further, the identification of *hpr* and *crh* homologs (and whether or not they are expressed under specific growth conditions) can provide insights into molecular engineering targets for purposes of alleviating CCR effects. However, it is important to note that while gene deletion of *crh* is a suitable engineering strategy, the multiple roles of HPR *in vivo* does not allow a similar strategy for mitigating the CCR effects of HPr. HPr knockout strains may alleviate CCR, but also simultaneously lose the ability to transport sugar through PTS-mediated mechanisms. Gene mutation strategies, whereby the Ser_{46} residue is mutated to prevent phosphorylation of HPr, have provided successful alternative approaches to remove CCR effects (Viana *et al*., 2000; Tsakraklides *et al*., 2012).

The co-utilization of diverse lignocellulose constituent saccharides is an important criterion to evaluate when selecting CBP-relevant microorganisms. However, strains with this ability, coupled with an extensive hydrolytic capacity, have not yet been identified. While such strains may exist and should be pursued, developing synergistic co-cultures may be valuable given the difficulties associated with genetic engineering of biofuel-relevant thermophiles (Chapter 1.3.5).

1.3.4 Biofuel production

Selecting microorganisms capable of achieving high biofuel titres, with the potential to approach theoretical maxima (e.g. 2 mol ethanol/mol glucose; 4 mol H_2/mol glucose; Carere *et al*., 2012), is as important as identifying strains capable of efficient lignocellulose hydrolysis. While under certain conditions, some CBP-relevant strains are capable of near maximal biofuel production (Table 1.5), the native fermentative pathways in these strains have evolved to also produce non-biofuel end-products (Fig. 1.2). While undesirable for biofuel production purposes, these branched pathways play important physiological roles in balancing global metabolites such as ATP/ADP, AMP/PPi, NAD^+/NADH, $NADP^+$/NADPH and acyl-CoAs (Lee *et al*., 2008; Mukhopadhyay *et al*., 2008). Therefore, identifying strategies that maximize biofuel production, while avoiding disruptions to globally connected processes, necessitates a thorough understanding of core metabolic pathways in developing platform organisms.

A recent study of genes involved with core metabolism in numerous CBP-relevant microorganisms correlated genome content, in the context of thermodynamics,

Table 1.5. Reported end-products and molar end-product yields for select thermophilic *Firmicutes* of interest for lignocellulosic biofuel production.

	End-products (mol product/mol hexose equivalent)										
Strain	H_2	CO_2	Ac	Bu	Bt	Et	Fo	La	Substrate	Type	Reference
Cal. subterraneus subsp. t*engcongensis* MB4	0.30	1.50	1.00	-	-	0.70	-	-	4 g/L glucose	B	Xue *et al*., 2001
	2.80	+	1.40	-	-	0.60	-	-	10 g/L starch	B	Soboh *et al*., 2004
	4.00	+	2.00	-	-	0.02	-	-	4.5 g/L glucose	S,H	Soboh *et al*., 2004
Ca. bescii DSM 6725	1.38	-	0.96	-	-	-	-	0.72	5 g/L cellulose[a]	B	Yang *et al*., 2009
Ca.kristjansonnii I77R1B	+	+	1.44	-	-	0.10	0.05	0.28	2 g/L avicel[b]	B	Bredholt *et al*., 1999
Ca. obsidiansis OB47	-	+	0.62	-	-	0.03	-	0.59	15 g/L avicel[c]	P	Hamilton-Brehm *et al*, 2010
	-	+	1.21	-	-	0.07	-	-	10 g/L switchgrass[c, d]	P	Hamilton-Brehm *et al*, 2010
Ca. saccharolyticus DSM 8903	3.15	1.71	1.71	-	-	-	-	0.11	10 g/L sucrose	P,H	van Niel *et al*., 2002
	3.60	1.50	1.60	-	-	-	-	-	4.4 g/L glucose[e]	C	de Vrije *et al*., 2007
	3.40	1.80	1.60	-	-	-	-	-	7 g/L glucose + 3 g/L xylose	P,H	de Vrije *et al*., 2009
C. stercorarium DSM 8532	+	+	+	-	-	+	-	+	Glucose[f]	B	Madden, 1983
C. thermocellum ATCC 27405	0.98	0.83	0.79	-	-	0.56	0.42	0.36	4.5 g/L cellobiose	B	Islam *et al*., 2006
	0.82	1.06	0.66	-	-	0.76	0.33	0.02	1.1 g/L cellobiose	B	Rydzak *et al*., 2009
	1.24	0.98	0.52	-	-	0.56	0.30	-	2 g/L cellobiose[a]	B	Rydzak *et al*.,

2012

Table 1.5 cont.

Strain	H_2	CO_2	Ac	Bu	Bt	Et	Fo	La	Substrate	Type	Reference
G. thermoglucosidasius NCIMB 11955	-	-	0.17	-	-	0.39	0.15	1.20	30 g/L glucose	P	Cripps *et al*., 2009
T. italicus Ab9	+	+	+	-	-	+	-	+	5 g/L glucose	B	Kozianowski *et al*., 1997
T. mathranii subsp. *mathranii* A3	1.08	2.17	0.48	-	-	1.32	-	0.07	2 g/L xylose	B	Larsen *et al*., 1997
T. pseudethanolicus 39E	-	1.63	0.08	-	-	1.55	-	0.23	4 g/L glucose	B	Lovitt *et al*., 1984
	0.45	1.72	0.10	-	-	1.41	-	0.15	10 g/L glucose	B	Lovitt *et al*., 1988
	-	-	0.24	-	-	1.25	-	0.32	5 g/L xylose	B	He *et al*., 2009
	-	-	0.17	-	-	1.10	-	0.35	5 g/L glucose	B	He *et al*., 2009
	-	+	0.38	-	-	1.11	-	0.30	Glucose[f]	B	Hemme *et al*., 2011
	-	+	0.29	-	-	1.52	-	0.10	2 g/L xylose	B	Hemme *et al*., 2011
Thermoanaerobacter sp. X514	-	+	0.33	-	-	1.25	-	0.17	2.1 g/L glucose[a]	B	Feng *et al*., 2009
	-	+	0.51	-	-	1.02	-	0.29	Glucose[f]	B	Hemme *et al*., 2011
	-	+	0.25	-	-	1.55	-	0.12	2 g/L xylose	B	Hemme *et al*., 2011
Th. saccharolyticum JW/SL-YS485	+	+	0.56	-	-	1.32	-	0.29	5 g/L glucose	B	Desai *et al*., 2004
	+	+	0.47	-	-	1.16	-	0.06	5 g/L xylose	B	Desai *et al*., 2004
	1.06	+	0.50	-	-	1.45	-	0.04	4 g/L xylose	B	Shaw *et al*., 2008a
	1.08	+	0.52	-	-	1.49	-	0.04	4 g/L xylose	B	Shaw *et al*., 2008b

Table 1.5 cont.

Strain	H_2	CO_2	Ac	Bu	Bt	Et	Fo	La	Substrate	Type	Reference
Th. thermosaccharolyticum DSM 571	2.30	1.76	0.49	-	-	0.59	-	0.26	10 g/L glucose	B	Sjolander, 1937
	-	+	0.50	0.03	0.27	0.22	-	0.19	20 g/L glucose	B	Freier-Schroder *et al*., 1989
	-	+	0.39	0.03	0.43	0.24	-	0.09	20 g/L starch	B	Freier-Schroder *et al*., 1989
	-	-	0.43	0.03	0.60	0.31	-	0.03	10 g/L cellobiose	B	Bhandiwad *et al*., 2013
Th. xylanolyticum LX-11	+	+	+	-	-	+	-	-	NR	B	Lee *et al*., 1993

[a]Substrate utilization and end-product data used for calculations estimated from graphical data. End-point values used for calculations.

[b]Extent of substrate consumption not reported. For calculations, total consumption of substrate was assumed.

[c]Ethanol also concentrated and reported in reactor condensors, but omitted in calculations as values would exceed theoretical limits.

[d]Substrate consumption calculated using the molecular weight of glucose only.

[e]Dilution rate of 0.01 h^{-1} used.

[f]Concentration of substrate used not reported.

Abbreviations and symbols: (Ac) acetate; (Bu) butanol; (Bt) butyrate; (Et) ethanol; (Fo) formate; (La) lactate; (B) batch culture; (S) stirred batch culture; (H) sparged headspace; (P) pH-controlled batch culture; (C) continuous culture; (NR) specific growth substrate not reported (+) end-product observed, but not quantified; (-) end-product not observed or reported.

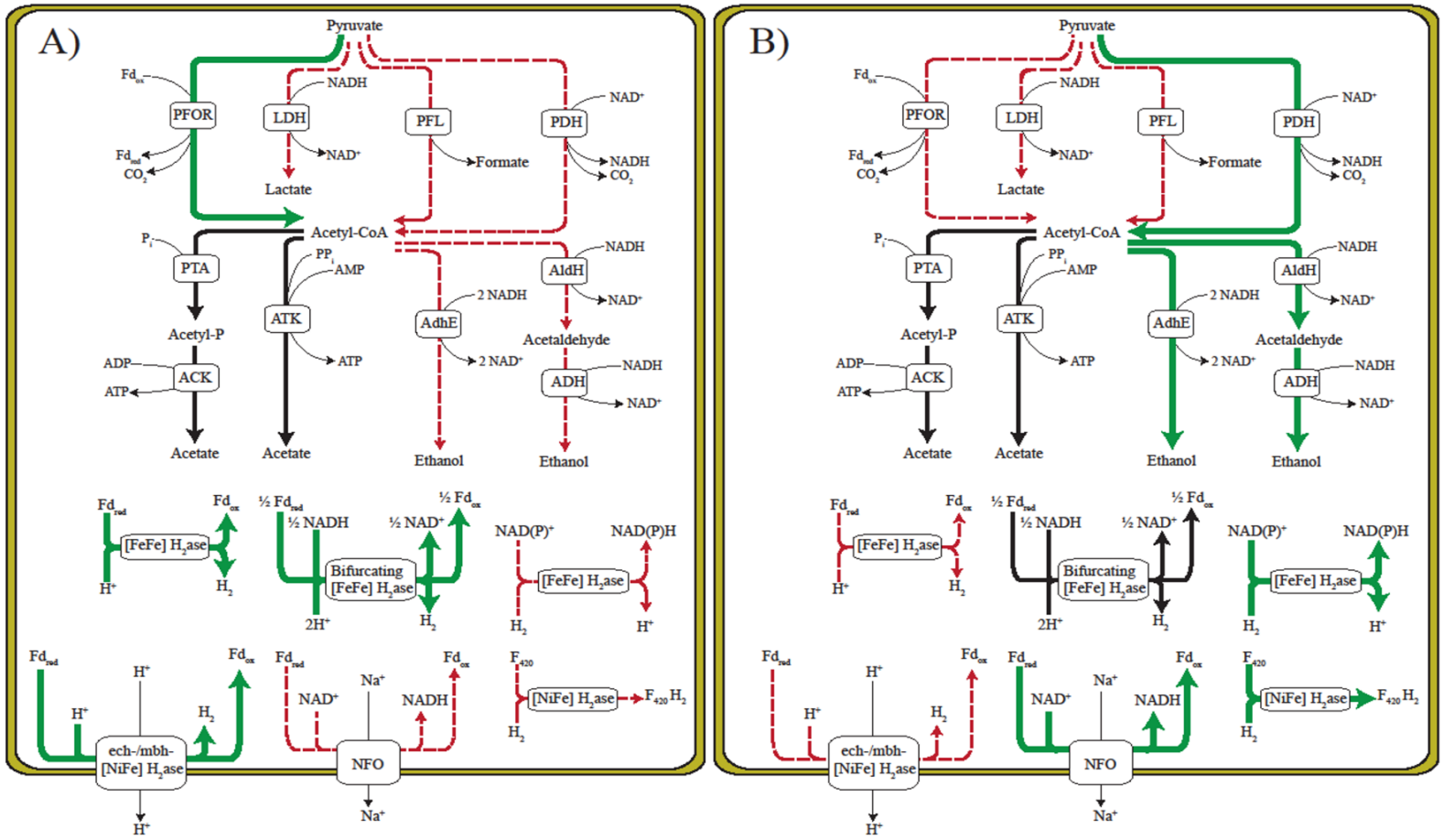


Fig 1.2. Pathways of pyruvate catabolism that favour hydrogen or ethanol production based on comparative analysis of select microorganisms (Carere *et al*., 2012). (A) Favour hydrogen. (B) Favour ethanol. Pathways that favour (green lines), disfavour (broken red lines), and appear to have little impact (black lines) on production of H_2 or ethanol are indicated. Abbreviations are as defined (p. xxv). Figure reproduced with permission: © 2012, Carere CR, University of Manitoba.

with reported biofuel yields (Carere *et al*., 2012). In that study, maintaining metabolic balances is identified as a key component that influences the type and amount of biofuel produced.

For example, Carere *et al*., (2012) correlate the ethanologenic capabilities of *T. pseudethanolicus* 39E, which has one of the highest native molar ethanol yields reported (Table 1.5; Lovitt *et al*., 1988; Hemme *et al*., 2011), with the presence of an NADH: ferredoxin oxidoreductase (*nfo*), an NAD(P)H-consuming bifunctional acetaldehyde: alcohol dehydrogenase (*adhE*) and the absence of a discernible ferredoxin- (Fd) or NAD(P)H-dependent hydrogenases (though a potential bifurcating Fd: NAD(P)H hydrogenase is identified). Using these genes, the potential for near optimal ethanol production is proposed based on the consumption of reduced ferredoxin (Fd_{red}). The generation of Fd_{red} by pyruvate: ferredoxin oxidoreductase (POR) in *T. pseudethanolicus* 39E (Lovitt *et al*., 1988; Hemme *et al*.., 2013), may be preferentially consumed by NFO (producing NAD(P)H), rather than being consumed by a hydrogenase. The *adhE* gene product, or alternative alcohol dehydrogenase encoding genes, would subsequently consume NAD(P)H leading to the formation of ethanol. In this scenario, maintaining low Fd_{red} levels, in combination with high NAD(P)H levels, could potentially explain the high ethanol yields observed (Fig. 1.2).

While genetic engineering targets/strategies for maximizing either H_2 or ethanol production in numerous strains are proposed (Carere *et al*., 2012), challenges still remain in that the genomes of many of these organisms encode redundancies in core metabolic genes. Further, many of the annotated genes lack sufficient characterization to confirm *in vivo* functionality. As such, global expression analyses (e.g. transcriptomics, proteomics)

is being increasingly investigated to help refine engineering targets as well as to elucidate connected physiological processes (Mukhopadhyay *et al.*, 2008).

Such approaches have recently been applied to *C. thermocellum* ATCC 27405 (Raman *et al.*, 2011; Rydzak *et al*., 2012), which is known to produce acetate, ethanol, formate and, under certain conditions, lactate, as major non-gaseous end-products (Table 1.5). Proteomic analysis of cellobiose grown *C. thermocellum* ATCC 27405 has shown that, while all four annotated iron-dependent alcohol dehydrogenases were synthesized, the AdhE encoding gene, which has been proposed to be a key enzyme for ethanol synthesis in *C. thermocellum* ATCC 27405 (Raman *et al.*, 2011), was the most abundantly expressed (Rydzak *et al.*, 2012). A prominent role for adhE in acetaldehyde formation, and potentially ethanol production, is further supported by the fact that a standalone acetaldehyde dehydrogenase was not observed. The same study correlated expression of the pfl-encoding gene with formate production, while both proteomic (Rydzak *et al.*, 2012) and transcriptomic (Raman *et al.*, 2011) analyses confirmed low expression of the lactate dehydrogenase (*ldh*) (Cthe_1053) correlating with no detectable lactate reported. Proteomic analysis was unable to discern based on relative expression values whether acetate production in *C. thermocellum* ATCC 27405 is likely catalyzed via a putative acetate thiokinase (Cthe_0551) or via a two-step phosphotransacetylase (*pta*) (Cthe_1029) – acetate kinase (*ack*) (Cthe_1028) reaction. The *pta-ack* process for acetate formation was still proposed though based on reported activities of these enzymes in *C. thermocellum* (Lamed & Zeikus, 1980a) as well as favorable thermodynamics using the *pta-ack* pathway (Rydzak *et al.*, 2012).

Based on these data, and the engineering strategies proposed (Carere *et al.*, 2012), preventing carbon flow through the proposed *pfl* or through *pta-ack* pathways (and potential deletion of *ldh* dependening on growth condition) could seemingly help direct carbon flux towards ethanol and improve overall yields in *C. thermocellum* ATCC 27405. However, *C. thermocellum* ATCC 27405 also reportedly catalyzes pyruvate oxidation through a POR enzyme(s) generating Fd_{red} (Carere *et al*., 2008; Rydzak *et al*., 2009). Directing total carbon flux through POR could potentially generate excess Fd_{red} for which, the proposed electron sinks are hydrogenases. Therefore, to improve ethanol production in *C. thermocellum* ATCC 27405 near theoretical maxima, it would require an alternative mechanism to dispose of Fd_{red}. Such approaches could include the heterologous expression of a pyruvate dehydrogenase complex (yielding NADH instead of Fd_{red} from pyruvate catabolism) as proposed (Fig. 1.2; Carere *et al.*, 2012) or a means of converting Fd_{red} to NAD(P)H (via an NFO type system) such as that found in *T. pseudethanolicus* 39E. It is interesting to note, that while the *C. thermocellum* ATCC 27405 genome encodes an NFO-type system, expression of all the constituent gene products have not been detected (Raman *et al*., 2011; Rydzak *et al.*, 2012). Deletion of the acetate and lactate forming pathways in *C. thermocellum* DSM 1313 did increase ethanol yields, but also led to the extrusion of pyruvate (Argyros *et al*., 2011), or increased amino acid expulsion (van der Veen *et al*., 2013) in the mutant strain. In both cases, the observed effects were attributed to metabolic imbalances.

Under certain conditions, near theoretical maximal hydrogen production have been reported for *Cal. subterraneus* subsp. *tengcongensis* MB4 (Table 1.5), which has also been investigated in many gene expression profiling studies (Wang *et al*., 2004;

Wang *et al*., 2007a; Meng *et al*., 2009; Chen *et al*., 2013; Tong *et al*., 2013). *Cal. subterraneus* subsp. *tengcongensis* MB4 also catalyzes pyruvate oxidation via POR (Wang *et al*., 2007a; Tong *et al*., 2013), but lacks obvious NFO-encoding genes within its genome to convert Fd_{red} to NAD(P)H (van de Werken *et al*., 2008). Both a Fd-dependent and a NADH-dependent hydrogenases have been characterized in the strain (Soboh *et al*., 2004). The NADH-dependent hydrogenase has also been recently proposed to be a potential bifurcating hydrogenase, which consumes both Fd_{red} and NAD(P)H (Schut & Adams, 2009; Carere *et al*., 2012). Recent quantitative proteomic evidence identifies that the potential bifurcating hydrogenase is most abundantly expressed in cultures grown at 55^{o}C and is expressed at decreasing levels as growth temperatures are elevated to 65^{o}C, 75^{o}C and 80^{o}C (Chen *et al*., 2013). This suggests that the Fd-dependent hydrogenase, which is also an energy conserving hydrogenase (Ech) and is expressed at 75^{o}C (Soboh *et al*., 2004, Wang *et al*., 2004), may play a greater role in hydrogen production at optimal growth temperatures (optimal growth is 75^{o}C for this strain; Xue *et al*., 2001). Unfortunately, relative expression levels of Ech were not reported (Chen *et al*., 2013).

Biofuel production is a by-product of core metabolism in many potential novel platform microorganisms. Given that many of the metabolites and co-factors involved with core metabolism have global implications on cellular physiology, improving biofuel yields through pathway manipulation is not a straightforward process. However, continuing to develop an understanding of the network connectivity will ultimately help in designing engineering approaches for successful strain improvements.

1.3.5 Genetically tractable microorganisms

The challenges associated with lignocellulose hydrolysis, biomass utilization and low biofuel yield has identified a need to find genetically tractable microorganisms in which, engineering strategies may be employed to address current physiological limitations. While amenability to genetic modification is not a readily evident trait that can be used in the screening of novel microorganisms, it is a criterion upon which characterized strains should be evaluated.

Many thermophilic CBP-relevant organisms have generally been considered to be recalcitrant to genetic modification (Taylor *et al*., 2009). Much of the basis for this has been attributed to the fact that often, the genetic tools developed to work with mesophilic organisms lack the thermostability needed for strains grown at elevated temperatures (Lin & Jian, 2013). However, recent advances in the development of these tools, allowing for the creation of more readily manipulated parent strains (Mai & Wiegel, 2000; Peng *et al*., 2006; Lin *et al.*, 2010; Tripathi *et al*., 2010; Chung *et al*., 2013a), has allowed for targeted engineering strategies to be employed with a few select organsims. Additionally, the realization that certain genera, such as *Thermoanaerobacter* and *Thermoanaerobacterium*, exhibit natural competence for the uptake of exogenous DNA, may even further accelerate their development into novel platform organisms (Shaw *et al*., 2010).

Most genetic engineering efforts to date have focused on central metabolism in attempts to improve overall biofuel yields (Table 1.6). This approach has been validated as many of the mutated strains are capable of producing biofuels approaching the theoretical maxima. While most strategies have tried to reduce or eliminate competing

end-products through gene deletion strategies (Table 1.6), others have tried to enhance production by over-expressing native pathways (Bandiwad *et al*., 2013). Increased expression of the native butanol synthesis pathways in *Th. thermosaccharolyticum* led to a 180% increase in total butanol yield. While butanol was still one of the least abundant end-products in the mutant strain, this initial attempt at genetically modifying *Th. thermosaccharolyticum* shows promise for its continued development as a thermophilic butanol producer.

Redirecting carbon flux may also have additional benefits apart from solely increasing biofuel yield. For example, engineered pathways in *Ca. bescii* to direct carbon solely to acetate not only improved H_2 yields, but also extended exponential phase and allowed cultures to reach higher cell densities (Cha *et al*., 2013). This effect was presumed attributable to the increased ATP levels associated with acetate production via the *pta-ack* pathway. While dense, efficient biofuel producing cultures are desirable, it is also important to note that increased acid production also leads to a need for pH-controlled systems for culturing biofuel producing microbes. The need for pH-control ultimately represents an undesirable operating cost (Fischer *et al*., 2008).

Addressing pH-control formed part of the basis for recent attempts by Shaw *et al*., (2012) to heterologously express a urease protein from *C. thermocellum* in *Th. saccharolyticum*. In that study, urea was investigated as a low cost nitrogen alternative to ammonia for use in lignocellulosic fermentations. Apart from cost reduction, the authors also hypothesized that medium acidification throughout growth could also be reduced through the use of urea. The basis for this rationale is that ammonia is typically in its protonated form in the pH-range suitable for *Th. saccharolyticum* growth. As such, it is

Table 1.6. Effect of genetic engineering[a] on biofuel yields in select thermophilic *Firmicutes*.

Species	Genetic modification	Substrate	Target biofuel	Biofuel yield (mol biofuel/ mol hexose equivalent)		Reference
				Parent strain	Mutant strain	
Ca. bescii	Δ*ldh*	Cellobiose	H_2	2.48	3.40	Cha *et al*., 2013[b]
C. thermocellum	Δ*pta*	Cellobiose	Ethanol	NC	NC	Tripathi *et al*., 2010
	Δ*pta*, Δ*ldh*	Cellulose	Ethanol	0.28	1.20	Argyros *et al*., 2011[b]
G. thermoglucosidasius	Δ*ldh*	Glucose	Ethanol	0.39	0.94	Cripps *et al*., 2009
	Δ*ldh*, Δ*pfl*	Glucose	Ethanol	0.39	1.10	Cripps *et al*., 2009
	Δ*ldh*, Δ*pfl*, *pdh*	Glucose	Ethanol	0.39	1.64	Cripps *et al*., 2009[c]
T. mathranii	Δ*ldh*	Glucose	Ethanol	1.15	1.80	Mikkelsen & Ahring, 2007[b]
	gldA	Glucose	Ethanol	1.31	1.68	Yao & Mikkelsen, 2010[d]
Th. aotearoense	Δ*ldh*	Xylose	H_2	0.88	1.74	Li *et al*., 2010b
Th. saccharolyticum	Δ*ldh*	Xylose	Ethanol	1.49	1.58	Shaw *et al*., 2008b
	Δ*pta*, Δ*ack*	Xylose	Ethanol	1.49	2.00	Shaw *et al*., 2008b
	Δ*pta*, Δ*ack*, Δ*ldh*	Xylose	Ethanol	1.49	2.07	Shaw *et al*., 2008b
	ureABCDEFG	Cellobiose	Ethanol	0.40	0.44	Shaw *et al*., 2012[e]
Th. thermosaccharolyticum	*Bcs*	Cellobiose	Butanol	0.03	0.09	Bandiwad *et al*., 2013[f]
	Δ*ldh*	Cellobiose	Butanol	0.03	0.03	Bandiwad *et al*., 2013
	Δ*adhE*	Cellobiose	Butanol	0.03	0.00	Bandiwad *et al*., 2013

[a]Table has been limited to studies involving targetted strategies and does not include random mutagenesis.

[b]Value estimated from graphical data (end-point values) or proportional increase reported.

[c]Increased expression of native *pdh*.

[d]Heterologous expression of *gldA* (NAD^+-dependent glyceraldehyde dehydrogenase) from *Thermotoga maritima*.

[e]Heterologous expression of a urease from *C. thermocellum*.

[f]Increased expression of the native *bcs* operon catalyzing butanol synthesis.

Abbreviations: (*ldh*) lactate dehydrogenase; (*pta*) phosphotransacetylase; (*pfl*) pyruvate formate lyase; (*pdh*) pyruvate dehydrogenase; (*ack*) acetate kinase; (*adhE*) bifunctional acetaldehyde-alcohol dehydrogenase; NC = not calculable due to missing substrate consumption data.

not readily diffusible across the cell membrane and *Th. saccharolyticum* must therefore uptake neutral ammonia ions (leaving behind protonated ammonium ions) resulting in medium acidification (Booth, 1985; Shaw *et al*., 2012). Urea however, is readily diffusible, but requires a urease for hydrolysis to ammonia (suitable for assimilation) and CO_2.

Expression of the urease not only permitted the use of urea as a nitrogen source for *Th. saccharolyticum*, but also increased overall ethanol titers (Table 1.6). It was proposed that the increased titers observed were due to the availability of a readily metabolized nitrogen source whose consumption did not result in concurrent medium acidification (Shaw *et al*., 2012). Its use also removed the need for base addition as a mechanism for pH-control in batch fermentations using the mutant *Th. saccharolyticum* strain.

Only a few attempts have been made so far at improving the hydrolytic ability of strains using targeted engineering strategies. A functional β-1,4-endoglucanase originating from *C. thermocellum* has been expressed in *Thermoanaerobacter* sp. X514 conferring cellulolytic ability on a native non-cellulolytic strain (Lin *et al*., 2010). Additionally, a modified version of *cipA*, the cellulosome anchoring protein from *C. thermocellum* (Kruus *et al*., 1995; Lynd *et al*., 2002), has been expressed in *Th. saccharolyticum* (Currie *et al*., 2013). When co-cultured with a Δ*cipA C. thermocellum* mutant, *Th. thermosaccharolyticum* was able to restore the formation of functional cellulosomes and improve cellulolytic activity near that observed for wild type co-cultures.

The improvement of CBP-relevant thermophiles through genetic engineering is in its infancy as tools to do so have only started to emerge in recent years. In addition to these advances, continuing to identify traits of genetically tractable microbes may help further progress engineering efforts. Recent evidence has shown that major barriers in the genetic modification of *C. thermocellum* (Guss *et al*., 2012) or *Ca. bescii* (Chung *et al*., 2012) are directly connected with their native restriction endonucleases. This has led to the use of comparative genomic analyses as a strategy to identify potential engineering barriers in many CBP-relevant *Firmicutes* (Carere, 2013; Chung *et al*., 2013b).

Consistent amongst strains that have been successfully engineered (Table 1.6) is the availability of genomic data from either the strain of interest, or a close phylogenetic relative. Genomics, particularly when coupled with expression datasets such as proteomics or transcriptomics, can provide powerful insights into global physiological phenomenon. While important insights are still routinely gained using non-"omics" approaches, the valuable role of "omics-based" strategies is increasingly evident for the continued development of novel monoculture, or designer co-cultures, for use in lignocellulosic biofuel production.

1.4 Thesis objectives

Microorganisms currently being developed for lignocellulosic biofuel production through CBP all have physiological limitations restricting their implementation as "industry-ready" strains. While distinct strategies for microorganism selection and development exist, identifying strains which natively possess a broad spectrum of biofuel producing capabilities is desirable. Using this logic, the broad objective of this research

was to identify novel microorganisms from a lignocellulosic community through bioprospecting and evaluate the CBP-relevant biofuel production potential of isolated strains. Newly identified strains may possess characteristic that are phenotypically advantageous for industrial implementation over current developing organisms, or they may help shed insight into the physiologies of currently investigated strains.

To achieve this, the following specific objectives were pursued:

1) To isolate and identify, from lignocellulose degrading environments, novel microorganisms possessing biofuel producing potential (Chapters 2 & 3)
2) To characterize and evaluate the newly isolated strains in relation to current developing platform organisms across multiple parameters including:
    a. Lignocellulose hydrolysis (Chapters 2 & 4)
    b. Substrate utilization (Chapters 2, 4 & 5)
    c. Biofuel production (Chapters 2, 4 & 5)
3) To assess strain limitations, and pursue physiological explanations for those limitations, through establishing links between observed metabolism, genome content and proteomic profiles (Chapter 4, Chapter 5)

1.5 Thesis outline

- **Chapter 1: Literature review**
    - **Submitted for publication:** Modified versions of section 1.3.2 and 1.3.3 are currently submitted for publication as part of a book chapter titled "Omics approaches for designing biofuel producing co-cultures for

enhanced microbial conversion of lignocellulosic substrates." (© 2013, David B. Levin, University of Manitoba)

- **Chapter 2: Isolates of *Thermoanaerobacter thermohydrosulfuricus* from decaying wood compost display genetic and phenotypic microdiversity**
    - **Published:** FEMS Microbiology Ecology. 2011. **78**: 473-487
    - **Summary:** The isolation, identification and physiological characterization of a microdiverse population of *Thermoanaerobacter thermohydrosulfuricus* strains from woodchip compost. Comparisons are made against select type strains from the genus *Thermoanaerobacter*.
- **Chapter 3: Predicting relatedness of bacterial genomes using the chaperonin-60 universal target (*cpn60* UT): Application to *Thermoanaerobacter* species**
    - **Published:** Systematic and Applied Microbiology. 2011. **34**: 171-179.
    - **Summary:** The development of a mathematical model for predicting genome relatedness between microorganisms based on sequence similarity of the *cpn60* UT region. Phylogenetic analysis of the genus *Thermoanaerobacter* and taxonomic assignment of strains isolated in Chapter 2 is provided.
- **Chapter 4: Genomic evaluation of *Thermoanaerobacter* spp. for the construction of designer co-cultures to improve lignocellulosic biofuel production**
    - **Published:** PLoS One. 2013. **8**: e59362.

- **Summary:** Comparative genomic analyses of sequenced *Thermoanaerobacter* spp. and assessment of genome content related to biomass hydrolysis, substrate utilization and biofuel production capabilities between strains. Includes analyses of a *de novo* genome from a novel isolate (Chapter 2).

- **Chapter 5: Metabolic and label-free quantitative proteomic analysis of *Thermoanaerobacter thermohydrosulfuricus* WC1 on single and mixed substrates**
- **Submitted for publication**
- **Summary:** Central metabolism of a novel isolate (Chapter 2) is investigated by linking observed fermentation end-products with label-free proteomic analyses. Expression of a carbon catabolite repression network, potentially influencing substrate utilization patterns, is investigated.
- **Chapter 6: Thesis conclusions and future perspectives**
- **Appendix**
- **Literature cited**

Chapter 2. Isolates of *Thermoanaerobacter thermohydrosulfuricus* from decaying wood compost display genetic and phenotypic microdiversity[7]

2.1 Abstract

In this study, 12 strains of *Thermoanaerobacter* were isolated from a single decaying wood compost sample and subjected to genetic and phenotypic profiling. The 16S rRNA encoding gene sequences suggested that the isolates were most similar to strains of either *Thermoanaerobacter pseudethanolicus* or *Thermoanaerobacter thermohydrosulfuricus*. Examination of the lesser conserved chaperonin-60 (*cpn60*) universal target showed that some isolates shared the highest sequence identity with *T. thermohydrosulfuricus*; however, others to *Thermoanaerobacter wiegelii* and *Thermoanaerobacter* sp. Rt8.G4 (formerly *Thermoanaerobacter brockii* Rt8.G4). BOX-PCR fingerprinting profiles identified differences in the banding patterns not only between the isolates and the reference strains, but also amongst the isolates themselves. To evaluate the extent these genetic differences were manifested phenotypically, the utilization patterns of 30 carbon substrates were examined and the niche overlap indices (NOI) calculated. Despite showing a high NOI (>0.9), significant differences existed in the substrate utilization capabilities of the isolates suggesting that either a high degree of niche specialization, or mechanisms allowing for non-competitive co-existence, were present within this ecological context. Growth studies showed that the isolates were physiologically distinct in both growth rate and fermentation product ratios. Our data

[7]A modified version of this chapter has previously been published. Reproduced with permission: © 2011, Federation of European Microbiological Societies. Found in: **Verbeke TJ, Dumonceaux TJ, Wushke S, Cicek N, Levin DB, Sparling R**. 2011. Isolates of *Thermoanaerobacter thermohydrosulfuricus* from decaying wood compost display genetic and phenotypic microdiversity. FEMS Microbiol Ecol **78**: 473—487.

indicate that phenotypic diversity exists within genetically microdiverse *Thermoanaerobacter* isolates from a common environment.

2.2 Introduction

Genomic and phenotypic heterogeneity within an operational taxonomic unit (OTU) is expanding our understanding of microbial diversity. Conventional diversity studies have often relied upon a ribosomal RNA gene sequence (*rrn*) based approach, but increasing evidence suggests that using the 16S rRNA-encoding gene can significantly underestimate (Staley, 2006; Hong *et al*., 2009) or overestimate (Acinas *et al*., 2004a, b) the true breadth of microbial diversity. Furthermore, clustering similar ribotypes into a single OTU fails to account for phenotypic diversity among ecotypes – physiologically distinct genetic lineages that occupy the same ecological niche and emerge through natural selection (Cohan, 2002; Rocap *et al*., 2003). Understanding the extent to which microdiversity exists in the microbial world dictates that a combined phenotypic and genetic approach be utilized.

Sub-species variation has previously been described as "microdiversity" (Schloter *et al*., 2000). It has been used to describe genotypic variation (Rocap *et al*., 2003; Brown & Fuhrman, 2004; Cuadros-Orellana *et al*., 2007), phenotypic variation (Moore *et al*., 1998; Haverkamp *et al*., 2009), or a combination of both (Jaspers & Overmann, 2004; Acinas *et al*., 2009; Peña *et al*., 2010; Vermette *et al*., 2010), among isolates that conventionally fall into a single OTU. Most of these studies investigating sub-species variation have involved the examination of strains from geographically widespread and heterogeneous environments with varying ecological pressures. The studies conducted by

Moore *et al*. (1998), Vermette *et al*. (2010), and Peña *et al*. (2010) report the co-occurrence of different ecotypes from a single environmental sample. Furthermore, it has been shown that the average distance between ecotypes inhabiting and exploiting distinct microenvironments may be greater than the average sequence divergence of the same strains (Ramsing *et al*., 2000; Vermette *et al*., 2010).

The co-occurrence of highly similar strains seemingly violates the traditional "one niche-one strain" philosophy in which two strains cannot occupy the same niche simultaneously (Staley, 2006). Competition for resources would, through time, select for a single variant, arisen from random mutation and genetic drift, best suited to occupy a specific niche. The existence of microdiverse communities therefore requires that strains exhibit either a high degree of niche specialization (Rainey & Travisano, 1998; Buckling *et al*., 2003; Bessen *et al*., 2005) or a mechanism of non-competitive niche diversification (Rainey *et al*., 2000). Selection based on niche differentiation would thus ultimately lead to phylogenetic divergence at the sub-species level (Cohan, 2002).

The mechanisms governing natural phenotypic variation have been attributed to spontaneous mutation (Schübbe *et al*., 2003), genomic rearrangement (Zverlov *et al*., 2008), and DNA acquisition (de la Cruz & Davies, 2000; Van Ham *et al*., 2000). Furthermore, it has been shown that under conditions of stress, these events increase as a coping mechanism (Woese, 1987). In coping with the stresses of the natural environment, frequent mutagenic events leading to niche diversification may act as an adaptive mechanism and lead to the establishment of a microdiverse community. Understanding the ecological significance of microdiversity has implications in terms of microbial ecology, evolutionary studies, taxonomy, bioprospecting, and biotechnology.

In this study, decaying wood compost was investigated as a potential source of novel lignocellulolytic, biofuel producing bacteria. From a single sample, 12 highly similar, co-existing strains of *Thermoanaerobacter* were isolated. All strains were subjected to genetic and phenotypic characterization to identify, and evaluate, differences between the strains. Small, albeit significant, differences were identified, consistent with previous studies in other environments (Acinas *et al*., 2009; Vermette *et al*., 2010), suggesting the existence of a natural microdiverse population within the compost. This is also the first known report of microdiversity involving the genus *Thermoanaerobacter*.

2.3 Materials and methods

2.3.1 Reference strains

Thermoanaerobacter brockii susbp. *brockii* DSM 1457^T (HTD4), *Thermoanaerobacter pseudethanolicus* DSM 2355^T (ATCC 33223; 39E), and *Thermoanaerobacter thermohydrosulfuricus* DSM 567^T were purchased from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ), Braunschweig, Germany.

2.3.2 Media and substrates

All experiments were performed using ATCC 1191 medium as described (Sparling *et al*., 2006) with 2 g/L cellobiose unless otherwise specified. Bottles containing medium were sealed using butyl rubber stoppers and made anaerobic through gassing: degassing (1 min: 4 min) four times with a final gas cycle using 100% N_2 in accordance with established protocols (Daniels *et al*., 1986). Reducing solution, 200 mM

sodium sulfide, was added at 1% (vol/vol) prior to autoclaving. Filter sterilized, anaerobic sugar solutions were added, as necessary, to the medium using a needle and syringe post-autoclaving.

Acid hydrolyzed hemicellulose (AHH) and neutral hydrolyzed hemicellulose (NHH) were gifts from Pin-Ching Maness at the National Renewable Energy Laboratory (NREL) in Golden, CO. Hydrolysates were prepared from corn-stover according to the protocol of Datar *et al*. (2007). Prior to use as a fermentation substrate, the AHH (pH = 1.6) required over-liming to neutralize acetic and phenolic acids. To over-lime, AHH was heated to 42°C and $Ca(OH)_2$ was added until a pH of 10 was reached. The alkaline solution was allowed to mix for 30 min forming calcium phenate, which was removed via filtration through a 0.22 μm filter. The pH of the filtrate was adjusted to 5.6 corresponding to that of NHH, which was not over-limed.

2.3.3 Enrichment and isolation

Environmental isolates were obtained from self-heating thermophilic, decaying compost in Arundel, PQ, Canada containing maple, pine, and spruce wood. Woodchips were collected from the center of the compost and stored in a 1 L stoppered Corning bottle (Corning Life Sciences, Lowell, MA) under an atmosphere of 100% N_2 at 4°C until use. Woodchips (~10 g) were transferred to 500 mL sterile ATCC 1191 medium inside an anaerobic chamber, stoppered, and incubated at 60°C until growth was evident via gas production. Inoculations (10% vol/vol) were then performed into serum bottles containing 45 mL of ATCC 1191 medium and one of arabinose, cellulose, cellobiose, glycerol, xylose, AHH, or NHH were added as enrichment substrates. All substrates were

added at concentrations of 5 g/L except for AHH and NHH, which were added at 2 g/L (total sugar content[8]) due to limited quantities. Cultures were plated in a Forma Scientific anaerobic glove chamber complete with incubator set at 60°C. Colonies were picked from plates containing medium with 2% (wt/vol) agar and a single enrichment substrate.

2.3.4 Microscopy

Cells were grown to an optical density (OD_{600}) of 0.75 ± 0.03, and wet mounts were prepared for visualization using an AXIO IMAGER.Z1 (Zeiss, Oberkochen, Germany). Images were captured using an AxioCam MRm camera (Zeiss) at a magnification of 1000×. Cell lengths were calculated using AXIOVISION version 4.7.2 software and averages calculated for each strain. Only individual cells, showing no signs of constriction or septum formation, and not those growing in chains, were used for measurements. A minimum of 65 measurements were used in calculating the average for each strain.

2.3.5 Substrate use and niche overlap

To determine the range of fermentable substrates for each isolate, 30 different carbon substrates were tested. Complex substrates tested included casamino acids, tryptone, and yeast extract. Nineteen sugar and sugar derivatives were used including arabinose, cellobiose, esculin, fructose, galactose, galacturonic acid, glucose, lactose, maltose, mannitol, mannose, melibiose, raffinose, rhamnose, ribose, salicin, sorbitol, sorbose, sucrose, and xylose. Tested carboxyl and amino acids included acetate, citrate,

[8] Calculated as the sum of the total sugar content for each saccharide as reported by Datar *et al*., (2007). Hydrolysates prepared using steam explosion under acidic conditions (200°C, 1 min) or neutral conditions (210°C, 3 min) were used.

isoleucine, lactate, pyruvate, succinate, and tartaric acid. All substrates tested were at a final concentration of 2 g/L unless otherwise specified.

To avoid potential growth on complex substrates found in ATCC 1191 medium, a chemically defined medium, MJ medium, (Johnson *et al*., 1981), supplemented with 2 g/L cellobiose was additionally tested as a potential alternative for the substrate utilization experiments. After 7 days, inoculated cultures showed no evidence of growth unless the medium was supplemented with yeast extract. Neither ATCC 1191 vitamin solution nor mineral solution (Sparling *et al*., 2006), casamino acids or tryptone were suitable replacements for yeast extract. As such, ATCC 1191 medium with reduced concentrations (0.67 g/L) of yeast extract was used for the substrate utilization experiments. Growth was quantified via OD_{600}. Carbon utilization was considered positive if the OD_{600} value of the test condition ± standard deviation was greater than that of the inoculated non-substrate control ± standard deviation.

Initial experiments were conducted in 96-well polystyrene plates in a manner similar to that described by Jaspers and Overmann (2004). Sterile plates were placed into an anaerobic chamber 3 days prior to use to remove trace oxygen. The design of the plates was such that wells contained either: (i) no substrate + cells + medium; (ii) substrate + medium (negative control); (iii) substrate + cells + medium. The three wells together comprise a single condition. Conditions were replicated four times in each row per plate and each row had a different carbon source. Seven test substrates and cellobiose (positive control) were tested on each plate. The total volume in the wells was 200 μL. Three plates for each set of conditions were tested and incubated at 60°C inside the anaerobic chamber for 24, 72, or 96 h. One plate was removed after each incubation

period (24, 72 or 96 h) and the growth determined by measuring the OD_{600}. Upon removal from the chamber, plates were left for 20 min prior to reading allowing for resazurin oxidation. Wells containing substrate + medium only (negative control) typically had OD_{600} values ranging from 0.05 to 0.08. Conditions in which the OD_{600} value of the negative control was >0.08 were discarded, as the possibility of contamination could not be discounted. Rows containing more than one negative control, in which the OD_{600} value was >0.08, were discarded and the experiment repeated. When contamination could not be discounted, new working stocks, derived from master stocks grown from single colonies, were prepared and used for subsequent experiments. OD_{600} values were read using a PowerWave-XS single channel spectrophotometer using KCJUNIOR software (BIO-TEK Instruments Inc., Winnoski, VT).

Substrate use was confirmed by repeating the above conditions in Balch (Bellco Glass Inc., Vineland, NJ) tubes using 10 mL cultures in triplicate. All conditions were independently verified a minimum of three times. Conditions where the OD_{600} reached low values (OD600 = ~ 0.10) were re-tested using culture serially transferred three times in 5 g/L test substrate prior to OD_{600} values being measured.

Test inocula were prepared by growing the cells in ATCC 1191 medium with 0.5 g/L cellobiose. Under these conditions, the maximum OD_{600} of all strains ranged from 0.25 to 0.30. All inocula were therefore grown to an OD_{600} of 0.20–0.25 to ensure near cellobiose depletion and to minimize the amount of residual cellobiose introduced into the test condition upon inoculation.

The niche overlap index (NOI), representing the potential for each isolate to occupy an identical carbon catabolism niche, was calculated according to the protocol of

Wilson and Lindow (1994). In brief, the NOI between any pair of strains is the ratio of the number of substrates utilized by both strains, divided by the sum of the total substrates used by either strain. Results were scored as a binary code matrix, where substrate consumption (i.e. potential niche occupation) was scored as a "1", whereas lack of substrate consumption was scored as a "0". Data analysis was performed as described (Jaspers & Overmann, 2004).

2.3.6 16S rRNA and *cpn60* UT amplification, sequencing and phylogenetic analysis

Isolated strains were grown overnight and DNA was extracted using the Wizard® Genomic DNA Purification kit (Promega Corp., Madison, WI) according to the manufacturer's protocol. Amplification of the 16S rRNA encoding gene was performed from genomic DNA using previously described primers (Löffler *et al*., 2000). PCR was performed using DreamTaq™ (Fermentas Canada Inc., Burlington, ON, Canada). The mixtures were comprised of 25 μL 2× PCR buffer, 40 ng DNA, 0.30 μM of each primer and H_2O to a final volume of 50 μL. The PCR cycling conditions were as follows: (i) initial denaturation at 96°C for 1 min; (ii) 35 cycles of 96°C for 45 s, 54°C for 1 min and 72°C for 1 min; and (iii) a final elongation step at 72°C for 10 min.

The chaperonin-60 universal target (*cpn60* UT) region was amplified using the H279/H280 primers previously described (Hill *et al*., 2002). The reaction mixture contained 0.5 μM of each primer, 2.5 mM $MgSO_4$, 500 nM each dNTP, 2.5 U Taq polymerase (HP Taq) and 100–1200 ng DNA template. The cycling conditions involved an initial denaturation step at 95°C for 3 min, followed by 40 cycles of 94°C for 30 s, 50°C for 30 s; 72°C for 45 s. A final extension of 72°C for 5 min was performed.

PCR product was purified prior to sequencing using Micron YM-30 ultrafiltration columns (Millipore, Billerica, MA) pre-rinsed with 200 μL water and centrifuged for 4 min at 14,000 × g. PCR product (30–80 μL) was added to the column, centrifuged for 12 min at 14,000 × g, and the flow through discarded. The purified product was recovered by adding 35 μL of H_2O to the column followed by centrifugation at 2,500 × g for 5 min. The concentration of the purified product was determined by NanoDrop (Thermo Scientific, Wilmington, DE), adjusted to 10–20 ng/μl, and sequenced. Contigs of the raw data were assembled using Pregap4 and Gap4 as part of the STADEN software package (http://sourceforge.net/projects/staden).

Sequences were aligned with the default parameters of CLUSTALW (Thompson *et al*., 1994). Phylogenetic trees were constructed using the maximum composite likelihood method in MEGA 4 (Tamura *et al*., 2007) by the neighbour-joining algorithm with Jukes–Cantor correction (Jukes & Cantor, 1969). Bootstrap values for 1,000 replicates were calculated. Phylogenetic relatedness was verified through Bayesian analysis using the MR. BAYES v3.1.2 program (Ronquist & Huelsenbeck, 2003). Analysis was run for two million generations sampling every 10^{th} tree. A majority rule consensus tree was constructed, and posterior probability analysis was performed. Sequence alignments were conducted and visualized using BioEdit v7.0.9.0 (Hall, 1999). Previously published sequences were accessed through GenBank or through the Joint Genome Institute's IMG database (Markowitz *et al*., 2012).

2.3.7 Genetic fingerprinting

BOX (Martin *et al*., 1992) fingerprinting profiles were investigated for all strains. PCR reactions were performed using the BOXA1R primer (Versalovic *et al*., 1994) as previously described (Urzì *et al*., 2001). Five microliters of PCR product was used for agarose gel electrophoresis (1% wt/vol).

2.3.8 Growth and metabolic profiling

Growth experiments were conducted in ATCC 1191 medium with 2 g/L cellobiose as described (Chapter 2.3.2) and was monitored via OD_{600} using a Biochrom Novaspec II (Biochrom Ltd, Cambridge,UK) spectrophotometer. Cultures were grown to an average $OD_{600} = 0.75 \pm 0.03$, at which time, three biological replicates were removed from incubation for end-product analyses and pH determination. Three additional tubes were removed immediately post-inoculation and were used for time zero measurements. Gas production (H_2 and CO_2) was measured using a Multiple Gas Analyzer #1 Gas Chromatograph System Model 8610-0070 (SRI Instruments, Torrance, CA) as previously described (Rydzak *et al*., 2009). Total production was determined by correcting for gas solubility (Sander, 1999) and the bicarbonate equilibrium (Darrett & Grisham, 1995). One milliliter culture samples were additionally removed, centrifuged (13,000 × *g*, 2 min), and the supernatant stored at -20°C until analysis of non-gaseous end-products could be performed. Biomass calculations, organic acids, ethanol, and residual sugars were determined according to established protocols (Rydzak *et al*., 2009). Residual sugars and organic acids were analyzed by high-performance liquid chromatography (Dionex, ICS 3000) using an anion-exchange Carbo-PacPA1 column (sugar analyses) or

an IonPac AS11-HC (organic acids) anion-exchange column. Ethanol measurements were determined spectrophotometrically at 340 nm using a R-Biopharm UV-Test kit (Darmstadt, Germany). All standards were prepared using ATCC 1191 medium to account for background signals. The supernatant from final tubes was diluted 1:4 to ensure that concentrations were in the detectable range. Reported end-product yields are reported as net production accounting for carryover during inoculation (T=0 measurements). All growth studies were replicated three independent times.

2.3.9 Nucleotide accession numbers

Nucleotide accession numbers for sequences determined in this study are accessible through GenBank. Accession numbers for 16S rRNA gene sequences are HM585213–HM585225, while those for *cpn60* UT regions are HM623896–HM623910. Accession numbers for all strains used for phylogenetic analyses can be found in Table A.1.1.

2.4 Results and discussion

2.4.1 Enrichment and isolation of strains

The production of H_2 or CO_2, used as an initial indicator of growth, was evident in all enrichment cultures. Gas production on arabinose, cellulose, and glycerol was much lower than for the other enrichment cultures, which was later attributed to fermentation of the yeast extract in the medium, rather than the enrichment substrate itself (Table A.1.2). Undiluted enrichment culture was transferred to plates, yielding relatively uniform colonies amongst all of the test substrates. Colonies were smooth,

uniformly round, mucoid, flat, and white. After 24 h incubation, the colonies could only be distinguished by size, ranging from 2 to 4.5 mm. Thirteen colonies (WC1–WC13) from different enrichment substrates (Table 2.1), and those that varied in size, were selected for subsequent genetic analysis.

2.4.2 Molecular identities and microdiversity of isolates

To begin identifying the isolated strains, DNA was extracted from all 13 isolates and the 16S rRNA-encoding gene sequences were determined. BLAST analysis identified a >99% shared sequence identity between the 16S rRNA-encoding sequence from isolates WC1-WC12 and *T. pseudethanolicus* 39E (DSM 2355) (Onyenwoke *et al*., 2007). The sequence from isolate WC13, which is significantly shorter than those determined for isolates WC1–WC12 (Table 2.1), shared the highest sequence identity (>99%) to *Thermoanaerobacterium thermosaccharolyticum* W16, an isolate from a hot spring in China (Ren *et al*., 2008).

Alignments of the 16S rRNA genes sequence from isolates WC1-WC12 with the four 16S gene copies found in *T. pseudethanolicus* 39E (GenBank accession no. CP000924.1) identified that the high shared sequence identity only exists with a single gene copy (found at nucleotide position 2,265,749 – 2,267,517) within the *T. pseudethanolicus* 39E genome. This particular 16S rRNA gene sequence is distinct in that it contains four intervening sequences (IVS) making it significantly longer than the other gene copies. IVS were also present in the 16S rRNA gene sequences for isolates WC1–WC12 as evidenced by sequence alignments (Fig. A.1.1). Alignment of the WC1–

Table 2.1. Enrichment substrate and nucleotide at three divergent loci within the 16S rRNA gene sequence determined for isolates WC1-WC13.

			Nucleotide position[a]		
Isolate	Enrichment substrate	Sequence length (base pairs)	1185	1382	1413
WC1	AHH	1658	T	C	G
WC2	NHH	1655	C	G	G
WC3	NHH	1661	T	G	G
WC4	NHH	1658	T	G	A
WC5	Xylose	1667	T	G	G
WC6	Xylose	1665	C	C	G
WC7	Xylose	1666	C	G	G
WC8	Arabinose	1668	C	G	A
WC9	Cellobiose	1661	T	G	A
WC10	Glycerol	1667	C	C	G
WC11	Cellulose	1665	C	G	A
WC12	Cellulose	1663	T	G	A
WC13	Cellobiose	1457	NA[b]	NA	NA

[a]Position determined in reference to WC8.

[b]NA = Not applicable as WC13 is a strain of *Thermoanaerobacterium thermosaccharolyticum.*

Lengths of published 16S rRNA gene sequences (in base pairs) for *T. brockii* subsp. *brockii* HTD4 = 1513, *T. pseudethanolicus* 39E = 1601, 1527, 1527, 1769 and *T. thermohydrosulfuricus* DSM 567 = 1768.

WC12 sequences identified three single nucleotide polymorphisms (SNPs) within the gene at conserved positions. Divergence at nucleotide positions 1185, 1382, and 1413 is such that, from the 12 sequences determined, six non-identical sequences exist and no more than three isolates share an identical sequence. Furthermore, no two isolates sharing a common sequence were derived from the same enrichment treatment (Table 2.1). Despite only three SNPs of variation, WC1–WC12 fell into two separate clades (Fig. 2.1) and clustered isolates WC1-WC12 with *T. thermohydrosulfuricus* DSM 567, another strain known to have IVS in its 16S rRNA encoding gene (Rainey *et al*., 1993).

The presence of IVS, which have also been described in numerous members of the *Thermoanaerobacteriaceae* family (Rainey *et al*., 1993), made accurate taxonomic assignment of the strains difficult. Sequence alignments identified that the positioning of all four IVS sequences within the rRNA gene is conserved amongst isolates WC1–WC12, *T. pseudethanolicus* 39E and *T. thermohydrosulfuricus* DSM 567. Furthermore, the sequence identity of all four IVS is 100% conserved across these strains (Table A.1.3) with the exception of IVS-1 and IVS-4 found in *T. thermohydrosulfuricus* DSM 567. The IVS-1 sequence from *T. thermohydrosulfuricus* DSM 567 shows 97.3% homology to that of *T. pseudethanolicus* 39E and the WC isolates, whereas IVS-4 shares 91.5% sequence identity across strains. However, these values may be artificially low as the published 16S rRNA gene sequence for *T. thermohydrosulfuricus* DSM 567 (GenBank accession no. L09161; Rainey *et al*., 1993) contains residues identified as "N" within these IVS regions. The 16S rRNA gene sequence of *T. brockii* subsp. *brockii* HTD4 contains no IVS.

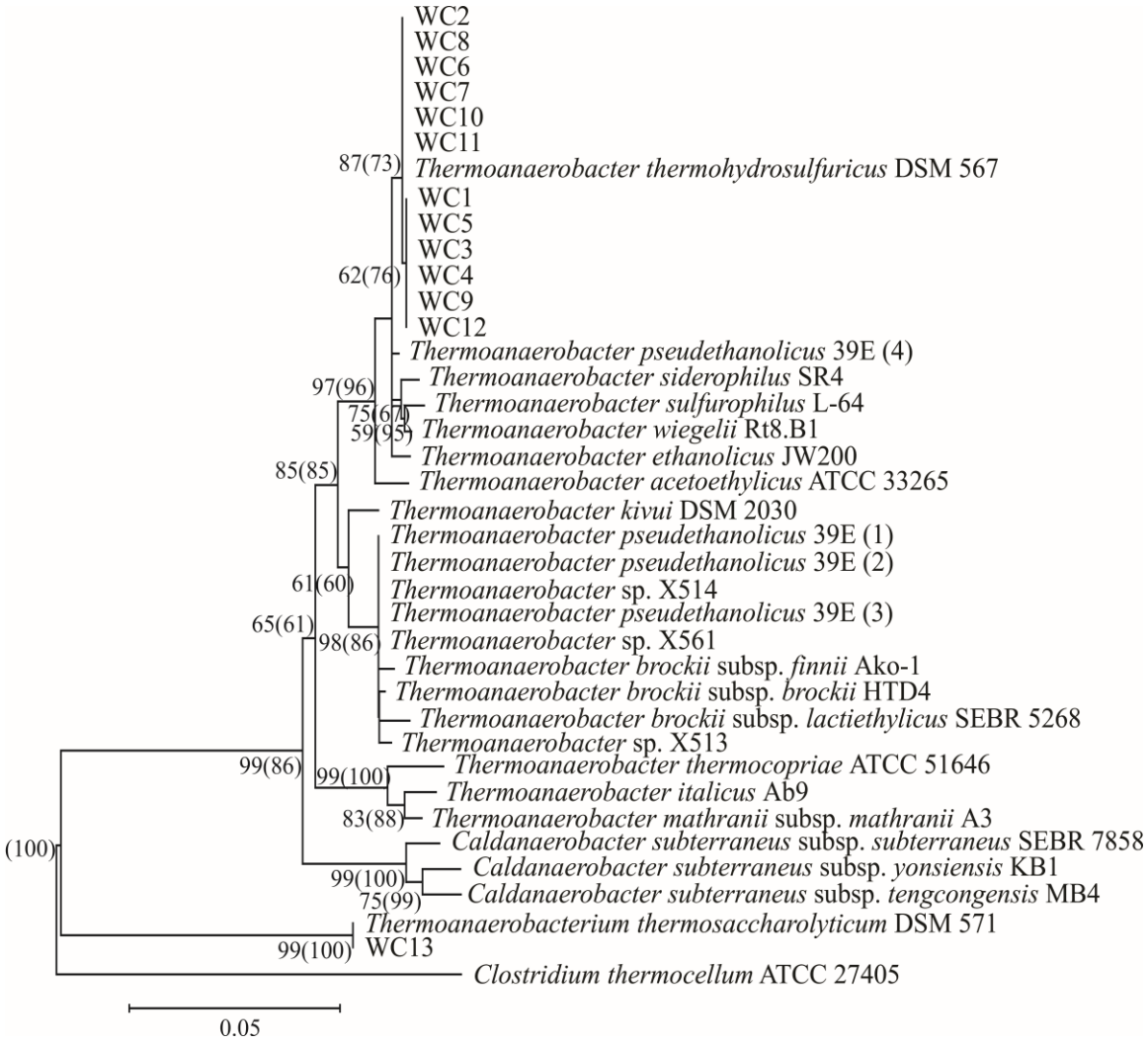


Fig 2.1. Neighbour-joining tree of 16S rRNA sequences from isolates WC1-WC13 and select reference strains. *Clostridium thermocellum* was used to root the tree. The tree was constructed as described (Chapter 2.3.6). Confidence levels >50% are shown after bootstrap values (outside of brackets) were calculated and posterior probability analyses (inside brackets) was performed. Nucleotide accession numbers for sequences used can be found in Table A.1.1. Numbers in brackets after *Thermoanaerobacter pseudethanolicus* 39E represent its four annotated 16S rRNA encoding sequences with the numerical values reflecting the relative position within the genome from nucleotide position +1. Scale bar represents the number of changes per nucleotide.

Additional evaluation of the published genomes of *Thermoanaerobacter* strains (available from the Joint Genome Institute's IMG database; Markowitz *et al*., 2012) has identified that not all gene copies within a common genome necessarily contain IVS. For example, of the four *T. pseudethanolicus* 39E *rrn* operons, one 16S rRNA gene sequence (position: 2,265,749 – 2,267,517) has four IVS, whereas another copy (position: 438,455 – 440,055) has only a single IVS. The additional two 16S rRNA gene copies have no IVS.

As IVS are not necessarily found in all 16S rRNA gene copies, yet the sequence is conserved across strains, it suggests that gene transfer of a particular rRNA gene sequence may have occurred in the evolutionary history of these strains. Typically, genes with core functions, such as *rrn* genes, are thought to be less susceptible to horizontal gene transfer (HGT) (Jain *et al*., 1999; Daubin *et al*., 2003). However, high levels of *rrn* inter-operonic variability among thermophiles, most notably within *Caldanaerobacter subterraneus* subsp. *tengcongensis* (formerly *Thermoanaerobacter tengcongensis*; Fardeau *et al*., 2004), have been attributed to HGT (Acinas *et al*., 2004b). The potential for horizontal gene acquisition is further supported by the recent realization that members of the *Thermoanaerobacteriaceae* family exhibit varying levels of natural competence (Shaw *et al*., 2010).

High interoperonic diversity observed in the *T. pseudethanolicus* 39E genome and in highly related strains (Fardeau *et al*., 2004; Pei *et al*., 2010a) suggests that using 16S rRNA gene sequences for phylogenetic classification of *Thermoanaerobacter* strains may not accurately reflect true evolutionary relationships. In addition, the possibility that *rrn* genes may have undergone HGT in this genus creates uncertainty in taxonomic

assignment of these isolates on the basis of the 16S rRNA gene sequence alone. Therefore, an alternative phylogenetic signature gene, the more variable *cpn60* UT region was investigated.

The *cpn60* UT region has previously been shown to accurately discriminate between closely related strains (Marston *et al*., 1999). Furthermore, it has been shown (Chapter 3) that sequence similarity of the *cpn60* UT region can predict whole genome relatedness with the same accuracy as some multi-locus and even whole genome strategies and correlates well with DNA–DNA hybridization values. The phylogenetic relatedness of isolate WC13 to *Th. thermosaccharolyticum* DSM 571 observed using 16S rRNA gene sequence analysis is further supported by the shared sequence identity of the *cpn60* UT region (97.8%), thus suggesting it is a strain of this species. Isolates WC2–WC7 shared the greatest *cpn60* UT sequence identity with *T. thermohydrosulfuricus* DSM 567 (>98.7%), whereas isolates WC1 and WC8–WC12 shared the greatest sequence identity (>98.9%) with *Thermoanaerobacter* sp. Rt8.G4 (Truscott & Scopes, 1998), a strain formerly identified as *T. brockii* (Truscott *et al*., 1994). In contrast, the *cpn60* UT sequence identity between any of the isolates and *T. pseudethanolicus* 39E was much lower ranging from 87.1 to 89.3%. Recently (June, 2010), the draft genome sequence for *Thermoanaerobacter wiegelii* Rt8.B1 (Cook *et al*., 1996) was released by the Joint Genome Institute and isolates WC1 and WC8–WC12 shared a greater *cpn60* UT sequence identity (>99.1%) with this strain. It should be noted that the original designation of *T. wiegelii* Rt8.B1 as a novel species (Cook *et al*., 1996) is largely based on phenotypic and 16S rRNA data without using the previously recognized standard (DNA–DNA hybridization) (Stackebrandt *et al*., 2002) for species designation.

Unfortunately, the lack of a partial or complete genome sequence for the type strain of *T. thermohydrosulfuricus* DSM 567 prevents whole genome comparisons of the type described by Richter & Rosselló-Móra (2009); however, mathematical models for predicting whole genome sequence identities, and thus phylogenetic relatedness, (Chapter 3) suggest that *T. wiegelii* Rt8.B1, *T. thermohydrosulfuricus* DSM 567, and the isolates (WC1–WC12) described here are probably related at the species level (data presented in Chapter 3).

Overall, sequence divergence was greater in the *cpn60* UT region than was found using the 16S rRNA-encoding gene. Twelve unique genetic signatures amongst isolates WC1–WC12 were identified suggesting they were members of a genetically microdiverse community. Further, different clustering patterns (Fig. 2.2) were observed using the *cpn60* UT region than were observed using 16S rRNA gene sequences. It would be expected that if isolates WC1–WC12 represented a continuum of genetic drift, the clustering patterns would remain constant. Thus, these findings suggest that within these isolates, divergence has occurred in multiple directions before the initial laboratory enrichment process.

2.4.3 Genetic fingerprinting of isolates

To help determine the extent of genetic diversity amongst the isolates, fingerprinting profiles based on the BOX (Versalovic *et al*., 1994) element were investigated. Genetic fingerprinting based on repetitive DNA sequences is effective for identifying genetic diversity at the sub-species level (Dombek *et al*., 2005; Tacão *et al*.,

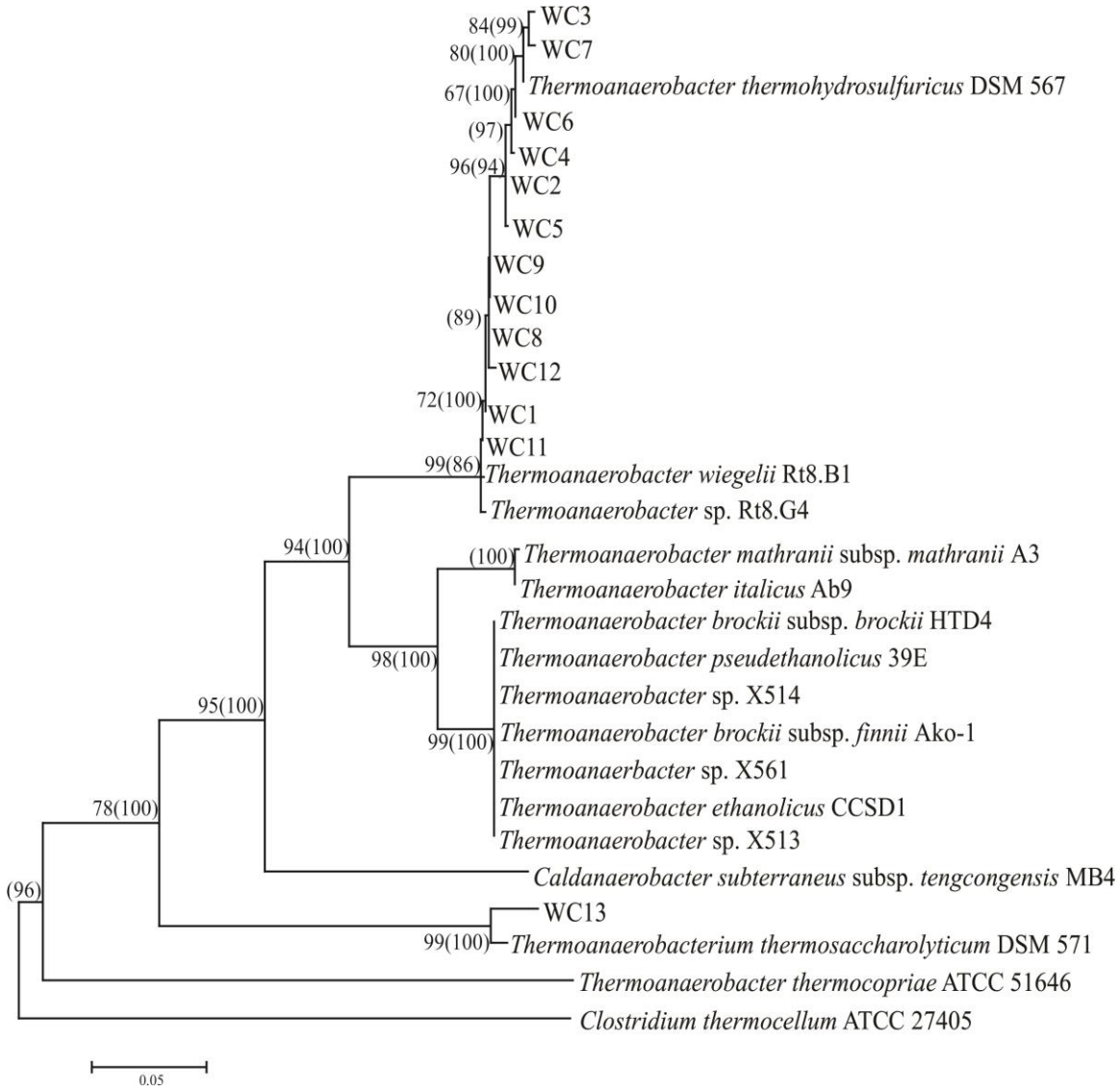


Fig. 2.2. Neighbour-joining tree of *cpn60* UT sequences from isolates WC1-WC13 and from select reference strains. *C. thermocellum* was used to root the tree. The tree was constructed as described in Chapter 2.3.6. Confidence levels >50% are shown after bootstrap values (outside of brackets) were calculated and posterior probability analyses (inside brackets) was performed. Nucleotide accession numbers for sequences used can be found in Table A.1.1. Scale bar represents the number of changes per nucleotide.

2005), and the BOX sequence has previously been effectively employed with strains of *Thermoanaerobacter* (Roh *et al*., 2002).

For fingerprinting analysis, the type *T. brockii* strain, *T. brockii* subsp. *brockii* HTD4 (Lee *et al*., 1993), *T. pseudethanolicus* 39E, *T. thermohydrosulfuricus* DSM 567, and *Th. thermosaccharolyticum* WC13 were used as reference strains. Distinct banding patterns were identified amongst isolates WC1– WC12 using BOX-PCR (Fig. 2.3). Isolate WC5 showed three distinct bands between 400 and 500 nucleotides in length, which was similar to both *T. brockii* subsp. *brockii* HTD4 and *T. pseudethanolicus* 39E, but distinct from *T. thermohydrosulfuricus* DSM 567. All other isolates showed banding patterns most comparable to *T. thermohydrosulfuricus* DSM 567, though distinct patterns amongst isolates were evident (e.g. WC3 and WC4). In addition, identical banding patterns existed between isolates that were not consistent with the phylogenetic clustering observed (Fig. 2.1, Fig. 2.2).

These results further support the proposal that these isolates form a genetically microdiverse group of *T. thermohydrosulfuricus* isolated from a common sampling environment. The differential banding patterns provide evidence that, among isolates WC1–WC12, mechanisms allowing for genomic diversification exist. Although the mechanism(s) of diversification (e.g. gene deletions, gene acquisitions, mobile element insertions, etc.) are unknown, they provide evidence of evolutionary progression potentially leading to strain differentiation (Cohan *et al*., 2002). Also, although the significance of these specific genomic variations is unknown in these isolates, previous studies have shown that chromosomal rearrangements can lead to strain diversification and niche specialization (Römling *et al*., 1997; Kresse *et al*., 2003).

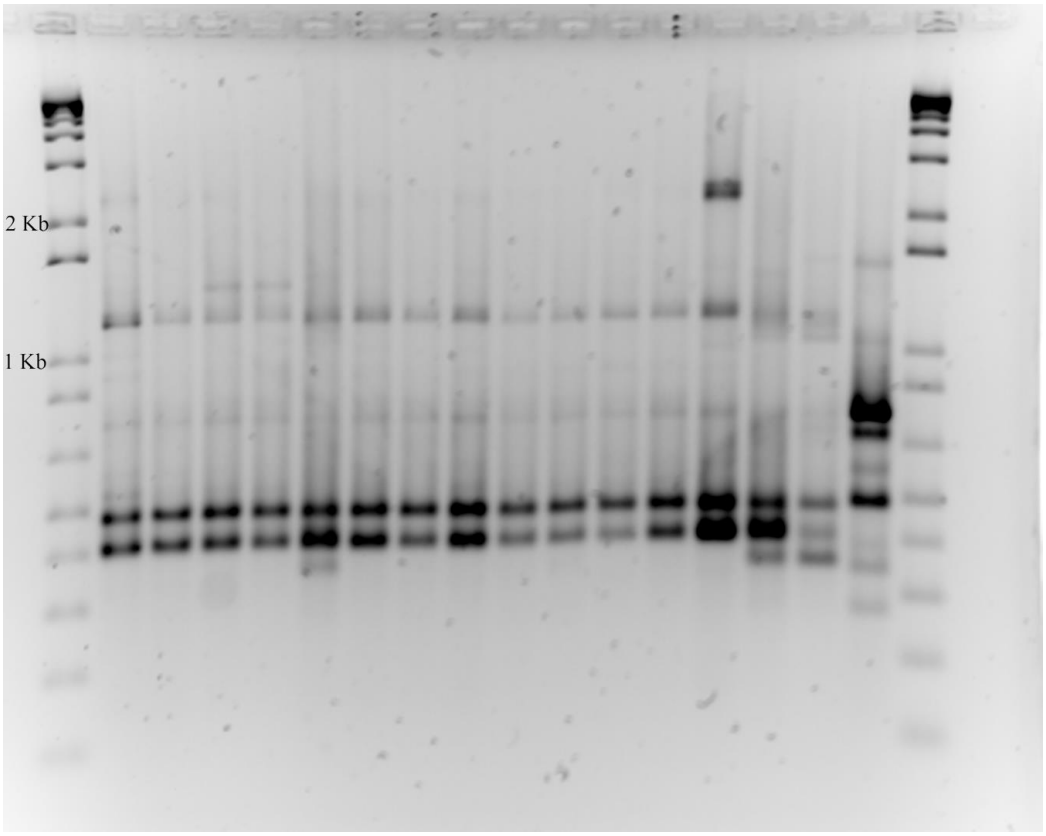


Fig. 2.3. Banding patterns of BOX-PCR profiles for isolates WC1-WC13, *T. pseudethanolicus* 39E, *T. brockii* subsp. *brockii* HTD4 and *T. thermohydrosulfuricus* DSM 567. Lanes (left to right): DNA ladder; WC1; WC2; WC3; WC4; WC5; WC6; WC7; WC8; WC9; WC10; WC11; WC12; *T. thermohydrosulfuricus* DSM 567; *T. pseudethanolicus* 39E; *T. brockii* subsp. *brockii* HTD4; *Thermoanaerobacterium thermosaccharolyticum* WC13; DNA ladder; water (negative control).

2.4.4 Cell morphology and size

Cells were grown to an OD_{600} = 0.75 ± 0.03 and then examined via differential interference contrast microscopy. Cultures WC1–WC12 showed indistinct morphologies as cells in all cultures were rod-shaped and arranged singly or in chains (Fig. A.1.2). Long filamentous cells were also periodically observed. All of the above characteristics are typical of cells within the *Thermoanaerobacter* genus (Wiegel *et al*., 1981) and are morphologically indiscriminate from the descriptions for *T. pseudethanolicus* 39E (Onyenwoke *et al*., 2007), *T. brockii* subsp. *brockii* HTD4 (Zeikus *et al*., 1979), and *T. thermohydrosulfuricus* DSM 567 (Lee *et al*., 1993). Cell division was often unequal as suggested by differences in lengths of cells arranged in chains. Cells removed from growth medium and immediately examined under the microscope were motile for short periods of time. The lengths of the cells varied from 1.68 to 17.83 μm. Average cell length was calculated for each culture and was shown to vary from 3.66 μm for WC4 to 6.24 μm for WC6 (Fig. A.1.3). The average lengths for isolates WC1–WC12 were greater than those observed for *T. pseudethanolicus* 39E and *T. brockii* subsp. *brockii* HTD4, but were smaller than those for *T. thermohydrosulfuricus* DSM 567.

2.4.5 Niche occupation and specialization

Considerable niche overlap (NOI >0.9) was found amongst strains WC1–WC12 (Fig. 2.4). Strain WC13, identified by DNA sequence analysis as a strain of *Th. thermosaccharolyticum*, also showed a NOI >0.9 in comparison to isolates WC1–WC12 confirming high potential niche overlap independent of species, and even genus, designation. Previous studies investigating the potential niche overlap of genetically

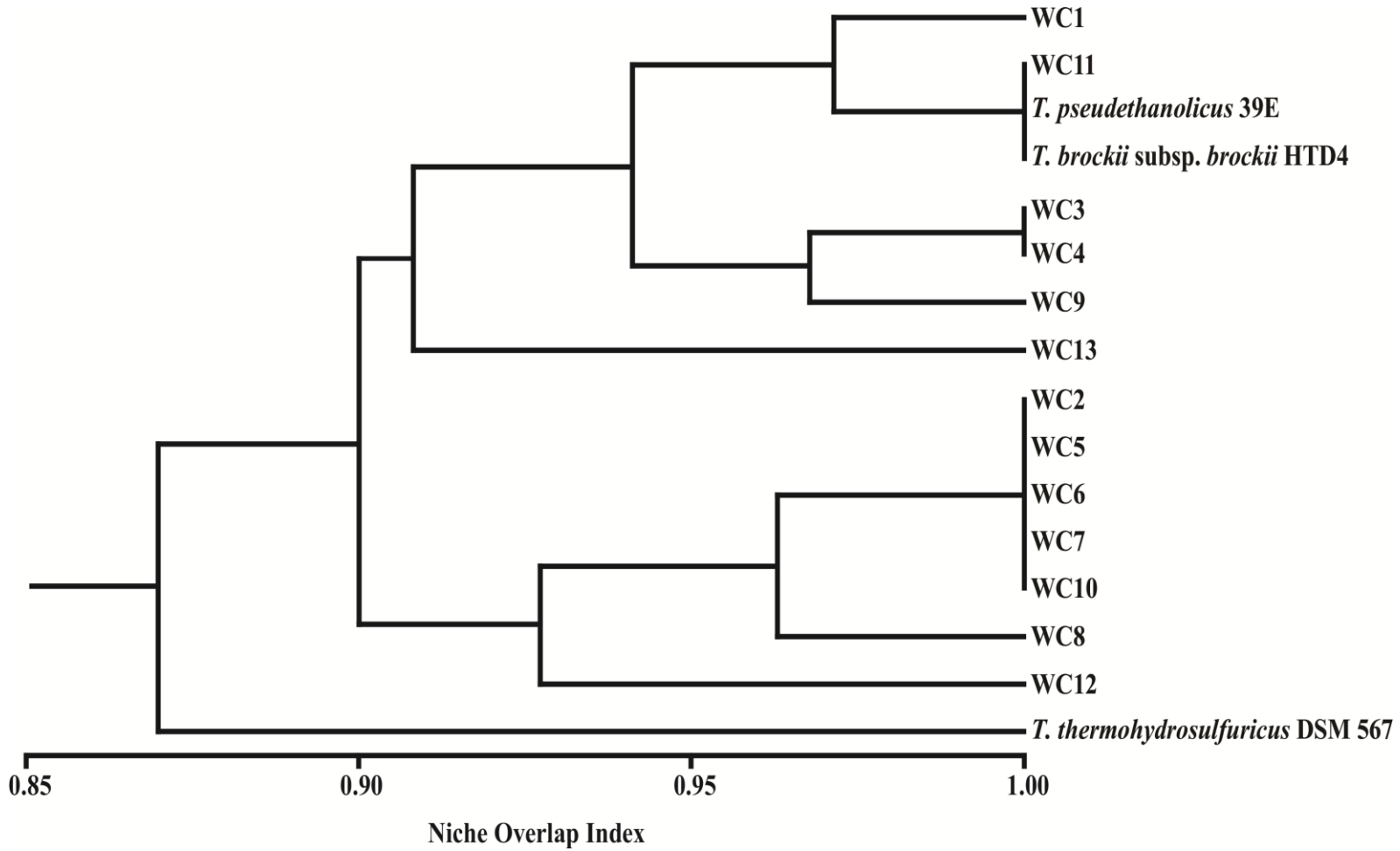


Fig. 2.4. Potential NOI of isolates WC1-WC12, *T. brockii* subsp. *brockii* HTD4, *T. pseudethanolicus* 39E and *T. thermohydrosulfuricus* DSM 567 as determined by substrate utilization patterns. WC13, identified as a strain of *Th. thermosaccharolyticum*, is included as a phylogenetic outgroup.

microdiverse strains have reported a wide range of NOIs ranging from 0.25 to 0.59 for 11 *Brevundimonas alba* strains (Jaspers & Overmann, 2004) to 1.00 in various *Pseudomonas syringae* strains (Wilson & Lindow, 1994).

Despite the high NOIs observed, seven utilization profiles (Fig. 2.4) for WC1–WC12 could be identified based upon the differential fermentation capabilities of only six carbon sources (Table 2.2). WC8 showed the least catabolic diversity as growth was supported using only 13 of the substrates, which all other strains were also capable of using. In contrast, WC1 was the most versatile, and was capable of fermenting 18 substrates.

Three distinct clusters, where all members of the cluster demonstrated identical utilization patterns, were observed. One cluster grouped isolate WC11 with *T. pseudethanolicus* 39E and *T. brockii* subsp. *brockii* HTD4 (Fig. 2.4). Isolates WC3 and WC4 comprised another cluster with both utilizing 15 test substrates, whereas WC2, WC5, WC6, WC7, and WC10 formed the final cluster capable of fermenting 14 substrates. This largest cluster was the only one observed that grouped all the isolates from a common enrichment together (WC5, WC6, WC7 – xylose enrichment). These isolates, with the exception of WC10, also formed a subgroup based on *cpn60* UT analysis (Fig. 2.2). The sugar utilization pattern exhibited by *T. pseudethanolicus* 39E agrees with previously published descriptions (Onyenwoke *et al*., 2007). *T. brockii* subsp. *brockii* HTD4 was shown to ferment xylose and mannose under these conditions in contrast to previous reports (Zeikus *et al*., 1979). This discrepancy may be due to differences in experimental design as this study included incubation periods of up to 96 h, whereas those of Zeikus *et al*. (1979) were only incubated for 16 h prior to measuring

Table 2.2. Differential substrate utilization of six carbon sources for isolates WC1-WC12, *T. brockii* subsp. *brockii* HTD4, *T. pseudethanolicus* 39E and *T. thermohydrosulfuricus* DSM 567[a].

Strain	Lactose	Mannitol	Melibiose	Raffinose	Sorbitol	Sucrose
WC1	+	+	+	+	+	+
WC2	-	-	+	+	-	-
WC3	+	+	+	-	-	-
WC4	+	+	+	-	-	-
WC5	-	-	+	+	-	-
WC6	-	-	+	+	-	-
WC7	-	-	+	+	-	-
WC8	-	-	+	-	-	-
WC9	+	+	+	+	-	-
WC10	-	-	+	+	-	-
WC11	+	+	+	+	-	+
WC12	+	-	+	+	+	-
T. brockii subsp. *brockii* HTD4	+	+	+	+	-	+
T. pseudethanolicus 39E	+	+	+	+	-	+
T. thermohydrosulfuricus DSM 567	-	+	-	-	-	+

Symbols denote substrate utilization (+) or a lack of substrate utilization (-).

[a]Complete utilization profile can be found in Table A.1.2.

growth. Long lag periods and slow generation times were commonly observed when cultures were transferred to a new carbon substrate in this study. If also true in the study by Zeikus *et al*., (1979), this may potentially account for the lack of observable growth. The substrate utilization capabilities of *T. thermohydrosulfuricus* DSM 567 agree with previous reports (Lee *et al*., 1993; Hollaus & Sleytr, 1972; Wiegel *et al*., 1979) except in the utilization of arabinose and raffinose. In this study, *T. thermohydrosulfuricus* DSM 567 did not utilize arabinose in agreement with Hollaus and Sleytr (1972), in contrast to the report by Wiegel *et al*. (1979). This inability may be partly explained by the use of lower amounts of yeast extract in the experiments performed here as it has been previously reported that yeast extract can significantly impact the extent of microbial growth for *T. thermohydrosulfuricus* DSM 567 (Wiegel *et al*., 1979). Surprisingly, despite the observed genetic similarities (Fig. 2.1 - Fig. 2.3), *T. thermohydrosulfuricus* DSM 567 showed the most distinct profile of all of the strains examined (Fig. 2.4).

The co-occurrence of such highly similar strains from a common environment suggests that, for strain survival, niche diversification and specialization may have existed in the natural community. Competition for resources in a natural environment is thought to deter similar ecotypes from sharing a common ecological niche and thus prevent their co-existence (Staley, 2006). The similar substrate utilization patterns (high NOIs) observed for strains WC1–WC12 though seemingly violates this principle. These results may be explained in two ways. First, niche specialization may involve parameters not investigated in the experiments performed here. Second, wood compost is a dynamic and nutritionally heterogeneous environment conducive to the formation of micro-niches. Spatial and temporal variation in the evolutionary and ecological forces exerted on these

micro-environments may ultimately result in diversified localized cell populations. Previous lab studies have shown that in heterogeneous environments, adaptive radiation through a combination of mutation and natural selection allows for phenotypic and niche diversification from an initial clonal population of cells (Rainey & Travisano, 1998; Brockhurst *et al*., 2007). While the *in situ* environmental and evolutionary pressures exerted on isolates WC1-WC12 are unknown, it is evident that these pressures have permitted the divergence, or maintenance of divergence, in the carbon catabolism potential of these strains resulting in a phenotypically microdiverse community.

2.4.6 Metabolic profiling

Growth rate and fermentation end-products on 2 g/L cellobiose were investigated to compare differences between the central metabolism of isolates WC1–WC12, *T. brockii* subsp. *brockii* HTD4, *T. pseudethanolicus* 39E, and *T. thermohydrosulfuricus* DSM 567. Significant differences were observed (Table 2.3; Fig. A.1.4 – Fig. A.1.10) in both growth rate and ratio of end-products formed. Although the O/R indices suggest accurate measurement of all end-products, the carbon recovery for all strains, except WC1, WC10, and *T. thermohydrosulfuricus* was higher than 100%. Presumably, this is due to the utilization of yeast extract in the medium, as all strains have shown to support growth using yeast extract alone (Table A.1.2). This is further supported by HPLC analyses indicating that the average cellobiose consumption was >98 ± 3.2% upon reaching an OD_{600} = 0.75 ± 0.03. As significant lag periods were not observed, we assume that the additional substrates were consumed simultaneously with the cellobiose

Table 2.3. Comparison of doubling times and fermentation end-product[a] averages on 2g/L cellobiose at 60°C amongst isolates WC1-WC12, *T. brockii* subsp. *brockii* HTD4, *T. pseudethanolicus* 39E and *T. thermohydrosulfuricus* DSM 567.

		Concentrations produced (mM)							
Culture	Doubling time (h/gen)	Acetate	Lactate	Ethanol	H_2	CO_2	Biomass	O/R index	% carbon recovery
WC1	1.86	8.30	3.80	7.12	6.60	13.95	3.92	1.04	99.14
WC2	2.50	2.05	17.00	3.03	1.73	6.59	3.82	0.97	102.64
WC3	2.32	3.10	16.23	4.49	2.10	8.22	3.78	0.98	108.95
WC4	6.74	7.11	8.22	7.96	4.82	13.38	3.96	1.00	107.14
WC5	2.16	4.12	13.06	3.64	1.75	7.02	3.79	0.95	109.43
WC6	2.28	3.62	15.10	2.82	1.56	6.09	3.81	0.94	110.88
WC7	2.40	2.01	14.89	3.10	1.78	6.42	3.81	0.94	107.11
WC8	2.65	6.80	13.53	4.94	2.88	9.98	3.77	1.08	115.39
WC9	3.01	2.85	14.17	2.84	2.10	6.11	3.66	0.92	111.59
WC10	2.62	2.84	12.83	2.78	2.52	6.83	3.82	0.99	97.34
WC11	1.88	2.46	11.89	2.21	1.39	5.40	3.85	0.93	101.17
WC12	2.45	0.98	16.52	3.94	2.23	7.23	3.84	0.91	109.49
T. pseudethanolicus 39E	2.14	6.12	4.74	12.23	0.98	15.62	3.77	1.00	107.95
T. brockii subsp. *brockii* HTD4	2.68	2.26	11.92	10.83	0.66	11.84	3.68	0.98	107.05
T. thermohydrosulfuricus DSM 567	5.57	9.33	4.03	7.63	6.88	13.09	3.96	0.93	96.67

[a]End product data correspond to measurements taken at an average OD_{600} = 0.75 ± 0.03 and corrected for carryover from inoculation. Values represent averages from three independent experiments using three biological replicates per experiment.

for direct biosynthetic use, or after the exponential phase of growth when cellobiose was depleted.

Acetate, lactate, ethanol, H_2 and CO_2 were the only observed fermentation products for isolates WC1–WC12 and the three reference strains. However, significant differences existed in both the ratio of end-products formed and the doubling time on 2 g/L cellobiose (Table 2.3; Fig. A.1.4). Most isolates produced lactate as the major end-product, whereas isolates WC1 and WC4 produced CO_2 (coupled to acetate and ethanol production) as the major product of fermentation. There also seemed to be no obvious correlation between fermentation products and growth rate. For example, WC1 and WC4 represented the extremes in doubling times (1.86 and 6.74 h/gen respectively), despite the fact that both strains had similar acetate and ethanol yields (Table 2.3). The two most similar isolates in terms of calculated doubling time, WC1 and WC11, exhibited a 4.3-fold difference in lactate production.

The reference strains produced the same end products as those of the isolates. For *T. brockii* subsp. *brockii* HTD4, lactate was the major fermentation end-product with only trace amounts of H_2 produced in agreement with previous findings (Lee *et al*., 1993; Ben-Bassat *et al*., 1981). The fermentation profile of *T. pseudethanolicus* 39E indicated that ethanol was the major reduced end-product formed, similar to previous reports (Zeikus *et al*., 1980; Lovitt *et al*., 1988). The nature of the end-products for *T. thermohydrosulfuricus* DSM 567 was consistent with those reported by Wiegel *et al*., (1979), although the ratio of end products differed in this study. Acetate was the major organic acid produced here, whereas ethanol was predominant in Wiegel and coworkers' experiment (1979). This difference may be due to variations in experimental design as

the incubation temperatures varied by 10°C between the two studies. Previous work with *Thermoanaerobacter* has shown that temperature can significantly impact the enzymatic activity of the core metabolic enzymes and thus impact end-product ratios (Lamed & Zeikus, 1980a).

The ecological significance of the metabolic differences observed between the strains is currently unclear. As the fermentation products were identical in all strains, variation in the ratios may be evidence of niche specialization, but also may simply indicate a certain level of genetic variation due to drift or mutation. Minor genetic variations may significantly impact factors including effector molecule binding affinities and rates of carbon flux. In addition, these effects may have either immediate or downstream effects. Previous studies have shown how mutations in metabolic enzymes may alter the rates of metabolic flux, while still maintaining reaction fidelity (Goupil *et al*., 1996).

Alternatively, the differences may also be due to divergence in protein expression patterns or regulatory control over protein activity, but the exact nature for the observed differences is not currently known. Further, it is unknown if these differences may have arisen through genetic drift at specific loci or through larger genome arrangements. For example, previous studies have shown how mobile element based chromosomal rearrangements can lead to phenotype variation, and ultimately differential niche occupation (Dekkers *et al*., 1998). Although mutagenesis of this type has been reported in the related anaerobic thermophile, *Clostridium thermocellum* (Zverlov *et al*., 2008), to date, no events have been described within the *Thermoanaerobacter* genus. An evaluation of the completed genomes from *Thermoanaerobacter* sp. X514, *T.*

pseudethanolicus 39E and the draft sequence of *Thermoanaerobacter brockii* subsp. *finnii* Ako-1 (as of July, 2010) identified multiple putative annotated transposase encoding sequences. Although the presence of mobile elements is unknown within isolates WC1–WC12, their frequent presence within the genus suggests they may exist.

2.5 Conclusions

Twelve genetically and phenotypically distinct, co-existing strains of *T. thermohydrosulfuricus* were isolated from a single decaying wood compost sample. Their co-existence suggests that within the compost, microbial diversification may be tightly linked to niche specialization, as has previously been reported in other environments (Rainey & Travisano, 1998; Ramsing *et al*., 2000). Such diversification and specialization has been shown to result from fitness "trade-offs" (Rainey *et al*., 2000). These "trade-offs" may have evolved amongst strains WC1–WC12 as a means of allowing for the spatial co-existence of strains derived from a single lineage, while permitting localized populations to fill unoccupied or non-competitive niches. In brief, a "tradeoff" may exist in the sense that one strain may lose the ability to ferment one substrate, yet optimize its utilization of another. A second strain may evolve in an inverse manner whereby it optimizes utilization of the substrate that the first strain no longer ferments. Mechanisms through which polymorphisms leading to fitness 'tradeoffs' and niche diversification include cross-feeding (Helling *et al*., 1987), spatial variation (Rainey & Travisano, 1998), and temporal variation (Turner *et al*., 1996; Horner-Devine *et al*., 2004). Furthermore, the effect that polymorphisms have on the fitness of a population is governed by the environmental context. The inability to

ferment a carbon source is not a fitness disadvantage if alternative substrates are sufficient to support growth during a period of time defined by additional growth parameters (e.g. temperature, pH, oxygen concentration, etc.). Therefore, differential substrate usage and phenotypic diversification may have evolved as a means of non-competitive co-existence amongst highly similar strains resulting in a microdiverse community.

However, it is important to realize that the niche occupation potential examined here may not be reflective of the true *in situ* activity. Wood compost is a heterogeneous environment, and the *in situ* activity of a strain at a given point in time will be greatly influenced by its immediate environment and the regulatory and biochemical processes of the strain itself. As such, although it is unclear to what extent these differences in niche occupation potential are realized in the natural environment, there is significant diversity between strains in their potential to fulfill various ecological niches.

Our results indicate that within the genus *Thermoanaerobacter*, significant genotypic and phenotypic microdiversity can exist at the sub-species level within a common environment. Although the specific mechanisms governing these differences are unclear, it is evident that varying evolutionary pressures have acted upon these co-existing strains in the form of blended phylogenies, potential niche specialization, and variations in their central metabolism. As such, this study highlights the importance of examining sub-species diversity in future ecological, evolutionary, and bioprospecting studies.

2.6 Authors' contributions

Experiments were conceived and designed by Tobin J. Verbeke, Nazim Cicek, David B. Levin and Richard Sparling. Strain enrichment and isolation, substrate use profiling, growth and end-product analyses, BOX-PCR profiling, 16S rRNA/*cpn60* UT sequence amplifications and phylogenetic analyses were performed by Tobin J. Verbeke. Additional strain isolation efforts, as well as microscopy, were performed by Scott Wushke. Tim J. Dumonceaux conducted 16S and *cpn60* UT sequencing reactions and assembly. The original manuscript, upon which this chapter is based, was authored by Tobin J. Verbeke, with input from Tim J. Dumonceax, David B. Levin and Richard Sparling.

2.7 Acknowledgements

We thank Dr. Pin-Ching Maness for supplying the hemicellulose hydrolysates. We also thank Dr. Bruce Ford for his assistance with the NTSYS-pc software used in evaluation of the NOI. We finally thank Marc Ransom, Florian Labat, and Andrea Wilkinson for technical contributions to this work.

This work was supported by a Natural Sciences and Engineering Research Council (NSERC) Strategic Grant (STPGP 365076), by the Manitoba Rural Adaptation Council (MRAC), Advancing Canadian Agriculture and Agri-Food (ACAAF) program (#309009), Genome Canada, and by the Cellulosic Biofuels Network (Agriculture and Agri-Food Canada).

Chapter 3. Predicting relatedness of bacterial genomes using the chaperonin-60 universal target (*cpn60* UT): Application to *Thermoanaerobacter* species[9]

3.1 Abstract

D.R. Zeigler determined that the sequence identity of bacterial genomes can be predicted accurately using the sequence identities of a corresponding set of genes that meet certain criteria using a three-gene model (Zeigler, 2003). This model for comparing bacterial genome pairs requires the determination of the sequence identities for *recN*, *thdF*, and *rpoA*. To do this, it involves the generation of approximately 4.2 kb of genomic DNA sequence from each organism to be compared, and also normally requires that oligonucleotide primers be designed for amplification and sequencing based on the sequences of closely related organisms. However, we have developed an analogous mathematical model for predicting the sequence identity of whole genomes based on the sequence identity of the 549–567 base pair chaperonin-60 universal target (*cpn60* UT). The *cpn60* UT is accessible in nearly all bacterial genomes with a single set of universal primers, and its length is such that it can be completely sequenced in one pair of overlapping sequencing reads via di-deoxy sequencing. Both mathematical models were applied to a set of *Thermoanaerobacter* isolates from woodchip compost and it was shown that both the one-gene *cpn60* UT-based model and the three-gene model based on

[9] A modified version of this chapter has previously been published. Reproduced with permission: © 2011, Elsevier GmbH. Found in: **Verbeke TJ, Sparling R, Hill JE, Links MG, Levin DB, Dumonceaux TJ**. 2011. Predicting relatedness of bacterial genomes using the chaperonin-60 universal target (*cpn60* UT): Application to *Thermoanaerobacter* species. Syst Appl Microbiol **34**: 171-179.

recN, *rpoA*, and *thdF* predicted that these isolates could be classified as *Thermoanaerobacter thermohydrosulfuricus*. Furthermore, it was found that the genomic prediction model using *cpn60* UT gave similar results to whole-genome sequence alignments over a broad range of taxa, suggesting that this method may have general utility for screening isolates and predicting their taxonomic affiliations.

3.2 Introduction

Determining the taxonomic identity of a bacterial isolate usually involves a combination of microscopy, phenotypic characterization, and DNA sequence analysis (Gevers *et al*., 2005). While the designation of a novel species requires whole genome comparisons based on DNA–DNA hybridization data (Stackebrandt *et al*., 2002) or, increasingly, by the comparison of whole or partial genome sequence data (Richter & Rosselló-Móra, 2009), less onerous approaches based on the unambiguous and highly reproducible determination of selected gene sequences can be viewed as an alternative for assigning an isolate to a pre-existing taxon or determining whether experimentation justifying the designation of a novel species is warranted (Stackebrandt *et al*., 2002; Ludwig *et* al., 2007; Pontes *et al*., 2007). The DNA sequence(s) to examine, however, remains a matter of choice. Although most isolates continue to be taxonomically classified based on the sequence of their 16S ribosomal RNA-encoding genes (Crocetti *et al*., 2000; Clavel *et al*., 2007), the limitations of this approach for discerning closely related bacteria are being increasingly recognized. Stackebrandt *et al*. (2002) suggested that "an informative level of phylogenetic data would be obtained from the determination of a minimum of five genes under stabilizing selection for encoded metabolic functions."

Protein-encoding genes are known to provide higher levels of taxonomic resolution than non-protein-encoding genes (Scholss *et al*., 2005) and classifications based on *gyrB* (Wang *et al*., 2007b), *recA* (Weng *et al*., 2009), *rpoB* (Adékambi *et al*., 2009; Meintanis *et al*., 2008), *recN* (Zeigler, 2005; Arahal *et al*., 2008), *cpn60* (Glazunova *et al*., 2009) and other genes have been proposed. Furthermore, Lorén *et al*. (2010) recently determined that the sequences of certain protein encoding housekeeping genes, including *cpn60*, can be used to predict with accuracy the genomic G + C content of species within the genus *Aeromonas*. A number of bacterial typing schemes based on the sequence comparisons of a number of protein-encoding genes from each isolate (multilocus sequence typing, or MLST) have been reported, but again the choice of genes to include in the MLST scheme is variable (Cooper & Feil, 2004).

Zeigler (2003) has reported the most comprehensive analysis to date of genes that are likely to be useful for predicting relatedness at the whole genome level. For this analysis, a computational algorithm was developed for determining whole genome sequence identities for 44 bacterial genomes distributed across 16 genera that were available at the time of analysis. The whole genome sequence identities corresponded well with available DNA–DNA hybridization data (Zeigler, 2003), and these sequence identity values were used to develop correlations and corresponding prediction models for individual genes and discrete sets of genes. These models thus facilitated the prediction of whole genome sequence identity for pairs of isolates based on the determination of the sequence identity of a set of genes. A scan of the genomic information for genes that are universal, lack paralogs and are phylogenetically informative led to the identification of a set of 32 candidate genes. These genes were

then evaluated for their abilities to predict whole genome relatedness, and examination of all these genes led to the development of mathematical models for predicting the sequence identity of pairs of genomes. The single-gene model that performed best (i.e. had the highest correlation between gene and genome sequence identities) was *recN*, while the best model overall involved three genes: *recN*, *rpoA*, and *thdF* (Zeigler, 2003). The gene that had the lowest correlation of sequence identity to genome sequence identity was the 16S rRNA-encoding gene.

The chaperonin-60 gene (*cpn60*, or *groEL* in *Escherichia coli*) is an approximately 1,650 base pair (bp) gene whose product functions to chaperone protein folding in prokaryotes and eukaryotes (Hemmingsen *et al*., 1988). It is present in virtually all Bacteria and a fragment of this gene, the 549–567 bp (183–189 amino acid) *cpn60* universal target (*cpn60* UT), is accessible from any isolate or from a microbial community (Vermette *et al*., 2010) with a set of universal amplification primers. Furthermore, the sequence that is amplified is highly distinct among organisms and an extensive database of *cpn60* UT sequence information is available (Hill *et al*., 2004). Since *cpn60* UT sequences are so distinct, short enough to be sequenced in a single reaction, and accessible using a universal set of amplification primers, we set out to determine how well *cpn60* UT sequence identities correlated to the genome sequence identities reported by Zeigler (2003). We also compared predictions of genome sequence identities using *cpn60* UT to the results determined by whole-genome alignments of a broad range of taxa, which is an emerging gold standard for species identification (Richter & Rosselló-Móra, 2009).

3.3 Materials and methods

3.3.1 Bacterial strains and growth

In lab stocks of *Thermoanaerobacter* spp. WC1-WC12 (Chapter 2), *Thermoanaerobacter brockii* subsp. *brockii* HDT4, *Thermoanaerobacter pseudethanolicus* 39E, and *Thermoanaerobacter thermohydrosulfuricus* DSM 567 were used for all experiments. Cultures were grown in anaerobic ATCC 1191 medium containing 2 g/L cellobiose as described (Chapter 2.3.2).

3.3.2 DNA extraction, amplification and sequencing of genes for determining taxonomic identity

DNA was extracted from overnight cultures of each strain using a Wizard Genomic DNA purification kit (Promega, Madison, WI), according to the recommended procedures (Chapter 2.3.6). To amplify each of the target genes (*rpoA*, *recN*, *thdF*) in the WC isolates and in the reference *Thermoanaerobacter* strains, PCR primers were designed for amplification and sequencing based on the identification of gene-specific orthologous regions in related *Thermoanaerobacter* genomes accessed using the Joint Genome Institute's IMG database (Markowitz *et al*., 2012) (Table A.2.1). Amplification of each target gene was performed using the following cycling conditions: (i) initial denaturation at 94^{o}C for 4 min; (ii) 35 cycles of 94^{o}C for 45 s, 1 min of annealing and 1min at 72^{o}C; and, (iii) final elongation at 72^{o}C for 10 min. Annealing temperatures used for *recN*, *rpoA* and *thdF* were 62^{o}C, 62^{o}C and 60^{o}C, respectively. Amplicons were purified for sequencing using Amicon YM-30 columns (Millipore).

3.3.3 Sequence alignments, identity determination and phylogenetic analyses

Sequences were assembled into contigs using Gap4 in the Staden software package (http://sourceforge.net/projects/staden/). Consensus sequences were exported as fasta files and used as input for multiple sequence alignments using CLUSTALW (Thompson *et al*., 1994) in BioEdit v7.0.9.0 (Hall, 1999). Sequences were aligned and gaps removed such that sequences were trimmed to that of the shortest high quality sequence obtained: for *recN*, 616 bp; *rpoA*, 754 bp; *thdF*, 1,214 bp. The sequence identity of each pairwise comparison for each gene was determined using the number of identical residues for each alignment divided by the total length of the alignment. Phylogenetic analysis was conducted using MEGA4 (Tamura *et al*., 2007). Consensus trees were inferred from 500 replicates and were constructed using the neighbour-joining method (Saitou & Nei, 1987). Calculations of average nucleotide identity (ANI) by MumMer[10] (ANIm), ANI by BLAST (ANIb), and tetranucleotide frequency correlations were performed using JSpecies, as described by Richter and Rosselló-Móra (2009).

3.4 Results and discussion

The genomic prediction model described by Zeigler (2003) provides a strong mathematical rationale for the selection of target genes for determining the taxonomic identity of an unknown isolate. The three-gene model or the one-gene *recN* based model has been successfully applied to a variety of isolates (Ziegler, 2005; Kuhnert *et al*., 2009). However, it should be noted that, while the criteria upon which genes were selected

[10] MuMmer (Maximal unique Match-mer) is a software program designed to align sequence data sets (e.g. whole genome) too large to be computed via traditional BLAST algorithms (Delcher *et al*., 1999; Kurtz *et al*., 2004). In the present work, MuMmer is part of the *JSpecies* software package used for analyses (Richter & Rosselló-Móra, 2009).

included those that were long enough to be phylogenetically informative, yet short enough to be sequenced in a reasonable number of reactions, the three-gene model still requires the generation of approximately 4.2 kb of sequence data from each isolate to be compared. Furthermore, with the exception of the 16S rRNA-encoding gene, none of the genes reported by Zeigler have a region that is currently universally accessible with a single primer set. Essentially, this means that amplification and sequencing primers must be designed based on reported sequences for close phylogenetic relatives whom encode orthologous gene sequences. We were therefore interested in determining whether the *cpn60* UT, which is relatively short (549–567 bp), and is accessible in almost all bacteria with a single set of universal primers (Hill *et al*., 2006), might be useful as a predictor of genomic identity in the same manner as the genes identified by Zeigler. Although *cpn60* is present as a single copy gene in the majority of organisms, distinct paralogs exist in some genomes (Lund, 2009), which may explain the exclusion of *cpn60* from Zeigler's original analysis. These paralogs are typically sufficiently distinct from one another that any of them can be a reasonable candidate for unambiguous taxonomic assignment.

Using the same set of organisms, the sequence identities were calculated for the *cpn60* UT and these values were correlated to the whole genome sequence identities determined by Zeigler. It was found that, using only a 552–555 bp fragment of the *cpn60* gene, the correlation of *cpn60* UT sequence identities to whole genome sequence identities was 0.89 (Fig. 3.1), which is just above that reported for *dnaB* and 10th best on the list of 32 genes examined (Zeigler, 2003). Linear regression analyses enabled the determination of a predictive model for whole genome relatedness of a pair of organisms

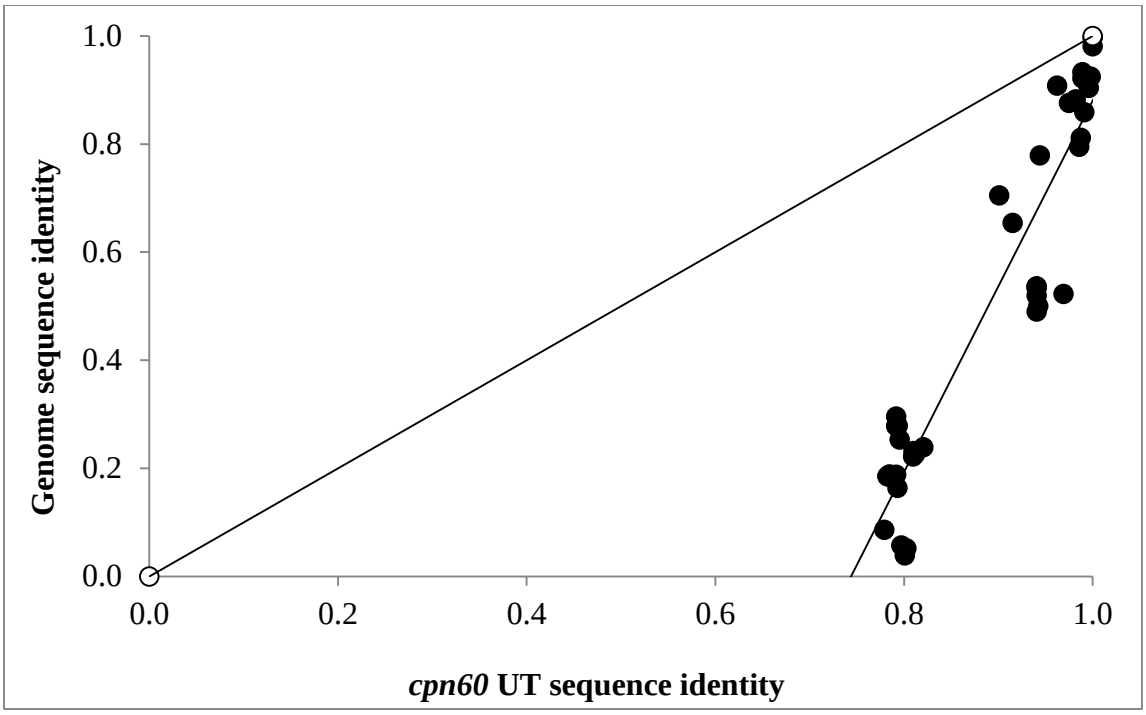


Fig. 3.1. Linear regression analysis of *cpn60* UT sequence identity compared to genome sequence identity. Genome sequence identities were from Zeigler (2003). The lines show the best fit of the data as well as a hypothetical 1:1 relationship between genome sequence identity and *cpn60* UT sequence identity.

based on the corresponding sequence identities of the *cpn60* UT:

$$SI_{genome} = -2.56 + 3.44(SI_{cpn60\ UT}) \text{ (Eq. 1)}$$

where SI_{genome} is the genome sequence identity as determined by Zeigler (2003), and $SI_{cpn60\ UT}$ is the sequence identity of the *cpn60* UT in the corresponding strains. The high correlation shown by the *cpn60* UT, combined with the fact that it is the shortest target region on the list and the only target besides 16S rRNA that is currently universally amplifiable with a single set of primers, led us to determine whether Eq. (1) could be used to predict the genome sequence identities of unknown isolates and determine their taxonomic classification. As a test set of organisms, we chose a series of *Thermoanaerobacter* isolates from woodchip compost, WC1–WC12, which showed varying degrees of taxonomic relatedness to *T. pseudethanolicus* 39E, *T. thermohydrosulfuricus* DSM 567 and *T. brockii* subsp. *brockii* HTD4 (Chapter 2). These organisms are especially difficult to type by their 16S rRNA sequences due to high sequence conservation across species, as well as the existence of multiple 16S rRNA copies that exhibit unusually high intragenomic divergence (Acinas *et al*., 2004b; Pei *et al*., 2010a).

The prediction based on Eq. (1) was compared to the genome sequence identities reported by Zeigler and it was found that *cpn60* UT provided a robust (mean absolute residual = 0.088) means of determining taxonomic identity (Fig. 3.2A). Using genomic sequence data for a series of *Thermoanaerobacter* species that are currently available through the JGI's IMG database (Markowitz *et al*., 2012), it was determined that, for

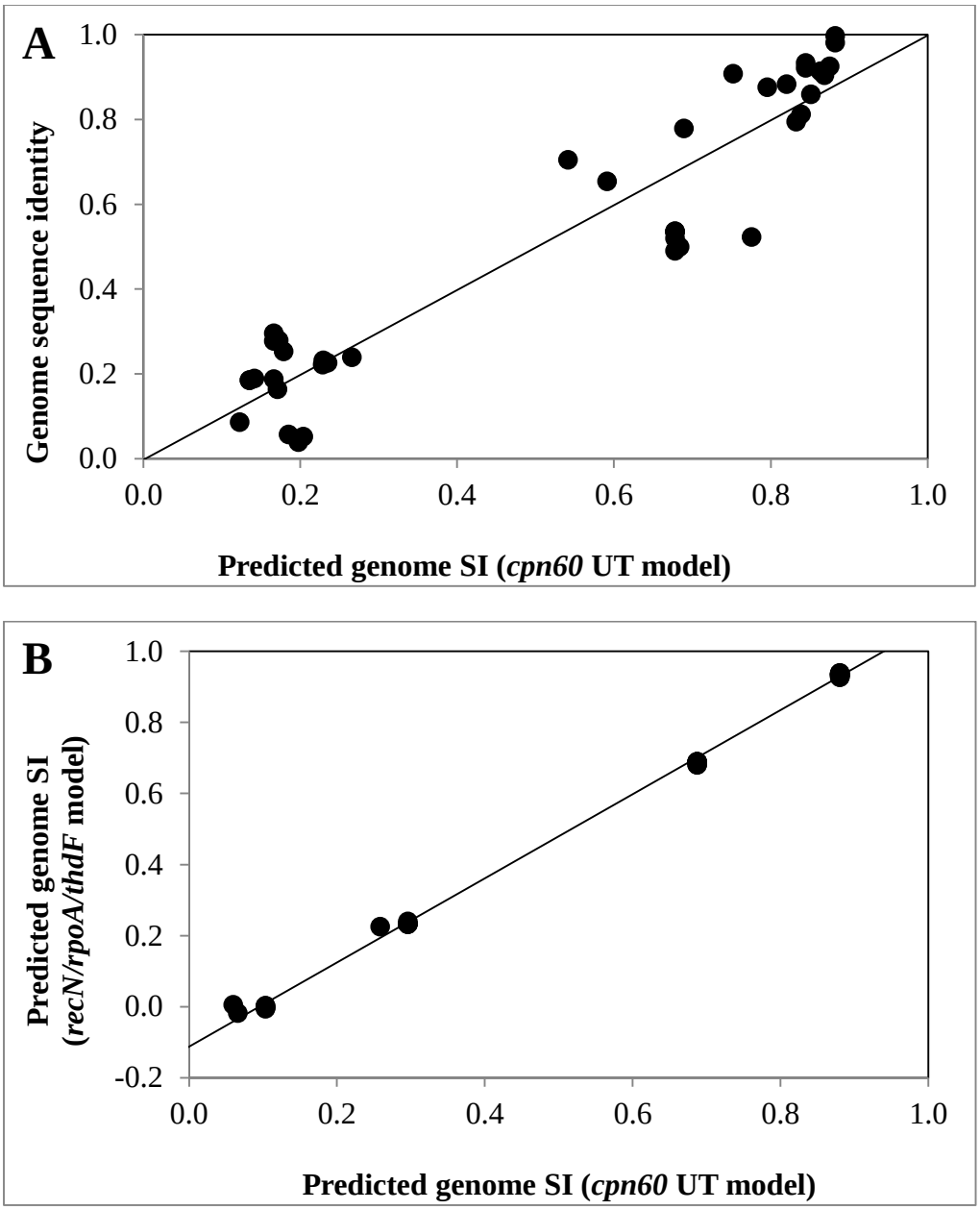


Fig. 3.2. Correlation of genome sequence identity predictions using the three-gene model of Zeigler (2003) and the one gene model based on the *cpn60* UT. (A) All species as described in Zeigler. (B) *Thermoanaerobacter* spp. and select reference organisms (*Cal. subterraneus* subsp. *tengcongensis* MB4 and *Th. thermosaccharolyticum* DSM 571).

these species, the prediction of whole genome sequence identity based on Zeigler's three-gene model and that based on the *cpn60* UT model reported here shared a very high level of agreement ($R^2 = 0.9987$; Fig. 3.2B). Furthermore, the species predictions using our one-gene model and Zeigler's three-gene models were compared to the determinations of average nucleotide identity and tetranucleotide frequency correlations (Richter & Rosselló-Móra, 2009) generated from the reported whole genome sequences of reference *Thermoanaerobacter* strains from GenBank and the JGI. Using an average nucleotide identity by MuMmer (ANIm) of 95% as the species-delineating line cut-off, as suggested by Richter and Rosselló-Móra (2009), it was found that, in all cases, the predictions using the gene models and the ANIm determinations resulted in the same species designations for these strains (Table 3.1). All models predicted that *T. brockii* subsp. *finnii* Ako-1, *T. pseudethanolicus* DSM 2355, *T. ethanolicus* CCSD1, and *Thermoanaerobacter* spp. X513, X514, and X561 should be considered the same species (Table 3.1).

To predict the genome sequence identities between isolates WC1-WC12 and the *Thermoanaerobacter* reference strains, the sequences of *recN*, *rpoA*, and *thdF* were determined. For each gene, including the *cpn60* UT whose sequences had previously been reported (Chapter 2), the sequence identities, as well as the predicted whole genome sequence identities were determined for each corresponding isolate and reference strain pair (Table 3.2; Table A.2.2). Consistent with Zeigler, a cut-off of 70% genomic identity was used for determining species identity. High sequence identities at all loci examined were found for the WC isolates, indicating that these isolates formed a taxonomically coherent group and could likely be considered a single bacterial species. It was also shown that the isolates had high sequence identities to *T. thermohydrosulfuricus* DSM

Table 3.1. Predictions of genomic sequence identities for *Thermoanaerobacter* and other select reference strains using the *cpn60* UT based one-gene model, the three-gene model of Zeigler (2003), and determination of average nucleotide identities and correlations of tetranucleotide frequences using JSpecies (Richter & Rosselló-Móra, 2009).

	prediction of genomic SI using									
Pair	*recN*	*rpoA*	*thdF*	*cpn60* UT	3-gene model	1-gene model	ANIm	ANIb	Tetra	Species prediction[a]
T. thermohydrosulfuricus DSM 567-*T. brockii* subsp. *finnii* Ako-1	0.966	0.969	0.879	0.877	0.691	0.456	ND[b]	ND	ND	Different species
T. thermohydrosulfuricus DSM 567-*T. ethanolicus* CCSD1	0.966	0.969	0.878	0.877	0.689	0.456	ND	ND	ND	Different species
T. thermohydrosulfuricus DSM 567-*T. italicus* Ab9	0.900	0.942	0.871	0.868	0.627	0.425	ND	ND	ND	Different species
T. thermohydrosulfuricus DSM 567-*T. mathranii* subsp. *mathranii* A3	0.902	ND	0.870	0.868	0.635	0.425	ND	ND	ND	Different species
T. thermohydrosulfuricus DSM 567-*T. pseudethanolicus* 39E	0.966	0.969	0.879	0.877	0.691	0.456	ND	ND	ND	Different species
T. thermohydrosulfuricus DSM 567-*Thermoanaerobacter* sp. X513	0.965	0.969	0.879	0.877	0.690	0.456	ND	ND	ND	Different species
T. thermohydrosulfuricus DSM 567-*Thermoanaerobacter* sp. X514	0.965	0.969	0.879	0.877	0.690	0.456	ND	ND	ND	Different species
T. thermohydrosulfuricus DSM 567-*Thermoanaerobacter* sp. X561	0.965	0.969	0.879	0.877	0.690	0.456	ND	ND	ND	Different species
T. thermohydrosulfuricus DSM 567-*Cal. subterraneus* subsp. *tengcongensis* MB4[c]	0.715	0.784	0.754	0.815	0.241	0.244	ND	ND	ND	Different species
T. thermohydrosulfuricus DSM 567-*Th. thermosaccharolyticum* DSM 571[d]	0.651	0.729	0.648	0.777	-0.008	0.113	ND	ND	ND	Different species

Table 3.1 cont.

Pair	*recN*	*rpoA*	*thdF*	*cpn60* UT	3-gene model	1-gene model	ANIm	ANIb	Tetra	Species prediction
T. brockii subsp. *finnii* Ako-1-*T. ethanolicus* CCSD1	1.000	1.000	0.999	1.000	0.939	0.880	0.980	0.977	0.999	Same species
T. brockii subsp. *finnii* Ako-1-*T. italicus* Ab9	0.912	0.949	0.896	0.944	0.682	0.687	0.922	0.917	0.997	Different species
T. brockii subsp. *finnii* Ako-1-*T. mathranii* subsp. *mathranii* A3	0.911	ND	0.899	0.944	ND	0.687	0.921	0.914	0.996	Different species
T. brockii subsp. *finnii* Ako-1-*T. pseudethanolicus* 39E	1.000	1.000	0.996	1.000	0.932	0.880	0.999	1.000	0.999	Same species
T. brockii subsp. *finnii* Ako-1-*Thermoanaerobacter* sp. X513	0.998	1.000	0.999	1.000	0.937	0.880	0.981	0.979	0.999	Same species
T. brockii subsp. *finnii* Ako-1-*Thermoanaerobacter* sp. X514	0.998	1.000	0.994	1.000	0.929	0.880	0.981	0.979	0.999	Same species
T. brockii subsp. *finnii* Ako-1-*Thermoanaerobacter* sp. X561	0.998	1.000	0.999	1.000	0.937	0.880	0.980	0.979	0.999	Same species
T. brockii subsp. *finnii* Ako-1-*Cal. subterraneus* subsp. *tengcongensis* MB4	0.711	0.790	0.749	0.830	0.234	0.294	0.863	0.769	0.959	Different species
T. brockii subsp. *finnii* Ako-1-*Th. thermosaccharolyticum* DSM 571	0.646	0.728	0.653	0.774	-0.003	0.101	0.862	0.697	0.902	Different species
T. ethanolicus CCSD1-*T. italicus* Ab9	0.912	0.949	0.895	0.944	0.681	0.687	0.922	0.977	0.999	Different species
T. ethanolicus CCSD1-*T. mathranii* subsp. *mathranii* A3	0.911	ND	0.898	0.944	ND	0.687	0.919	0.912	0.996	Different species
T. ethanolicus CCSD1-*T. pseudethanolicus* 39E	1.000	1.000	0.995	1.000	0.931	0.880	0.979	0.977	0.998	Same species
T. ethanolicus CCSD1-*Thermoanaerobacter* sp. X513	0.998	1.000	0.998	1.000	0.935	0.880	0.985	0.980	0.998	Same species
T. ethanolicus CCSD1-*Thermoanaerobacter* sp. X514	0.998	1.000	0.994	1.000	0.928	0.880	0.985	0.980	0.998	Same species

Table 3.1 cont.

Pair	*recN*	*rpoA*	*thdF*	*cpn60* UT	3-gene model	1-gene model	ANIm	ANIb	Tetra	Species prediction
T. ethanolicus CCSD1-*Thermoanaerobacter* sp. X561	0.998	1.000	0.998	1.000	0.935	0.880	0.985	0.980	0.998	Same species
T. ethanolicus CCSD1-*Cal. subterraneus* subsp. *tengcongensis* MB4	0.711	0.790	0.749	0.830	0.233	0.294	0.866	0.770	0.959	Different species
T. ethanolicus CCSD1-*Th. thermosaccharolyticum* DSM 571	0.646	0.728	0.654	0.774	-0.002	0.101	0.890	0.700	0.904	Different species
T. italicus Ab9-*T. mathranii* subsp. *mathranii* A3	0.979	ND	0.995	1.000	ND	0.880	0.981	0.980	0.999	Same species
T. italicus Ab9-*T. pseudethanolicus* 39E	0.912	0.949	0.899	0.944	0.689	0.687	0.925	0.918	0.997	Different species
T. italicus Ab9-*Thermoanaerobacter* sp. X513	0.912	0.949	0.899	0.944	0.689	0.687	0.925	0.916	0.997	Different species
T. italicus Ab9-*Thermoanaerobacter* sp. X514	0.912	0.949	0.899	0.944	0.689	0.687	0.925	0.916	0.997	Different species
T. italicus Ab9-*Thermoanaerobacter* sp. X561	0.912	0.949	0.899	0.944	0.689	0.687	0.925	0.916	0.997	Different species
T. italicus Ab9-*Cal. subterraneus* subsp. *tengcongensis* MB4	0.691	0.780	0.753	0.819	0.226	0.257	0.860	0.763	0.953	Different species
T. italicus Ab9-*Th. thermosaccharolyticum* DSM 571	0.644	0.730	0.658	0.761	0.006	0.057	0.873	0.696	0.909	Different species
T. mathranii subsp. *mathranii* A3-*T. pseudethanolicus* 39E	0.911	ND	0.895	0.944	ND	0.687	0.921	0.915	0.996	Different species
T. mathranii subsp. *mathranii* A3-*Thermoanaerobacter* sp. X513	0.912	ND	0.899	0.944	ND	0.687	0.923	0.915	0.997	Different species

Table 3.1 cont.

Pair	*recN*	*rpoA*	*thdF*	*cpn60* UT	3-gene model	1-gene model	ANIm	ANIb	Tetra	Species prediction
T. mathranii subsp. *mathranii* A3-*Thermoanaerobacter* sp. X514	0.912	ND	0.895	0.944	ND	0.687	0.923	0.914	0.997	Different species
T. mathranii subsp. *mathranii* A3-*Thermoanaerobacter* sp. X561	0.912	ND	0.899	0.944	ND	0.687	0.923	0.915	0.997	Different species
T. mathranii subsp. *mathranii* A3-*Cal. subterraneus* subsp. *tengcongensis* MB4	0.693	ND	0.748	0.819	ND	0.257	0.855	0.765	0.955	Different species
T. mathranii subsp. *mathranii* A3-*Th. thermosaccharolyticum* DSM 571	0.644	ND	0.652	0.761	ND	0.057	0.865	0.700	0.904	Different species
T. pseudethanolicus 39E-*Thermoanaerobacter* sp. X513	0.998	1.000	0.999	1.000	0.937	0.880	0.981	0.978	0.999	Same species
T. pseudethanolicus 39E-*Thermoanaerobacter* sp. X514	0.998	1.000	0.999	1.000	0.937	0.880	0.981	0.977	0.999	Same species
T. pseudethanolicus 39E-*Thermoanaerobacter* sp. X561	0.998	1.000	0.999	1.000	0.937	0.880	0.981	0.977	0.999	Same species
T. pseudethanolicus 39E-*Cal. subterraneus* subsp. *tengcongensis* MB4	0.711	0.790	0.753	0.830	0.240	0.294	0.865	0.773	0.958	Different species
T. pseudethanolicus 39E-*Th. thermosaccharolyticum* DSM 571	0.646	0.728	0.656	0.774	0.002	0.101	0.859	0.696	0.904	Different species
Thermoanaerobacter sp. X513-*Thermoanaerobacter* sp. X514	1.000	1.000	0.996	1.000	0.932	0.880	1.000	1.000	1.000	Same species
Thermoanaerobacter sp. X513-*Thermoanaerobacter* sp. X561	1.000	1.000	1.000	1.000	0.940	0.880	0.999	1.000	1.000	Same species
Thermoanaerobacter sp. X513-*Cal. subterraneus* subsp. *tengcongensis* MB4	0.711	0.790	0.748	0.830	0.232	0.294	0.866	0.770	0.956	Different species

Table 3.1 cont.

Pair	*recN*	*rpoA*	*thdF*	*cpn60* UT	3-gene model	1-gene model	ANIm	ANIb	Tetra	Species prediction
Thermoanaerobacter sp. X513-*Th. thermosaccharolyticum* DSM 571	0.646	0.728	0.652	0.774	-0.006	0.101	0.880	0.699	0.905	Different species
Thermoanaerobacter sp. X514-*Thermoanaerobacter* sp. X561	1.000	1.000	0.996	1.000	0.931	0.880	0.999	1.000	1.000	Same species
Thermoanaerobacter sp. X514-*Cal. subterraneus* subsp. *tengcongensis* MB4	0.711	0.790	0.748	0.830	0.232	0.294	0.866	0.770	0.956	Different species
Thermoanaerobacter sp. X514-*Th. thermosaccharolyticum* DSM 571	0.646	0.728	0.652	0.774	-0.006	0.101	0.880	0.699	0.905	Different species
Thermoanaerobacter sp. X561-*Cal. subterraneus* subsp. *tengcongensis* MB4	0.711	0.790	0.748	0.830	0.232	0.294	0.867	0.770	0.955	Different species
Thermoanaerobacter sp. X561-*Th. thermosaccharolyticum* DSM 571	0.646	0.728	0.657	0.774	0.003	0.101	0.879	0.699	0.905	Different species
Cal. subterraneus subsp. *tengcongensis* MB4-*Th. thermosaccharolyticum* DSM 571	0.635	0.716	0.651	0.763	-0.018	0.064	0.883	0.680	0.865	Different species

[a]Species prediction is consistent for all models. A cutoff of 70% genome sequence identity was used to determine species identity, consistent with Zeigler (2003). For determination of species identity using JSpecies, a cutoff of 95% ANIm was used to determine species identity (Richter & Rosselló-Móra, 2009).

[b]ND, not determined due to the lack of full or partial gene/genome sequence at time of analyses.

[c]*Cal. subterraneus* subsp. *tengcongensis* is a reclassification of the strain *Thermoanaerobacter tengcongensis* (Fardeau *et al*., 2004).

[d]*Th. thermosaccharolyticum* DSM 571 was selected as a phylogenetically related outgroup for comparison purposes.

Table 3.2. Gene sequence identities and prediction of genome sequence identities for the *Thermoanaerobacter* isolates using the three-gene model of Zeigler and the one-gene model based on the *cpn60* UT: summary data showing the range exhibited by WC1-WC12.

	prediction of genomic SI using						
Pair	SI *recN*	SI *rpoA*	SI *thdF*	SI *cpn60* UT	3-gene model	1-gene model	Species prediction[a]
Isolates as a group (low -- WC3-WC8)[b]	0.998	0.999	0.998	0.975	0.936	0.793	Same species
Isolates as a group (high -- WC9-WC10)	1.000	0.997	0.998	0.998	0.936	0.874	Same species
T. thermohydrosulfuricus DSM 567-isolates (low -- WC8)	0.995	0.999	0.998	0.978	0.934	0.805	Same species
T. thermohydrosulfuricus DSM 567-isolates (high -- WC6)	0.989	0.999	0.999	0.996	0.932	0.868	Same species
T. brockii subsp. *brockii* HTD4-isolates (low -- WC3)	0.966	0.969	0.875	0.873	0.684	0.444	Different species
T. brockii subsp. *brockii* HTD4-isolates (high -- WC1)	0.969	0.967	0.873	0.893	0.681	0.512	Different species
T. pseudethanolicus 39E-isolates (low -- WC3)	0.966	0.969	0.879	0.873	0.691	0.444	Different species
T. pseudethanolicus 39E-isolates (high -- WC1)	0.969	0.967	0.877	0.893	0.688	0.512	Different species

[a]Species prediction is consistent for both models. A cutoff of 70% genome sequence identity was used to determine species identity, consistent with Zeigler (2003).

[b]"low" and "high" refer to the extremes of the ranges of sequence identities observed in the groups of sequences compared. Based on the sequence identities of the *cpn60* UT. The isolate(s) that exhibited these ranges are indicated.

567, and both the three-gene model of Zeigler and the one-gene *cpn60* UT-based model predicted that the isolates should be classified as *T. thermohydrosulfuricus*. Furthermore, both models predicted that *T. thermohydrosulfuricus* was a distinct species from both *T. pseudethanolicus* 39E and *T. brockii* subsp. *brockii* HTD4 as the type strains shared a genome sequence identity of under 70% (Table A.2.2).

Of the genes studied, *recN* and *rpoA* showed very high sequence conservation between *T. brockii* subsp. *brockii* HTD4, *T. pseudethanolicus* 39E, *T. thermohydrosulfuricus* DSM 567 and the WC isolates, while *thdF* and *cpn60* UT were much more diverse. Zeigler reported that *rpoA* was the most highly conserved in the three-gene model, followed by *thdF* and *recN*, and that a one-gene model based solely on *recN* sequences was nearly as good a predictor as the three-gene model (2003). We found that, for the *Thermoanaerobacter* species analyzed, *recN* was highly conserved and genomic identity prediction based only on *recN* predicted that *T. brockii* subsp. *brockii* HTD4, *T. pseudethanolicus* 39E, *T. thermohydrosulfuricus* DSM 567, and all the isolates belonged together in a single species (Table A.2.3).

To resolve this discrepancy, phylogenetic trees were constructed based on the sequences of *recN*, *rpoA*, and *thdF*. As shown in Fig. A.2.1 – Fig. A.2.3, the phylogeny for each locus showed with high bootstrap support that the WC isolates clustered with *T. thermohydrosulfuricus* DSM 567, while *T. brockii* subsp. *brockii* HTD4, *T. ethanolicus* CCSD1, and *T. pseudethanolicus* 39E formed a separate clade. This is in agreement with the phylogeny that was constructed based on the *cpn60* UT (Chapter 2). The phylogeny predicted using the one-gene model based on *cpn60* UT was consistent with the three-gene model based on *recN*, *rpoA*, and *thdF*. The high sequence identities observed at the

recN locus were not due to the shorter fragments of the *recN* gene that were analyzed (all sequences were trimmed to 616 bp), since the same results were observed using as much as 1,543 bp (88% of the gene length) of the *recN* sequence. The SI_{recN} between *T. thermohydrosulfuricus* DSM 567 and *T. brockii* subsp. *brockii* HTD4 was 0.970 when comparing 88% of the gene and 0.966 when comparing the 616 bp fragment.

To determine if the *cpn60* UT genomic prediction model was applicable to other taxa, the predicted species identity was calculated for 39 comparisons of the whole genomes reported by Richter and Rosselló-Móra (2009), and compared to the reported ANIm, ANIb, and Tetra values. These comparisons represented a broad taxonomic range of five bacterial phyla. As shown in Table 3.3, the predictions based only on the *cpn60* UT agreed with those based on whole genome comparisons in nearly all cases, with only two discrepancies in the 39 comparisons. Moreover, the two discrepancies that were noted were both borderline cases where the ANIm or the predicted genome identities were near the cut-off threshold. The comparison of *Bacillus cereus* ATCC 10987 to *B. cereus* 14579^T was particularly interesting, since whole genome sequencing of the former strain revealed high genetic relatedness to *Bacillus anthracis* Ames (Rasko *et al*., 2004). This relatedness was reflected in the higher *cpn60* UT sequence identities of *B. cereus* 10987 and *B. anthracis* Ames ($SI_{cpn60\ UT} = 0.995$). Comparison of *Shewanella baltica* OS155 to *Shewanella putrefaciens* resulted in a predicted genomic sequence identity of 0.73 (just above the threshold defined by Zeigler), while whole genome alignments revealed these to be distinct species. In all other cases, the prediction of genomic sequence identities based on the *cpn60* UT prediction model and the species designation based on ANIm were in perfect agreement.

Table 3.3. Application of the *cpn60* UT species prediction model to other taxa[a].

			Species prediction		Genomic parameters (JSpecies)			
Pair	$SI_{cpn60\ UT}$	SI_{genome} (*cpn60* UT)	1-gene model	3-gene model[b]	ANIm	Tetra	ANIb	Species prediction
Phylum: *Firmicutes*								
Bacillus cereus ATCC 10987 - *Bacillus cereus* ATCC 14579^T	0.975	0.793	Same species	Same species	0.924	0.999	0.909	Different species
Streptococcus agalactiae 2603 V/R - *Streptococcus agalactiae* A909	0.987	0.836	Same species	Same species	0.994	0.999	0.988	Same species
Streptococcus agalactiae 2603 V/R - *Streptococcus agalactiae* NEM316	0.998	0.874	Same species	Same species	0.994	0.998	0.984	Same species
Streptococcus agalactiae A909 - *Streptococcus agalactiae* NEM 316	0.980	0.812	Same species	Same species	0.993	0.998	0.986	Same species
Phylum: *Bacteroidetes/Chloribi* group								
Bacteroides fragilis NCTC 9343 - *Bacteroides fragilis* YCH46	0.998	0.874	Same species	Same species	0.992	1.000	0.987	Same species
Bacteroides fragilis NCTC 9343 - *Bacteroides thetaiotomicron* VPI-5482	0.873	0.442	Different species	Different species	0.843	0.965	0.740	Different species
Bacteroides thetaiotomicron VPI-5482 - *Bacteroides fragilis* YCH46[c]	0.876	0.455	Different species	Different species	0.847	0.966	0.740	Different species
Phylum: *Alphaproteobacteria*								
Brucella canis ATCC 23365 - *Brucella melitensis* ATCC 23457 biovar 2	0.998	0.874	Same species	Same species	0.996	1.000	0.997	Same species
Brucella canis ATCC 23365 - *Brucella ovis* ATCC 25840	0.998	0.874	Same species	Same species	0.996	1.000	0.996	Same species
Brucella canis ATCC 23365 - *Brucella suis* 1330	1.000	0.880	Same species	Same species	0.999	1.000	0.999	Same species
Brucella melitensis ATCC 23457 biovar 2 - *Brucella ovis* ATCC 25840	0.996	0.868	Same species	Same species	0.995	1.000	0.996	Same species

Table 3.3 cont.

Pair	$SI_{cpn60\ UT}$	SI_{genome} (*cpn60* UT)	1-gene model	3-gene model[b]	ANIm	Tetra	ANIb	Species prediction
Brucella melitensis ATCC 23457 biovar 2 - *Brucella suis* 1330	0.998	0.874	Same species	Same species	0.997	1.000	0.997	Same species
Brucella ovis ATCC 25840 - *Brucella suis* 1330	0.998	0.874	Same species	Same species	0.996	1.000	0.996	Same species
Phylum: *Betaproteobacteria*								
Nitrosomonas europea ATCC 19718 - *Nitrosomonas eutropha* C91	0.856	0.384	Different species	Different species	0.832	0.925	0.789	Different species
Phylum: *Gammaproteobacteria*								
Escherichia coli CFT073 - *Escherichia coli* E2348/69	0.993	0.855	Same species	Same species	0.984	0.999	0.976	Same species
Escherichia coli CFT073 - *Escherichia coli* K12 MG1655	0.984	0.824	Same species	Same species	0.972	0.997	0.961	Same species
Escherichia coli CFT073 - *Escherichia coli* O157:H7 EDL933	0.982	0.818	Same species	Same species	0.967	0.998	0.960	Same species
Escherichia coli CFT073 - *Escherichia coli* 042	0.987	0.837	Same species	Same species	0.970	0.999	0.959	Same species
Escherichia coli CFT073 - *Shigella flexneri* 2a 2457^T	0.986	0.830	Same species	Same species	0.970	0.998	0.959	Same species
Escherichia coli CFT073 - *Shigella sonnei* 53G	0.984	0.824	Same species	ND[d]	0.970	0.996	0.959	Same species
Pseudomonas aeruginosa PAO1 - *Pseudomonas fluorescens* Pf-5	0.879	0.465	Different species	Different species	0.840	0.943	0.758	Different species
Pseudomonas aeruginosa PAO1 - *Pseudomonas fluorescens* PfO1	0.877	0.459	Different species	Different species	0.839	0.911	0.749	Different species
Pseudomonas aeruginosa PAO1 - *Pseudomonas fluorescens* SBW25	0.883	0.477	Different species	Different species	0.838	0.916	0.745	Different species
Pseudomonas aeruginosa PAO1 - *Pseudomonas syringae* pv. Syringae B728a	0.868	0.428	Different species	Different species	0.837	0.820	0.739	Different species

Table 3.3 cont.

Pair	$SI_{cpn60\ UT}$	SI_{genome} (*cpn60* UT)	1-gene model	3-gene model[b]	ANIm	Tetra	ANIb	Species prediction
Pseudomonas aeruginosa PAO1 - *Pseudomonas syringae* pv. Tomato DC3000	0.863	0.409	Different species	Different species	0.836	0.806	0.736	Different species
Pseudomonas fluorescens PfO1 - *Pseudomonas fluorescens* Pf-5	0.923	0.613	Different species	Different species	0.858	0.911	0.806	Different species
Pseudomonas fluorescens PfO1 - *Pseudomonas fluorescens* SBW25	0.930	0.638	Different species	Different species	0.859	0.940	0.802	Different species
Shewanella amazonensis SB2B - *Shewanella oneidensis* MR-1	0.843	0.341	Different species	Different species	0.830	0.696	0.709	Different species
Shewanella amazonensis SB2B - *Shewanella putrefaciens* CN-32	0.827	0.285	Different species	Different species	0.823	0.649	0.705	Different species
Shewanella amazonensis SB2B - *Shewanella woodyi* ATCC 51908	0.805	0.211	Different species	Different species	0.831	0.647	0.692	Different species
Shewanella baltica OS155 - *Shewanella baltica* OS185	0.977	0.799	Same species	Same species	0.969	0.999	0.966	Same species
Shewanella baltica OS155 - *Shewanella baltica* OS195	0.980	0.812	Same species	Same species	0.968	0.999	0.965	Same species
Shewanella baltica OS155 - *Shewanella baltica* OS223	0.984	0.824	Same species	Same species	0.971	0.999	0.966	Same species
Shewanella baltica OS155 - *Shewanella putrefaciens* CN-32	0.955	0.725	Same species	Different species	0.867	0.968	0.826	Different species
Shewanella baltica OS155 - *Shewanella oneidensis* MR-1T	0.901	0.539	Different species	Different species	0.858	0.972	0.783	Different species
Phylum: *Thermotogae*								
Thermotoga maritima MSB8 - *Thermotoga neapolitana* DSM 4359	0.850	0.363	Different	ND[e]	0.872	0.961	0.800	Different species
Thermotoga maritima MSB8 - *Thermotoga petrophila* RKU-1	0.955	0.724	Same species	ND[e]	0.953	0.998	0.948	Same species
Thermotoga neapolitana DSM 4359 - *Thermotoga petrophila* RKU-1	0.832	0.300	Different species	ND[e]	0.857	0.964	0.790	Different species

Table 3.3 cont.

Pair	$SI_{cpn60\ UT}$	SI_{genome} (*cpn60* UT)	1-gene model	3-gene model[b]	ANIm	Tetra	ANIb	Species prediction
Phylum: *Deinococcus-Thermus*								
Deinococcus deserti VCD115 - *Deinococcus radiodurans* R1	0.888	0.494	Different species	Different species	0.841	0.764	0.738	Different species

[a]Pairs and JSpecies-generated ANIm, ANIb, and Tetra values were taken from Richter and Roselló-Móra (2009).

[b]SI values for *recN, rpoA and thdF* can be found in Table A.2.4.

[c]Comparison not reported by Richter and Rosselló-Móra (2009), but provided in the present study.

[d]ND, not determined. Genome sequence for *S. sonnei* 53G not available in GenBank at time of analyses.

[e]ND, not determined. DNA recombination and repair in *Thermotoga* species occurs via *rad*-type genes (Harvey *et al*., 2002) and no *recN* ortholog was identified.

For two strains to have a predicted genome sequence identity of 70%, and therefore being considered the same species, a *cpn60* UT sequence identity threshold between 0.94-0.95 is necessary. This is in very good agreement with Richter and Rosselló-Móra, who suggested that two strains sharing an ANIm of 95% (along with a correspondingly high Tetra score) should be considered the same species (2009). Species prediction using the *cpn60* UT was quite accurate compared to whole genome comparisons, but was somewhat less conservative. This suggests that species prediction using the *cpn60* UT may be an excellent "screening tool" for determining candidate strains for whole or partial genome sequencing in certain cases. In screening a large number of environmental isolates, whole genome sequence identity prediction by *cpn60* UT can be a complementary tool to select candidates for larger genome sequencing efforts.

In summary, a mathematical model was developed for predicting genome sequence identities using the methods of Zeigler (2003) based on the determination of the sequence identities of the 549–567 bp *cpn60* UT. This method is rapid and simple, the sequences can be easily determined using a single set of sequencing reactions and is accessible using a universal set of amplification primers. Further, the results are in agreement with the best three-gene model determined by identifying candidate genes from complete genomes, as well as with determinations of ANI over the whole genome, which has been designated as a "new gold standard" for species identity determinations (2009). While this method would not be expected to replace established and emerging methodologies for the determination of novel bacterial species, the prediction of genome identities based on *cpn60* UT sequence identity can be a rapid way to determine if an

isolate belongs in an established species or if further experimentation to designate a novel species is warranted.

3.5 Authors' contributions

Experiments were conceived and designed by Tobin J. Verbeke, Tim J. Dumonceaux, Janet E. Hill, David B. Levin and Richard Sparling. DNA extraction, primer design and PCR amplification, phylogenetic analyses and pairwise-comparisons were additionally performed by Tobin J. Verbeke. DNA sequencing reactions and determination were done by Tim J. Dumonceaux. Model development was done jointly by Tobin J. Verbeke and Tim J. Dumonceaux, while JSpecies analyses were done by Mathew G. Links. The original manuscript, upon which this chapter is based, was jointly authored by Tobin J. Verbeke and Tim J. Dumonceaux, with input from David B. Levin and Richard Sparling.

3.6 Acknowledgements

We thank Florian Labat for technical assistance and Sean Hemmingsen for critical comments on this manuscript. This work was funded by a Natural Sciences and Engineering Research Council (NSERC) Strategic Grant (STPGP 365076), Genome Canada, and by the Cellulosic Biofuels Network (Agriculture and Agri-Food Canada).

Chapter 4. Genomic evaluation of *Thermoanaerobacter* spp. for the construction of designer co-cultures to improve lignocellulosic biofuel production[11]

4.1 Abstract

The microbial production of ethanol from lignocellulosic biomass is a multi-component process that involves biomass hydrolysis, carbohydrate transport and utilization, and finally, the production of ethanol. Strains of the genus *Thermoanaerobacter* have been studied for decades due to their innate abilities to produce comparatively high ethanol yields from hemicellulose constituent sugars. However, their inability to hydrolyze cellulose limits their usefulness in lignocellulosic biofuel production. As such, co-culturing *Thermoanaerobacter* spp. with cellulolytic organisms is a plausible approach to improving lignocellulose conversion efficiencies and yields of biofuels. To evaluate native lignocellulosic ethanol production capacities relative to competing fermentative end-products, comparative genomic analysis of 11 sequenced *Thermoanaerobacter* strains, including a *de novo* genome, *Thermoanaerobacter thermohydrosulfuricus* WC1, was conducted. Analysis was specifically focused on the genomic potential for each strain to address all aspects of ethanol production mentioned through a consolidated bioprocessing approach. Whole genome functional annotation analyses identified three distinct clades within the genus.

[11] A modified version of this chapter has previously been published. Reproduced with permission: © 2013, Tobin J. Verbeke. Found in: **Verbeke TJ, Zhang X, Henrissat B, Spicer V, Rydzak T, Krokhin OV, Fristensky B, Levin DB, Sparling R**. 2013. Genomic evaluation of *Thermoanaerobacter* spp. for the construction of designer co-cultures to improve lignocellulosic biofuel production. PLoS One **8**: e59362.

The genomes of Clade 1 strains encode the fewest extracellular carbohydrate active enzymes and also show the least diversity in terms of lignocellulose relevant carbohydrate utilization pathways. However, these same strains reportedly are capable of directing a higher proportion of their total carbon flux towards ethanol, rather than non-biofuel end-products, than other *Thermoanaerobacter* strains. Strains in Clade 2 show the greatest diversity in terms of lignocellulose hydrolysis and utilization, but proportionately produce more non-ethanol end-products than Clade 1 strains. Strains in Clade 3, in which *T. thermohydrosulfuricus* WC1 is included, show mid-range potential for lignocellulose hydrolysis and utilization, but also exhibit extensive divergence from both Clade 1 and Clade 2 strains in terms of cellular energetics. The potential implications regarding strain selection and suitability for industrial ethanol production through a consolidated bioprocessing co-culturing approach are examined throughout the chapter.

4.2 Introduction

Consolidated bioprocessing (CBP), whereby microbial enzyme production, biomass hydrolysis and substrate conversion to biofuels all occurs in a single bioreactor, offers improved economic feasibility and process efficiencies compared with alternative approaches to lignocellulosic biofuel production (Lynd *et al*., 2005; Cardona & Sánchez, 2007; Olson *et al*., 2012). However, no single organism has yet been identified that is capable of performing all of these tasks at industrially significant levels (Desvaux, 2006). The utilization of designer co-cultures, allowing for potential complementary or synergistic phenotypes between multiple organisms to improve process efficiencies and

yields, is an alternative strategy to potentially overcome these limitations (reviewed by Brenner *et al*. 2008; Zuroff & Curtis, 2012).

From a CBP standpoint, the use of thermophilic, anaerobes belonging to the *Firmicutes* offers many advantages. Growth at elevated temperatures reduces energy costs by avoiding repeated heating and cooling steps associated with cycling between microbial growth and both upstream pre-processing as well as downstream product recovery. Additionally, the native capacity for many strains to produce lignocellulose hydrolytic enzymes *in situ* may reduce or eliminate the need for enzymatic pre-treatment of biomass. Finally, the ability of multiple strains to ferment a broad range of lignocellulose constituent saccharides into ethanol allows for efficient conversion and utilization of the biomass.

To date, the organisms garnering the most attention for thermophilic CBP include strains in the orders *Clostridiales* or *Thermoanaerobacteriales* (Lynd *et al*., 2008). One of these organisms, *Clostridium thermocellum*, has been at the forefront of lignocellulosic ethanol research for decades due to its high cellulolytic capabilities (Lynd *et al*., 1989; Levin *et al*., 2006; Raman *et al*., 2009). However, the inability of *C. thermocellum* to ferment pentose sugars resulting from hemicellulose hydrolysis reduces potential biomass conversion efficiencies and represents a major limitation in its development as an industrial microorganism. Various strains of the genus *Thermoanaerobacter*, on the other hand, are known to hydrolyze hemicellulose and ferment hemicellulose constituent sugars (Wiegel & Ljungdahl, 1981; Kozianowski *et al*., 1997; Larsen *et al*., 1997; Onyenwoke *et al*., 2007; Chapter 2), naturally produce (Wiegel & Ljungdahl, 1981; Zeikus *et al*., 1980) and tolerate (Larsen *et al*., 1997; Georgieva *et al*., 2007a)

comparatively high ethanol concentrations, and are amenable to genetic manipulation for purposes of further improving biofuel yields (Shaw *et al*., 2010; Yao & Mikkelsen, 2010). Furthermore, previous studies have reported that cellulose degradation rates and overall biofuel yields are improved by *C. thermocellum* - *Thermoanaerobacter* sp. co-cultures as compared to *C. thermocellum* mono-cultures (Ng *et al*., 1981; Sadder & Chan 1984; He *et al*., 2011). Therefore, the use of a *Thermoanaerobacter* strain as an industrially relevant co-culture partner for *C. thermocellum* shows great potential as a CBP strategy.

The genomes of multiple *Thermoanaerobacter* spp. are currently available publicly. The purpose of the current chapter is to conduct genus wide comparative genomic analysis of all available genomes, including a *de novo* genome from *Thermoanaerobacter thermohydrosulfuricus* WC1 (Chapters 2, 3), and evaluate the potential of each strain as a *C. thermocellum* co-culture partner. Factors evaluated include the capacity for lignocellulose hydrolysis, transport of the resulting hydrolysis products, carbohydrate utilization and the potential to produce ethanol relative to other fermentation end-products. Further, given that these processes are interconnected with cellular energy metabolism (Fig. 4.1), the potential mechanisms for energy conservation will also be evaluated.

4.3 Materials and Methods

4.3.1 DNA extraction, genome sequencing and assembly

In lab glycerol stock cultures of *T. thermohydrosulfuricus* WC1 were revived, grown overnight and plated as previously described (Chapter 2.3.3). A single colony was

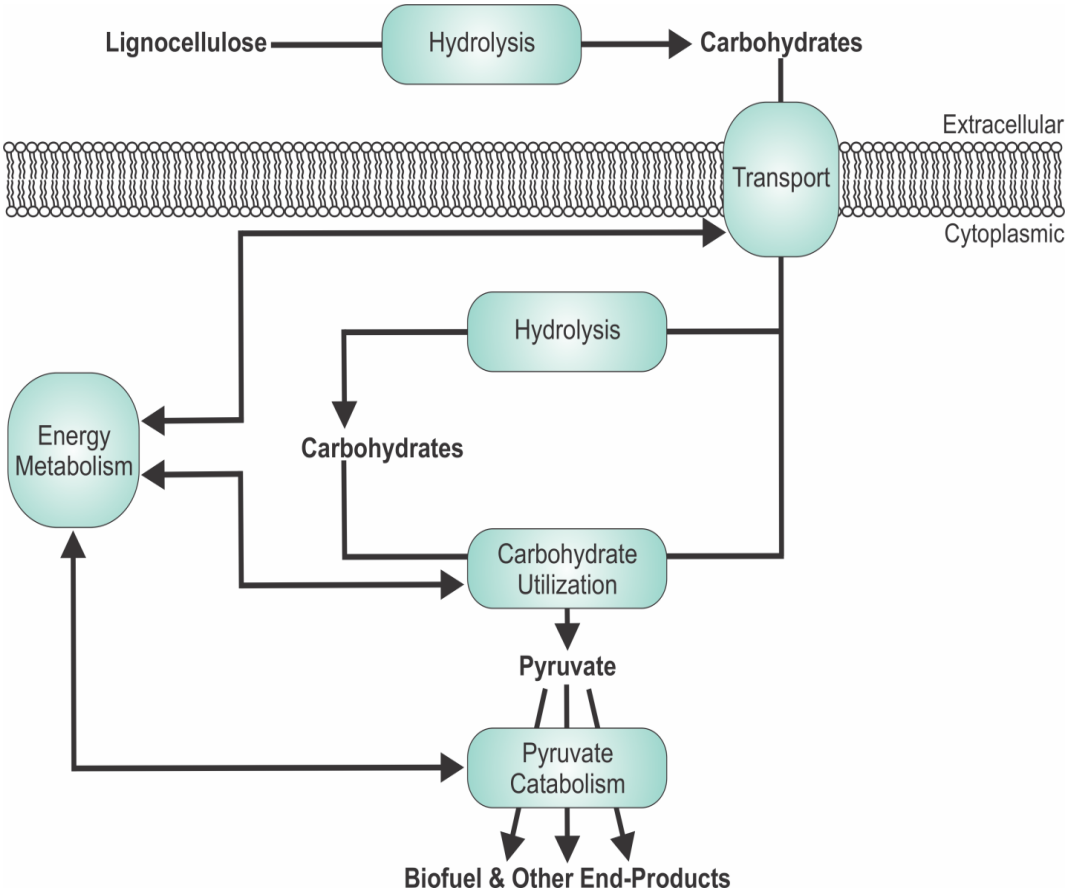


Fig. 4.1. Schematic representation of key physiological processes pertinent to lignocellulosic ethanol production in a CBP system. Physiological processes are identified as the text in blue boxes.

picked and used to inoculate fresh ATCC 1191 liquid medium (10 ml) containing 2 g/L cellobiose. The resulting culture was allowed to grow overnight and gDNA was extracted as done previously (Chapter 2.3.6).

The genome of *T. thermohydrosulfuricus* WC1 was first sequenced at the McGill University and Genome Quebec Innovation Centre using shotgun 454 pyrosequencing (Margulies *et al*., 2005) to get 121,690 reads with an average length of 626 bp. Genome assembly by Newbler v2.6 generated 123 contigs (size >100 bp) with the longest contig 182,175 bp with ~15x depth coverage. To improve gap closure and perform scaffolding of the contigs, genome re-sequencing was conducted with Illumina HiSeq 2000 paired-end technologies generating 122,042,203 reads of 100 bp from each end with an average insert size of 277 bp. Illumina reads were pre-processed by adaptor clipping, quality trimming with the FASTX-Toolkit (Pearson *et al*., 1997) and random subset data selection. Multiple assembly pipelines combining both 454 and Illumina data were evaluated including Optimized-Velvet (Zerbino & Birney, 2008), Ray (Boisvert *et al*., 2010) and Newbler v.2.6. Based on statistics metrics, Newbler v2.6 generated the longest contigs with the highest N50 and the resulting assembly was composed of 47 scaffolds plus an additional 15 unscaffolded contigs (size >100 bp with the longest contig 261,431 bp) with ~56x depth coverage.

4.3.2 Genome annotation and proteogenomic analysis

The assembled DNA scaffolds and contigs were submitted to the Joint Genome Institute's (JGI) Integrated Microbial Genomes-Expert Review (IMG-ER) platform (Markowitz *et al*., 2009) for gene calling and annotation using their automated pipeline

(http://img.jgi.doe.gov/w/doc/img_er_ann.pdf). The annotated genome was subsequently submitted to the JGI's GenePRIMP pipeline (Pati *et al*., 2010), which reported 358 anomalies. All anomalies were manually curated.

Proteogenomic analysis provided supporting evidence for manual curation decisions regarding reported GenePRIMP anomalies. A mechanism of genome assembly independent proteomics, similar to the approach by Krug *et al*. (2011), was implemented for this analysis. In brief, raw genomic sequencing reads for *T. thermohydrosulfuricus* WC1, rather than an assembled genome, were transcribed into amino acid sequences in all six possible reading frames with all potential STOP codons being reported as new elements. These elements were subjected to *in silico* tryptic digestion to create a "naïve" peptide database. This database was used to search a 2D-HPLC-MS/MS experimental output, generated from mid-exponential phase *T. thermohydrosulfuricus* WC1 cells grown in liquid ATCC 1191 medium (as described above), using a high-performance GPU-based identification engine developed for this project (McQueen *et al*., 2012). Correlating observed peptide retention times against their computed hydrophobicities as described by Krokhin (2006) further supported the assignment confidences. As the resultant *in silico* naïve peptide collection was wholly disconnected from their source proteins, the peptide database served as a validation for the genome assembly and annotation workflows. Observed peptides were compared against the annotation-derived protein database and those unique to the "naïve" database search were reported. Unique peptides (869), corresponding to unannotated protein coding sequences (CDS) within the genome, were used to support or refute GenePRIMP identified annotation anomalies. Modifications to the annotation based upon these peptide data are reported as "notes"

within the GenBank file associated with the genome. The genome has been deposited at DDBJ/EMBL/GenBank databases under accession number AMYG00000000. The version described in this paper is the first version, AMYG01000000.

4.3.3 Comparative genomic analyses

Comparative genomic analyses were conducted on genomes and gene annotations available (as of July, 2012) using the IMG-ER platform unless specified elsewhere. Genes of interest, with the exception of transporters and carbohydrate active enzymes (CAZymes), were identified within the IMG-ER annotated genomes using independent searches for the Clusters of Orthologous Groups (COG) (Tatusov *et al*., 2000), KEGG Orthology (KO) (Kanehisa *et al*., 2008) and TIGRFAMs (Haft, 2003) unique identifiers. Transporters were identified using the assignment criteria given by the Transporter Classification Database (TCDB) (Saier *et al*., 2009) as part of the IMG-ER system and substrate specificities were inferred based upon KO annotations of the same genes. The annotation accuracy for all genes identified using the above search methods was manually assessed using a combination of genomic contextual analysis, sequence alignments and through literature and database searches. Gene annotations using additional databases made available through the IMG-ER annotation pipeline including Pfams (Finn *et al*., 2010), IMG Terms (Markowitz *et al*., 2009), the SEED (Overbeek *et al*., 2009) and Interpro (Hunter *et al*., 2009) were used in the manual assessment of COG, KO and TIGRFAM designations when appropriate. All functional annotations designations should be considered to be putative assignments, unless specifically discussed.

Analysis of CAZymes was conducted by accessing the CAZy database (Cantarel *et al*., 2009) for all *Thermoanaerobacter* genomes available in the database or analysed *de novo*. For *Thermoanaerobacter ethanolicus* CCSD1 and *Thermoanaerobacter* sp. X561 the unfinished draft genomes were downloaded from the NCBI database and were analyzed using the standard detection pipeline of the CAZy database. Substrate specificities for all CAZymes were not inferred unless specifically discussed (Chapter 4.4.3) and were instead limited to substrates reported for each enzyme class by the CAZy database. The subcellular localization of identified CAZymes was predicted by uploading fasta files of all identified gene amino acid sequences into the PSORTb 3.0 database (Yu *et al*., 2010) and using the final localization predictions.

4.3.4 Sequence and phylogenetic analysis

Sequence alignments were performed using BioEdit v.7.0.9.0 (Hall, 1999). Phylogenetic analysis of individual sequences was performed using MEGA 4 (Tamura *et al*., 2007). Phylogeny was inferred using the neighbour-joining method with evolutionary distances calculated via the Poisson correction as described (Calusinska *et al*., 2010). Alignment gaps were deleted via pair wise sequence comparisons only and clusters grouped using the bootstrapping method (10,000 replicates).

4.4 Results and discussion

4.4.1 Genome properties of *T. thermohydrosulfuricus* WC1

The draft genome sequence of *T. thermohydrosulfuricus* WC1 is comprised of 2,573,514 bp and shows a G + C content of 34.35%, which is consistent with other

sequenced *Thermoanaerobacter* strains (Table A.3.1). There are 2,655 genes annotated with 2,552 predicted to be CDS. The chaperonin-60 universal target (*cpn*60 UT) nucleotide sequence, which has been shown to be a more reliable phylogenetic indicator than 16S rRNA encoding DNA for the *Thermoanaerobacter* genus (Chapter 3), agrees 100% with the previously published sequence (Chapter 2). Based on the standards for genomic sequencing projects (Chain *et al*., 2009), the *T. thermohydrosulfuricus* WC1 permanent draft-genome belongs to the "Annotation-Directed Improvement" classification.

4.4.2 Whole genome comparative analysis

The genomes of 11 *Thermoanaerobacter* strains (10 publicly available as of July, 2012 + *T. thermohydrosulfuricus* WC1) were included in our genus wide comparison. The *Thermoanaerobacter tengcongensis* MB4 genome (Bao *et al*., 2002), corresponding to GenBank accession number AE008691, was not included based on the reclassification of the strain into the genus *Caldanaerobacter* (Fardeau *et al*., 2004). The available genome sequences for the genus, in draft, permanent draft and finished states, range in size from 2.20 Mb – 2.78 Mb and are annotated to contain anywhere from 2,286 – 2,800 CDS (Table A.3.1).

To determine the extent that similar functional profiles exist between strains, hierarchical clustering, based upon COG, KO and TIGRFAM qualifiers, was conducted independently for each qualifier. As shown in Fig. 4.2, three distinct clades exist within the genus. Clade 1 contains the most genomes and could also potentially be divided into sub-clades. However, it was previously shown (Chapter 3) that all strains in Clade 1

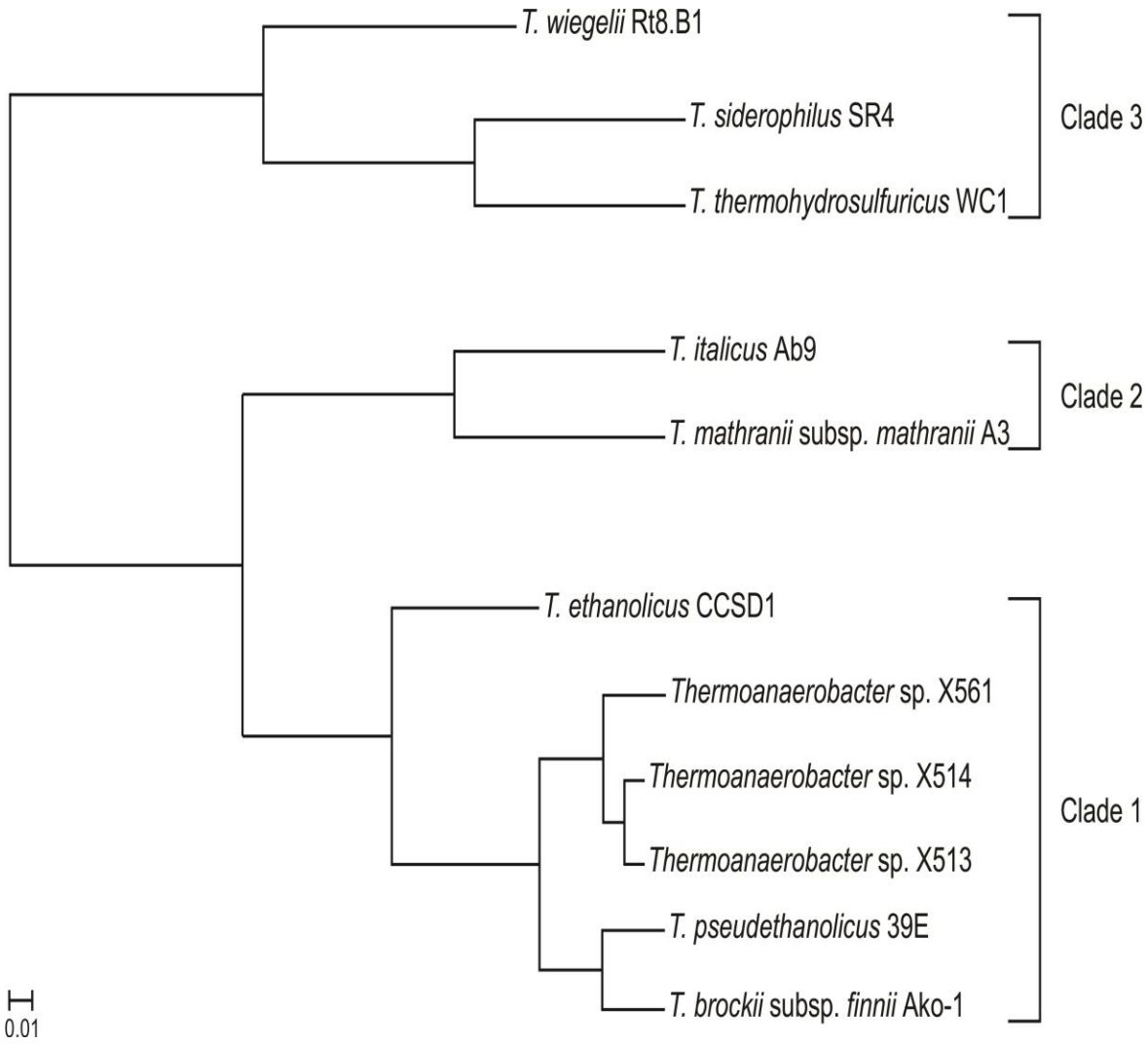


Fig. 4.2. Phylogram of annotated COG functional profiles for sequenced *Thermoanaerobacter* strains. Cluster analysis, based on Cluster 3.0 analysis software (de Hoon *et al*., 2004), was conducted within the IMG-ER platform using the COG profiles for each genome. Branch lengths correspond to calculated distances between functional profiles. Similar clade architectures are also observed when using KO (Fig. A.3.1) or TIGRFAM (Fig. A.3.2) descriptors.

share an average nucleotide identity (ANIm) score (Richter & Rosselló-Móra, 2009) greater than 0.97 and thus, may represent strains of the same species. Previous work by Hemme *et al*., (2011) with *Thermoanaerobacter pseudethanolicus* 39E and *Thermoanaerobacter* sp. X514, have shown that genomes with ANIm scores greater than 0.97 may have significant differences that impact relative strain suitability for lignocellulosic ethanol production. Inter-clade functional divergence may therefore have an even more drastic impact on co-culturing potential than intra-clade divergence. Specifics regarding strain suitability are discussed below.

4.4.3 CAZyme analyses

The availability of fermentable sugars from lignocellulosic biomass is dependent on CAZymes capable of degrading the insoluble, extracellular carbohydrate polymers (Fig. 4.1). In designer co-cultures, the multiple hemicellulase enzymes in the *C. thermocellum* cellulosome may help to liberate hemicellulose constituent sugars from the polymeric backbone. The liberation of such saccharides, most notably pentoses, but also a mixture of hexoses, may subsequently be fermented to ethanol by a *Thermoanaerobacter* strain. However, expression analyses studies of *C. thermocellum* grown on various substrates by Raman *et al*. (2009) have shown that xylanase expression levels were at their lowest when grown on a hemicellulose containing substrate (switchgrass) in comparison to growth on cellobiose or cellulose. Therefore, under CBP conditions, the utilization of a *Thermoanaerobacter* strain, which has xylan hydrolysis capabilities, may help facilitate biomass hydrolysis similar to what has been reported in

co-cultures of *C. thermocellum* and *Caldicellulosiruptor bescii* grown on switchgrass or cellulose plus xylan (Kridelbaugh *et al*., 2013).

To examine the potential that a *Thermoanaerobacter* strain may contribute to biomass hydrolysis, all CAZyme genes were identified within the genus. As shown in Table A.3.2, between 45-61 independent CAZyme gene sequences, belonging to 27-40 distinct CAZyme classes, were identified within the sequenced strains.

A small subset (GH23 and various glycosyltransferases) of these genes encode proteins typically associated with cellular maintenance functions such as peptidoglycan processing and glycogen synthesis. Furthermore, based on the PSORTB 3.0 predictions, only a small fraction of each CAZyome is predicted to be localized extracellularly (Table 4.1) such that only a few gene products per strain would be capable of contributing to the extracellular hydrolysis of complex polymers.

4.4.3.1 Extracellular, lignocellulose hydrolyzing CAZymes.

As shown in Table 4.1, no predicted extracellular lignocellulose hydrolyzing CAZymes were identified in Clade 1 strains. Moreover, only 3 strains, *T. italicus* Ab9, *T. mathranii* subsp. *mathranii* A3 and *T. thermohydrosulfuricus* WC1, possess extracellular endo-xylanases capable of hydrolyzing the xylan backbone of many hemicelluloses. All three strains contain a secretable multi-component CDS with the modular structure CBM22-CBM22-GH10-CBM9-CBM9-SLH-SLH-SLH. To date, proteins belonging to the GH10 family are only known to act on xylan, while CBM22 and CBM9 modules are considered to be primarily xylan binding. All orthologs show high amino acid sequence similarity (>74%) and similar modular structure to the functionally characterized *xyn*A

Table 4.1. Predicted extracellular CAZymes[a] involved with lignocellulosic biomass hydrolysis within sequenced *Thermoanaerobacter* spp.

		CAZyme designation and modular structure						
Clade	Strain	GH10	GH43	GH52	CE4	PL9	GH66-CBM35-CBM35-GH15	CBM22-CBM22-GH10-CBM9-CBM9-SLH-SLH-SLH
1	*T. brockii* subsp. *finnii* Ako-1							
	T. ethanolicus CCSD1							
	T. pseudethanolicus 39E							
	Thermoanaerobacter sp. X513							
	Thermoanaerobacter sp. X514							
	Thermoanaerobacter sp. X561							
2	*T. italicus* Ab9	0188		0190		1727		0192
	T. mathranii subsp. *mathranii* A3	0247	1695	0249				0251
3	*T. siderophilus* SR4				1967		0902	
	T. thermohydrosulfuricus WC1			1012	1808		0529	1010
	T. wiegelii Rt8.B1				00017610			

[a]Only the numerical suffix of locus tags are identified. Corresponding prefixes for each strain are identified in Table A.3.1.

gene in *Thermoanaerobacterium saccharolyticum* NTOU1 (Hung *et al*., 2011). *T. italicus* Ab9 (Thit_0188) and *T. mathranii* subsp. *mathranii* A3 (Tmath_0247) also contain an additional extracellular GH10 family gene showing high amino acid sequence similarity (>77%) with the cloned and characterized xylanase/β-xylosidase from *Caldicellulosiruptor saccharolyticus* DSM 8903 (Lüthi *et al.*, 1990).

The hydrolysis products resulting from GH10 mediated endo-xylanase activity would be xylo-oligomers. The generation of xylose monomers from xylo-oligomers requires extracellular and/or intracellular β-xylosidase activity. The *T. italicus* Ab9, *T. mathranii* subsp. *mathranii* A3 and *T. thermohydrosulfuricus* WC1 genomes also encode for an extracellular GH52 family enzyme (Table 4.1, Table A.3.2), which, to date, are solely reported to have β-xylosidase activity. Additionally, unique to *T. mathranii* subsp. *mathranii* A3 is a putatively cell bound GH43 enzyme, which may have further pentose releasing hydrolysis capabilities.

The identification that only three of the strains evaluated contain GH10 family enzymes is consistent with the reports that only *T. italicus* Ab9 (Kozianowski *et al*., 1997), *T. mathranii* subsp. *mathranii* A3 (Larsen *et al*., 1997) and *T. thermohydrosulfuricus* WC1 (Appendix A.5) are capable of growing on xylan polymers. Other sequenced strains lacking extracellular xylanolytic enzymes, specifically *T. siderophilus* SR4 (Slobodkin *et al*., 1999), *Thermoanaerobacter* sp. X514 (Hemme *et al*., 2011) and *T. pseudethanolicus* 39E (Ng *et al*., 1981), have been specifically reported to not grow on xylan. The ability to use, or not use, xylan by the remaining *Thermoanaerobacter* strains has not yet been reported.

T. italicus Ab9 has the sole genome containing a pectate lyase (Kozianowski *et al*., 1997), while the genomes of *T. thermohydrosulfuricus* WC1 and *T. wiegelii* Rt8.B1 are the only two that encode a putative extracellular acetyl-xylan esterase for removal of acetyl-groups from xylan polymers. A final unique extracellular enzyme found in *T. thermohydrosulfuricus* WC1 and *T. siderophilus* SR4 (TthWC1_0529; ThesiDRAFT1_0902) has a modular structure of GH66-CBM35-CBM35-GH15. No database (GenBank – non redundant protein) entry yet is homologous over the entire length of these CDS, as <60% query coverage is observed when comparing all database entries with either the TthWC1_0529 or ThesiDRAFT1_0902 queries. Reported activities associated with GH66 and GH15 family proteins are reported to act on α-glucan linkages, while CBM35 modules reportedly bind β-linkages (xylans, mannans, galactans), but the functional role of these enzymes is unknown.

4.4.3.2 Intracellular, characterized lignocellulose hydrolyzing CAZymes

Only a few glycoside hydrolases with lignocellulose hydrolysis potential have been functionally characterized in any *Thermoanaerobacter* strain and these are all predicted to be localized intracellularly. One of these, a GH3 family protein *xarB* from *T. ethanolicus* JW200, was initially cloned and characterized and was reported to cleave β-1,4-xylobiose linkages as well as remove arabinosyl groups from xylo-oligomers (Mai *et al*., 2000). Orthologous sequences, showing conserved genomic organization to *T. ethanolicus* JW200, are found in *T. italicus* Ab9 (Thit_0198), *T. mathranii* subsp. *mathranii* A3 (Tmath_0257) and *T. thermohydrosulfuricus* WC1 (TthWC1_1005) (also see Fig. 6 in Hemme *et al*., 2011). Alternatively, a near identical, though N-terminally

truncated *xarB* ortholog, *xglS*, was characterized in *T. brockii* subsp. *brockii* HTD4 (Breves *et al*., 1997) and also shown to have β-xylosidase activity. Similarly to *T. brockii* subsp. *brockii* HTD4, the *xglS* orthologs in *T. brockii* subsp. *finnii* Ako-1 (Thebr_2099) and *T. pseudethanolicus* 39E (Teth39_2055) sequences are co-localized with a *cglT* (GH1) ortholog (Thebr_2098; Teth39_2054) immediately downstream. The *xglS* ortholog in *T. wiegelii* Rt8.B1 (Thewi_00009950) is separated from its *cglT* ortholog (Thewi_00009970) by a hypothetical protein, while the GH3 enzymes of *Thermoanaerobacter* spp. X513, X514 and X561, as well as *T. siderophilus* SR4, do not show conserved genomic organization to the experimentally characterized *xarB* or *xglS* sequences

The present analyses indicate that none of the GH3 CDS are predicted to be secreted extracellularly is in agreement with the findings for *T. thermohydrosulfuricus* JW 102 (Mai *et al*., 2000). The predicted cytosolic location of this enzyme suggests that xylo-oligosaccharides, and not just xylose, are capable of being transported into the cell. The presence of both extracellular GH52 and cytosolic GH3 β-xylosidase orthologs in the *T. italicus* Ab9, *T. mathranii* subsp. *mathranii* A3 and *T. thermohydrosulfuricus* WC1 genomes is an interesting redundancy regarding xylan/xylo-dextrin hydrolysis patterns within these organisms, though the potential impact it may have on xylan hydrolysis and utilization in raw substrates is unknown.

The study by Breves *et al*. (1997) additionally characterized a GH1 family protein, designated *cglT*, which is also predicted to be localized intracellularly. In *T. brockii* subsp. *brockii* HTD4, *cglT* is capable of hydrolyzing the β-1,4-glucosidic linkages of cellodextrins (up to cellopentaose tested) as well as the β-1,3- and β-1,2-

glucosidic linkages of laminaribose and sophorose respectively, and is co-localized with *xglS* (see above). Only three sequenced strains, *T. brockii* subsp. *finnii* Ako-1, *T. pseudethanolicus* 39E and *T. wiegelii* Rt8.B1 share similar genomic organization. GH1 domain containing sequences, highly similar to the *cglT* gene in *T. brockii* subsp. *brockii* HTD4, can be identified in all other *Thermoanaerobacter* strains, but are not co-localized with *xglS* and thus, may simply represent *cglT* homologs, rather than orthologs.

The release of fermentable sugars from lignocellulosic biomass is often considered the rate limiting step in simultaneous saccharification and fermentation processes (Piccolo & Bezzo, 2009; Xu *et al*., 2010; Das *et al*., 2012). Enzymatic pre-treatment of biomass is a common strategy for improving lignocellulose saccharification (Sun & Chen, 2002), though the accumulation of soluble sugars generated through hydrolysis are well known to have an inhibitory effect on continued enzymatic activity (Ramos *et al*., 1993; Xiao *et al*., 2004). Simultaneous incubation of enzyme cocktails and sugar fermenting bacteria is one strategy to relieve this inhibition and continue driving hydrolysis. However, Podkaminer *et al*. (2012) have recently shown that in some cases, the activity of free enzymes is reduced when incubated under conditions amenable for bacterial growth. Thus, using the hydrolytic machinery native to ethanol producing strains is favorable in terms of maintaining enzymatic activity and limiting costs associated with exogenous enzyme addition. The present analyses of the *Thermoanaerobacter* spp. genomes has identified that only *T. italicus* Ab9, *T. mathranii* subsp. *mathranii* A3 and *T. thermohydrosulfuricus* WC1 possess the potential enzymatic machinery needed to hydrolyze complex xylan polymers and potentially facilitate lignocellulose hydrolysis under CBP conditions.

4.4.4 Carbohydrate transport

The extracellular hydrolysis of insoluble carbohydrate polymers requires that the resulting soluble saccharides be transported into the cell prior to fermentation (Fig. 4.1). Based upon genome annotations, *Thermoanaerobacter* spp. import carbohydrates via ABC-type transporters, phosphotransferase system (PTS) transporters and via cationic symporters (Table 4.2). Of the three annotated systems, ABC-type transporters, are the most abundant as *Thermoanaerobacter* spp. contain anywhere from 123-150 total genes belonging to the ATP-binding Cassette (ABC) Superfamily (TC:3.A.1) based on TCDB designations. Only a subset of these will be involved with carbohydrate import, and to date, only a few of this subset have been characterized within any strain of the genus.

4.4.4.1 ABC-type transporters.

Despite the presence of annotated poly- and oligosaccharide transporters within *Thermoanaerobacter* genomes (Table 4.2), few studies have investigated the ability of *Thermoanaerobacter* strains to import these saccharides. Working with *Thermoanaerobacter* strains not analyzed here, the study by Wiegel *et al*. (1985) noted strain specific differences in xylo-oligomer utilization patterns and rates, yet to date, no correlation between genome content and poly- or oligosaccharide utilization has been established for any strain within the genus. The prediction that most *Thermoanaerobacter* CAZymes are intracellular also suggests that these strains have the innate capacity to transport complex soluble carbohydrates as there is little evolutionary advantage to maintain intracellular CAZymes if the saccharides upon which they act cannot be transported into the cell. However, experimental characterization of

Table 4.2. Selected[a] transporters associated with carbohydrate import identified within sequenced *Thermoanaerobacter* spp.

	Clade 1						Clade 2		Clade 3		
	T. brockii subsp. *finnii* Ako-1	*T. ethanolicus* CCSD1	*T. pseudethanolicus* 39E	*Thermoanaerobacter* sp. X513	*Thermoanaerobacter* sp. X514	*Thermoanaerobacter* sp. X561	*T. italicus* Ab9	*T. mathranii* subsp. *mathranii* A3	*T. siderophilus* SR4	*T. thermohydrosulfuricus* WC1	*T. wiegelii* Rt8.B1
Cationic symporters											
2-Keto-3-Deoxygluconate	+	+	+	-	-	-	-	-	-	-	-
Glycoside/Pentoside/ Hexuronide Cation Symporter	+	+	+	+	+	+	+	+	+	+	+
ATP binding family											
Lactose/L-arabinose	-	-	-	-	-	-	+	+	+	+	+
Maltose/Maltodextrin	+	+	+	+	+	+	+	+	+	+	+
Methyl-Galactoside	-	+	-	-	-	-	-	-	-	-	-
Multiple Sugar (unspecified)	+	+	+	+	+	+	+	+	+	+	+
Oligogalacturonide	-	-	-	-	-	-	+	-	+	-	-
Putative Multiple Sugar (unspecified)	-	-	-	+	+	+	-	-	+	+	+
Putative Sugar (unspecified)	-	-	-	-	-	-	+	-	-	-	-
Ribose	+	+	+	+	+	+	+	+	-	+	+
Simple Sugar (unspecified)	+	+	+	+	+	+	+	+	+	+	+
Xylose	-	+	-	+	+	+	+	+	+	+	+
PTS family											
Cellobiose	+	+	+	+	+	+	+	+	+	+	+
Fructose	+	+	+	+	+	+	+	-	+	+	+
Galactitol	+	+	+	+	+	+	+	+	+	+	+
Galactosamine	-	-	-	-	-	-	-	-	+	+	+
Glucitol/Sorbitol	-	-	-	-	-	-	-	-	+	+	+
Mannitol	+	+	+	+	+	+	+	+	+	+	+
Mannose	+	+	+	+	+	+	+	+	+	+	+
N-Acetylglucosamine	+	+	+	+	+	+	+	+	+	+	+
Sucrose	-	-	-	+	+	+	-	-	+	+	-

Symbols denote the presence (+) or absence (-) of a particular annotated transporter. Substrate specificity is inferred based upon KO annotated specificity of the substrate binding protein (ABC transporters) or the membrane linked EIIC components (PTS transporters).

[a]Transport systems presented are limited to complexes showing co-localization of all genes needed to form a functional complex. Complexes lacking annotation of a single component are not included. Redundancy in transport systems exists, but is not identified.

Thermoanaerobacter transport systems is needed. This is particularly true given that the mean degree of polymerization for cellulose hydrolysis products generated by *C. thermocellum* is four (Zhang & Lynd, 2005), and that *C. thermocellum* xylan hydrolysis yields principally xylo-oligomers (Morag *et al*., 1990; Hayashi *et al*., 1999; Zverlov *et al*., 2005; Izquierdo *et al*., 2012). The ability, or lack thereof, to transport, poly- and oligosaccharides in a *Thermoanaerobacter* strain may significantly impact carbohydrate utilization in a co-culture with *C. thermocellum*.

Xylose transport has been one of the most thoroughly investigated sugar import mechanisms within the genus. An ABC-type xylose binding protein was first identified in *T. pseudethanolicus* 39E by Erbeznik *et al*. (1998b) and a *xylFGH* operon, coding for the entire ABC-type transporter, was later determined in the same strain (Erbeznik *et al*., 2004). However, the amino acid sequences reported do not agree with those identified in the *T. pseudethanolicus* 39E genome. Nevertheless, the reported partial *xylF* sequence and complete *xylGH* sequences reported do show 99.5%, 100% and 100% sequence identity, respectively, and similar genomic architecture, with the annotated xylose ABC-transport system in *T. italicus* Ab9.

In *T. italicus* Ab9, as well as *T. mathranii* subsp. *mathranii* A3, *T. thermohydrosulfuricus* WC1 and *T. wiegelii* Rt8.B1, the annotated xylose transport genes are not co-localized with the xylose isomerase and xylulokinase (*xylAB*) operon. They are however, co-localized in *Thermoanaerobacter* spp. X513, X514 and X561, *T. ethanolicus* CCSD1 and *T. siderophilus* SR4 as is reported elsewhere (Hemme *et al*., 2011). *Thermoanaerobacter* spp. X513, X514, and X561, as well as *T. siderophilus* SR4, all contain a secondary gene cluster with an annotated xylose binding protein located

further downstream. Recently, Lin *et al*. (2011) report that in *Thermoanaerobacter* sp. X514, the xylose transport genes found co-localized with the *xylAB* operon are induced when grown on xylose giving support to the annotation of these genes as xylose transporters. The response of the secondary gene cluster was not discussed and its role in xylose transport is not yet confirmed. The presence of xylose specific ABC-transport systems in some, but not all, *Thermoanaerobacter* strains (Table 4.2) has been proposed to potentially account for increased xylose uptake and utilization by the strains encoding the transporter (Hemme *et al*., 2011). Assuming this translates throughout the genus, only two strains, *T. brockii* subsp. *finnii* Ako-1 and *T. pseudethanolicus* 39E, may be limited in xylose uptake efficiency based on their annotated genomes, as they may rely on non-specific mechanisms for xylose uptake.

The transport of cellobiose and/or glucose has also been proposed to occur via an ABC-type transport system designated *cglF-cglG*(*xarG*) with the substrate specificity inferred based upon its proximal location to a glycoside hydrolase (*cglT* – discussed in Chapter 4.4.3.2) showing high activity towards cellobiose in *Thermoanaerobacter brockii* subsp. *brockii* HTD4 (Breves *et al*., 1997). However, Mai *et al*. (2000) note that in *T. ethanolicus* JW200 the orthologous transporter encoding genes are co-localized with a bifunctional xylosidase-arabinosidase (*xarB*/*xglS*) and not with a *cglT* ortholog. The gene products may therefore be involved with transporting xylan hydrolysis products. Only six *Thermoanaerobacter* genomes, *T. brockii* subsp. *finnii* Ako-1, *T. pseudethanolicus* 39E, *T. italicus* Ab9, *T. mathranii* subsp. *mathranii* A3, *T. thermohydrosulfuricus* WC1 and *T. wiegelii* Rt8.B1, contain the *cglF-cglG*/*xarG* gene cluster and in all cases, show similar genomic organization to *T. ethanolicus* JW200.

4.4.4.2 PTS-mediated transport.

Genes annotated to transport nine different substrates via PTS-mediated mechanisms are identified within the genus (Table 4.2), though none of these gene sequences have been functionally characterized. Alternative to the ABC-dependent glucose uptake proposed (Chapter 4.4.4.1), Lin *et al*. (2011) have identified genes in *Thermoanaerobacter* sp. X514 which may form a glucose specific PTS complex. Expression analysis of the gene cluster identified (Teth514_0412-Teth514_0414) was specifically linked to mid-exponential phase growth of cells grown on glucose only. However, the KO annotation (KO:K02803/KO:K02804) suggests that this gene cluster is specific for N-acetyl-glucosamine (GlcNAc) rather than glucose. The KO annotation is further supported by the fact that the phosphotransferase specificity of the orthologous EIICB gene in *Caldanaerobacter subterraneus* subsp *tengcongensis* MB4 (87.7% shared amino acid identity with *Thermoanaerobacter* sp. X514) has a V_{max} 4-fold higher for GlcNAc than it does for glucose (Navdaeva *et al*., 2011). Additionally, the studies by Ng and Zeikus (1982) on *T. pseudethanolicus* 39E and Cook *et al*. (1993) on *T. wiegelii* Rt8.B1 have identified that glucose import is via non-PTS-mediated transporters. Thus, the genes responsible for glucose uptake in any *Thermoanaerobacter* strain are not yet confirmed.

4.4.4.3 Cationic symporters.

Only two distinct carbohydrate-relevant types of cationic symporters are annotated within the genus (Table 4.2). The study by Hemme *et al*. (2011) has proposed that Na^+ gradient-linked transport may be a mechanism for xylose uptake in *T.*

pseudethanolicus 39E, which lacks an annotated xylose specific ABC-type transport system. While the substrate binding specificity of the annotated cationic symporters is not yet known, experimental characterization of these enzymes may shed valuable insights into *Thermoanaerobacter* spp. carbon transport, particularly in strains lacking annotated carbon-specific transport systems.

4.4.5 Carbohydrate utilization

While cellulose is comparatively chemically homogenous (β-1,4-glucose linkages), hemicellulose fractions are chemically and structurally diverse. The products of lignocellulose hydrolysis will therefore generate a mixed pool of saccharides available for utilization. While sugar composition ratios and linkages vary from hemicellulose to hemicellulose, the principal simple saccharides of hemicellulose are xylose, arabinose, mannose, glucose, galactose and glucuronic acid (Saha, 2003; Shallom & Shoham, 2003; Moreira *et al.*, 2011). As the conversion of hemicellulose to ethanol plays an important role in making lignocellulosic biofuels economically feasible (Demirbaş, 2005; Galbe *et al*., 2007), identifying a strain capable of fermenting multiple sugars, particularly multiple sugars simultaneously, into ethanol is essential.

4.4.5.1 Utilization of hexose sugars.

The potential for all sequenced strains within the genus to utilize the seven primary constituent sugars of lignocellulose was evaluated (Table 4.3, Table A.3.3). All strains contain a complete Embden-Meyerhoff-Parnas (EMP) pathway. Redundancy in genome annotations for the EMP pathway were only observed for genes encoding

Table 4.3. Identification of the genomic potential for sequenced *Thermoanaerobacter* strains to utilize the major carbohydrate hydrolysis products of lignocellulose degradation.

		Cellulose hydroylsis products[a]				Hemicellulose hydrolysis products										
		Glucose		Cellobiose		Arabinose		Galactose		Mannose		Xylose		Galacturonic acid		
Clade	Strain	GP	PR	GP	PR	GP	PR	GP	PR	GP	PR	GP	PR	GP	PR	Ref[b]
1	*T. brockii* subsp. *finnii* Ako-1	+	+	+	+	-	-	+	+	+	+	+	+	-	NR	1
	T. ethanolicus CCSD1	+	NR	+	NR	-	NR	+	NR	+	NR	+	NR	-	NR	-[c]
	T. pseudethanolicus 39E	+	+	+	+	-	-	+	+	+	+	+	+	-	NR	2
	Thermoanaerobacter sp. X513	+	+	+	NR	-	NR	+	NR	+	NR	+	+	-	NR	3
	Thermoanaerobacter sp. X514	+	+	+	NR	-	NR	+	NR	+	NR	+	+	-	NR	3
	Thermoanaerobacter sp. X561	+	+	+	NR	-	NR	+	NR	+	NR	+	+	-	NR	3
2	*T. italicus* Ab9	+	+	+	+	+	+	+	+	+	+	+	+	+	NR	4
	T. mathranii subsp. *mathranii* A3	+	+	+	+	+	+	+	-	+	+	+	+	+	NR	5
3	*T. siderophilus* SR4	+	+	+	+	-	-	+	NR	+	NR	+	+	-	NR	6
	T. thermohydrosulfuricus WC1	+	+	+	+	-	-	+	+	+	+	+	+	+	NR	7
	T. wiegelii Rt8.B1	+	+	+	+	-	-	+	+	+	+	+	+	-	NR	8

Symbols and abbreviations: Symbols denote the presence (+) or absence (-) of all protein coding genes needed for substrate hydrolysis to pyruvate or whether the substrate has been reported to be used (+) or not used (-) in the literature. GP = genes present; PR = phenotype reported; NR = substrate utilization has not yet been reported for a particular strain.

[a]Cellulose hydrolysis products have been limited to glucose and cellobiose and not higher order cellodextrins.

[b]References are as follows: **1** = Cayik *et al*., 1995; **2** = Onyenwoke *et al*., 2007; **3** = Roh *et al*., 2002; **4** = Kozianowski *et al*., 1997; **5** = Larsen *et al*., 1997; **6 =** Slobodkin *et al*., 1999; **7** = Chapter 2; **8** = Cook *et al*., 1996.

[c]Physiological data has not yet been reported for this strain.

glucokinase, 6-phopshofructokinase, fructose-1,6-bisphosphate aldolase and phosphoglyceromutase (Table A.3.3).

Genes for a complete Entner-Doudoroff pathway could not be identified, in agreement with previous findings (Hemme *et al*., 2011). Additionally, a 6-phosphogluconolactonase encoding gene, potentially allowing for hexose utilization through the oxidative pentose phosphate pathway, was not identified in any of the genomes. This pathway has been shown to be non-functional in *Thermoanaerobacter* sp. X514 (Feng *et al*., 2009). However, genes encoding enzymes common to both the Entner-Duodoroff and pentose phosphate pathways are identified within the genomes. These gene products may catalyze the conversion of specific sugars into saccharides suitable for catalysis via the EMP pathway.

All of the sequenced genomes contain a mannose-6-P isomerase responsible for converting mannose-6-P to fructose-6-P and needed for mannose utilization. Additionally, a conserved 5-gene cluster needed for conversion of galactose to glucose-6-P via the Leloir pathway was identified in all of the genomes. This cluster is orthologous to the novel *gal* operon characterized in *Cal. subterraneus* subsp. *tengcongensis* MB4 (Qian *et al*., 2009).

The entry points into the EMP pathway for the products of cellobiose hydrolysis are somewhat difficult to predict and would largely be dependent on the mode of transport into the cell (see Chapter 4.4.4), which is not yet known in any strain. Transport via an ABC-type system would import cellobiose, while transport via a PTS transporter would yield cellobiose-6-P. In the former situation, cellobiose could be hydrolyzed to two glucose molecules or, if hydrolyzed using a cellobiose phosphorylase

(Table A.3.3), could yield 1 glucose + 1 glucose-1-P using an inorganic phosphate for phosphorylation. If cellobiose import is via a PTS transporter, hydrolysis would yield 1 glucose + 1 glucose-6-P and could occur via any of the GH1, GH3, GH4 or GH5 enzymes identified in all *Thermoanaerobacter* strains (Table 4.2, Table A.3.3). To date, only the *cglT* gene characterized in *T. brockii* subsp. *brockii* HTD4, which has homologs in all other *Thermoanaerobacter* genomes, has been shown to have β-glucosidase (including cellobiose) activity (Breves *et al*., 1997). Transformation from glucose-1-P to glucose-6-P could occur via a phosphoglucomutase (COG0637; KO:K01838) found in all genomes.

Genes for glucuronic acid metabolism are not universally conserved throughout the genus. Only the Clade 2 strains, plus *T. thermohydrosulfuricus* WC1, have the necessary genes for conversion of glucuronic acid to glyceraldehyde-3-P + pyruvate in conserved 5-gene clusters. Multiple genes are annotated to encode for a 2-keto-3-deoxyphosphogluconate (KDPG) aldolase (COG0800; KO:K01625; TIGR01182) in all *Thermoanaerobacter* strains, though. This enzyme, which is common to the Entner-Doudoroff pathway, also serves as the entry point for glucuronic acid utilization into the EMP pathway. However, with the exception of the strains mentioned above, the necessary genes for glucuronic acid utilization are not identified and thus the role of these annotated KDPG aldolases is difficult to infer.

4.4.5.2 Utilization of pentose sugars.

The present analyses confirms that, in all strains, xylose is likely isomerized and phosphorylated via xylose-isomerase (*xylA*) and xylulose-kinase (*xylB*) reactions prior to

entering the pentose phosphate pathway, in agreement with previous findings (Hemme *et al*., 2011). Arabinose utilization genes seem to be limited to Clade 2 strains. Both *T. italicus* Ab9 and *T. mathranii* subsp. *mathranii* A3 have a conserved 3-gene cluster (Table A.3.3) annotated as L-arabinose-isomerase, L-ribulokinase and L-ribulose-5-P-4-epimerase needed to convert L-arabinose to D-xylulose-5-P prior to entering the pentose phosphate pathway.

Of the 77 *in silico* predictions for carbohydrate utilization (11 genomes x 7 carbohydrates), 45 agree with phenotypes reported in the literature (Cayol *et al*., 1995; Cook *et al*., 1996; Kozianowski *et al*., 1997; Larsen *et al*., 1997; Slobodkin *et al*., 1999; Roh *et al*., 2002; Onyenwoke *et al*., 2007; Chapter 2) (Table 4.3). Thirty-one of the strain-substrate combinations have not yet been investigated experimentally to either confirm or refute these predictions and one prediction (galactose utilization by *T. mathranii* subsp. *mathranii* A3) disagrees with reported phenotypes (Larsen *et al*., 1993). *T. mathranii* subsp. *mathranii* A3 has orthologs to the functionally characterized *gal* operon (Chapter 4.4.5.1) in *Cal. subterraneus* subsp. *tengcongensis* MB4. The arrangement of the genes is identical to *Cal. subterraneus* subsp. *tengcongensis* MB4 and the annotated genes share >89% amino acid sequence similarity to each respective ortholog. Therefore, the reason galactose utilization was not observed by Larsen and coworkers (1993) may be due to regulatory differences, a few select mutations affecting enzyme functionality or even an inability to transport galactose, but the exact reason is not clear at this time.

Only Clade 2 strains have the potential to utilize all of the major lignocellulose hydrolysis products (Table 4.3). However, the significance of this in CBP terms may

vary dependent on the nature of the lignocellulosic feedstock. As the composition of hemicellulose varies between feedstocks, the inability to utilize substrates in low abundance may represent acceptable losses for any single CBP system. Alternatively, using a strain with diverse substrate utilization capabilities affords flexibility in designing a CBP system, independent of the nature of the biomass feedstock, not present in strains lacking the ability to utilize specific hemicellulose-relevant saccharides.

4.4.6 Pyruvate catabolism and end-product synthesis

4.4.6.1 Pyruvate catabolism.

Fermentation of the above mentioned carbohydrates leads to the formation of pyruvate (Fig. 4.1), a key branch point in *Thermoanaerobacter* carbohydrate metabolism. All *Thermoanaerobacter* strains are reported to have branched catabolic pathways from pyruvate, which yield both ethanol and non-ethanol end-products in varying ratios (Table 4.4). Understanding pyruvate catabolism, and identifying mechanisms to maximize carbon flow towards ethanol, is therefore an important component in making *Thermoanaerobacter* strains industrially relevant.

Pyruvate decarboxylation in all *Thermoanaerobacter* strains, forming acetyl-CoA + CO_2 + reducing equivalents, appears to proceed through a pyruvate: ferredoxin oxidoreductase (POR) enzyme. This is supported by the identification of genes (Table A.3.4) homologous to the single subunit characterized POR in the phylogenetically related *Moorella thermoacetica* (Furdui & Ragsdale, 2002; Pierce *et al*., 2008). Further, multiple investigations of Clade 1 strains have reported a significant role for Fd, as well

Table 4.4. Reported end-product yields and related growth conditions for sequenced *Thermoanaerobacter* strains[a] grown on glucose, xylose or cellobiose.

		End products (mol/mol hexose equivalent)					Culture Conditions			
Clade	Strain	H_2	CO_2	Ac	Et	La	Substrate	[g/L]	Type	Reference
1	*T. brockii* subsp. *finnii* Ako-1	0.3	1.7	0.3	1.7	0.4	Glucose	5	B	(Schmid *et al*., 1986)
	T. pseudethanolicus 39E	NI	NQ	0.4	1.1	0.3	Glucose	NR	B	(Hemme *et al*., 2011)
		NI	NQ	0.3	1.5	0.1	Xylose	2	B	(Hemme *et al*., 2011)
		0.5	1.7	0.1	1.4	0.2	Glucose	10	B	(Ben-Bassat *et al*., 1981)
		NI	1.6	0.1	1.6	0.2	Glucose	4	B	(Lovitt *et al*., 1982)
		NI	NI	0.2	1.3	0.3	Xylose	5	B	(He *et al*., 2009)
		NI	NI	0.2	1.1	0.4	Glucose	5	B	(He *et al*., 2009)
	Thermoanaerobacter sp. X513	NI	NI	NI	NQ	NI	Xylose	1.5	B	(Roh *et al*., 2002)
	Thermoanaerobacter sp. X514	NI	NQ	0.5	1.0	0.3	Glucose	NR	B	(Hemme *et al*., 2011)
		NI	NQ	0.3	1.6	0.1	Xylose	2	B	(Hemme *et al*., 2011)
		NI	NQ	0.3	1.3	0.2	Glucose	2.1	B	(Feng *et al*., 2009)[b]
	Thermoanaerobacter sp. X561	NI	NI	NI	NQ	NI	Xylose	1.5	B	(Roh *et al*., 2002)
2	*T. italicus* Ab9	NQ	NC	NC	NC	NC	Glucose	5	B	(Kozianowski *et al*., 1997)[c]
	T. mathranii subsp. *mathranii* A3	1.1	2.2	0.5	1.3	0.1	Xylose	2	B	(Larsen *et al*., 1997)
3	*T. siderophilus* SR4	NQ	NQ	NI	NQ	NQ	Glucose	5	B	(Slobodkin *et al*., 1999)
	T. thermohydrosulfuricus WC1	0.6	1.2	0.7	0.6	0.3	Cellobiose	2	B	Chapter 2
	T. wiegelii Rt8.B1	NI	NI	0.5	0.9	0.2	Glucose	8.6	B	(Cook, 2000)[b]
		NI	1.6	0.5	1.1	0	Glucose	9	P	(Cook *et al*., 1994a)[d]
		NI	1.8	0.5	1.3	0.1	Xylose	7.5	P	(Cook *et al*., 1994a)[d]
		NI	1.7	0.6	1.1	ND	Cellobiose	17	P	(Cook *et al*., 1994a)[d]

Abbreviations: NI = end-product not investigated; ND = end-product investigated, but not detected; NQ = end-product detected, but not quantified; NC = end-product quantified, but ratio not calculable due to missing substrate consumption data; B = batch culture; P = pH-controlled batch culture.

[a]*T. ethanolicus* CCSD1 omitted as no physiological data is available.

[b]End-product and substrate utilization data used for calculations approximated from graphical data.

[c]Succinate production reported as a minor product.

[d]Propionate detected on xylose grown cells (molar ratio = 0.12) and cellobiose grown cells (molar ratio = 0.06).

as ferredoxin: NAD(P)H reductase activity (Lamed & Zeikus, 1980a; Ben-Bassat *et al*., 1981; Lovitt *et al*., 1988). Four-gene clusters, annotated as the alpha, beta, gamma and delta subunits of a multi-subunit POR complex are also identified within all strains (Table A.3.4). It is difficult to predict on an *in silico* basis which gene or gene clusters encode the primary POR responsible for pyruvate catabolism and which encode gene products that may act on alternative keto-acids such as indolepyruvate, 2-ketoisovalerate or 2-ketoglutarate.

4.4.6.2 Lactate synthesis.

The production of lactate has been reported for all sequenced *Thermoanaerobacter* strains with physiological data available (Table 4.4) and occurs via the reduction of pyruvate using a lactate dehydrogenase (LDH) enzyme. Strains in all 3 clades have a single gene annotated as a *ldh* (KO:K00016; TIGR01771), though by COG annotation (COG0039), these same genes are designated as malate/lactate dehydrogenases. Distinguishing between *ldh* and malate dehydrogenase (*mdh*) genes *in silico* can be difficult, though the CDS identified in all genomes (Table A.3.4), with the exception of *Thermoanaerobacter* sp. X561 (sequence truncated due to contig break), share >86% amino acid sequence similarity with the characterized *ldh* from *Thermoanaerobacterium saccharolyticum* (Shaw *et al*., 2011a). As no other obvious *ldh* is identified, and lactate production is reported throughout the genus, the genes identified in Table A.3.4 are proposed to catalyze *Thermoanaerobacter* lactate formation.

4.4.6.3 Acetate synthesis.

Strains of *Thermoanaerobacter* are also reported to produce acetate (Table 4.4). POR mediated pyruvate catabolism will yield acetyl-CoA, which can be converted to acetate + 1 ATP via phosphotransacetylase (*pta*) and acetate kinase (*ack*) gene products. In all strains, the PTA and ACK enzymes are co-localized within the genome (Table A.3.4). Three strains, *Thermoanaerobacter* spp. X513, X514 and X561 have additional *ack* genes annotated, though these are not co-localized with *pta* genes. Working with a non-sequenced *Thermoanaerobacter* strain, *Thermoanaerobacter thermohydrosulfuricus* DSM570, Mayer *et al*. (1995) observed a severe reduction in acetate production and enzyme activity in PTA^- and ACK^- mutants and additionally proposed that the *pta* and *ack* genes were co-localized and formed an operon. The residual acetate production observed may be in part due to the fact that in all sequenced strains, additional gene sequences annotated as phosphate butyryltransferases and butyrate kinase genes are also identified (Table A.3.4), and the substrate specificity of these genes is not yet known.

4.4.6.4 Ethanol synthesis.

Ethanol production occurs via the reduction of acetyl-CoA to acetaldehyde via an acetaldehyde dehydrogenase followed by a second reduction to ethanol via an alcohol dehydrogenase. Three functionally characterized alcohol dehydrogenase (ADH) genes, *adhA, adhB* and *adhE,* have been reported to be principally involved with ethanol formation and a model describing the physiological roles of each gene has been proposed (Pei *et al*., 2010b). The *adhA* gene product from *T. ethanolicus* JW200 is a reported Zn-binding NADPH-dependent primary alcohol dehydrogenase (Bryant *et al*., 1988; Holt *et*

al., 2000). In comparison, the *adhA* gene product from *T. pseudethanolicus* 39E, which is capable of utilizing both NADH and NADPH, showed a higher catalytic efficiency for NADH oxidation over NADPH oxidation (Burdette & Zeikus, 1994). Pei *et al*. (2010b) demonstrated, *in vitro*, that the *adhB* and *adhE* gene products from *T. ethanolicus* JW200 displayed bifunctional acetaldehyde: alcohol dehydrogenase activity despite the fact that only the *adhE* gene contained two independent domains related to aldehyde dehydrogenase and alcohol dehydrogenase families, respectively. However, when assayed using measured intracellular concentrations of $NAD(P)^+$ and NAD(P)H, the *adhE* gene product displayed only aldehyde dehydrogenase activity (NADH dependent), while the *adhB* gene strongly favored acetaldehyde reduction over acetyl-CoA reduction.

Only five *Thermoanaerobacter* genomes contain genes annotated as standalone aldehyde dehydrogenases (Table A.3.4), but no evidence yet exists to suggest that these genes function as acetaldehyde dehydrogenases. Furthermore, genomic context provides no further insights into substrate specificity. As such, it is likely that the reduction of acetyl-CoA to acetaldehyde via *adhE* is a conserved physiological process throughout the genus. The *adhB* encoding gene in *T. pseudethanolicus* 39E is considered to be NADPH-dependent (Burdette *et al*., 1996) and, upon ethanol accumulation, has shown a higher specific activity towards ethanol oxidation as opposed to ethanol formation (Pei *et al*., 2010b). Its role *in vivo* is not yet confirmed.

The three ADH encoding genes, *adhA*, *adhB* and *adhE*, in conjunction with a recently described redox-sensing transcriptional regulator in *T. ethanolicus* JW200 (Pei *et al*., 2011), have largely formed the basis for our understanding of ethanologenesis in a few select *Thermoanaerobacter* strains. However, given that the sequenced

Thermoanaerobacter genomes have annotated anywhere from 5-9 putative alcohol dehydrogenases, most of which have unknown specificity, this model of ethanol metabolism may not fully encompass all ethanol producing reactions within the cell.

The three characterized *Thermoanaerobacter* alcohol dehydrogenase genes belong to COG1454 - Class IV alcohol dehydrogenase (*adhA*, *adhE*) and COG1063 - Threonine dehydrogenase and related Zn-dependent dehydrogenases (*adhB*). Additional sequences belonging to each COG designation were identified, as well as sequences belonging to COG1979 - Uncharacterized oxidoreductase, Fe-dependent alcohol dehydrogenase family (Table A.3.4). To identify whether the genomes encode potential additional ethanol producing alcohol dehydrogenases, we conducted phylogenetic analysis of all gene sequences identified in COG1063, COG1454 and COG1979 (Fig. 4.3) as a means of inferring specificity.

Of the 80 gene sequences analyzed, 8 distinct clades, and an additional 4 sub-clades were identified. Gene sequences homologous to *adhA* were identified in Clusters 4A and 4B. Additionally, 5 genomes contained paralogous pairs within these clusters that showed >90.3% amino acid sequence identity within each respective pair. Microarray data from Hemme *et al*. (2011) have shown that in *Thermoanaerobacter* sp. X514, both genes in the paralagous pair (Teth514_0564 – Cluster 4A; Teth514_0654 – Cluster 4B) are expressed. Additionally, in one case (Teth514_0564), expression is dependent on growth conditions. All genomes contained sequences orthologous to the characterized *adhB* gene (Cluster 5) as well as the *adhE* gene (Cluster 1A). Three strains, *Thermoanaerobacter* spp. X513, X514 and X561 contained an additional *adh* gene that grouped near the *adhE* orthologs (Cluster 1B). The annotation of these sequences though

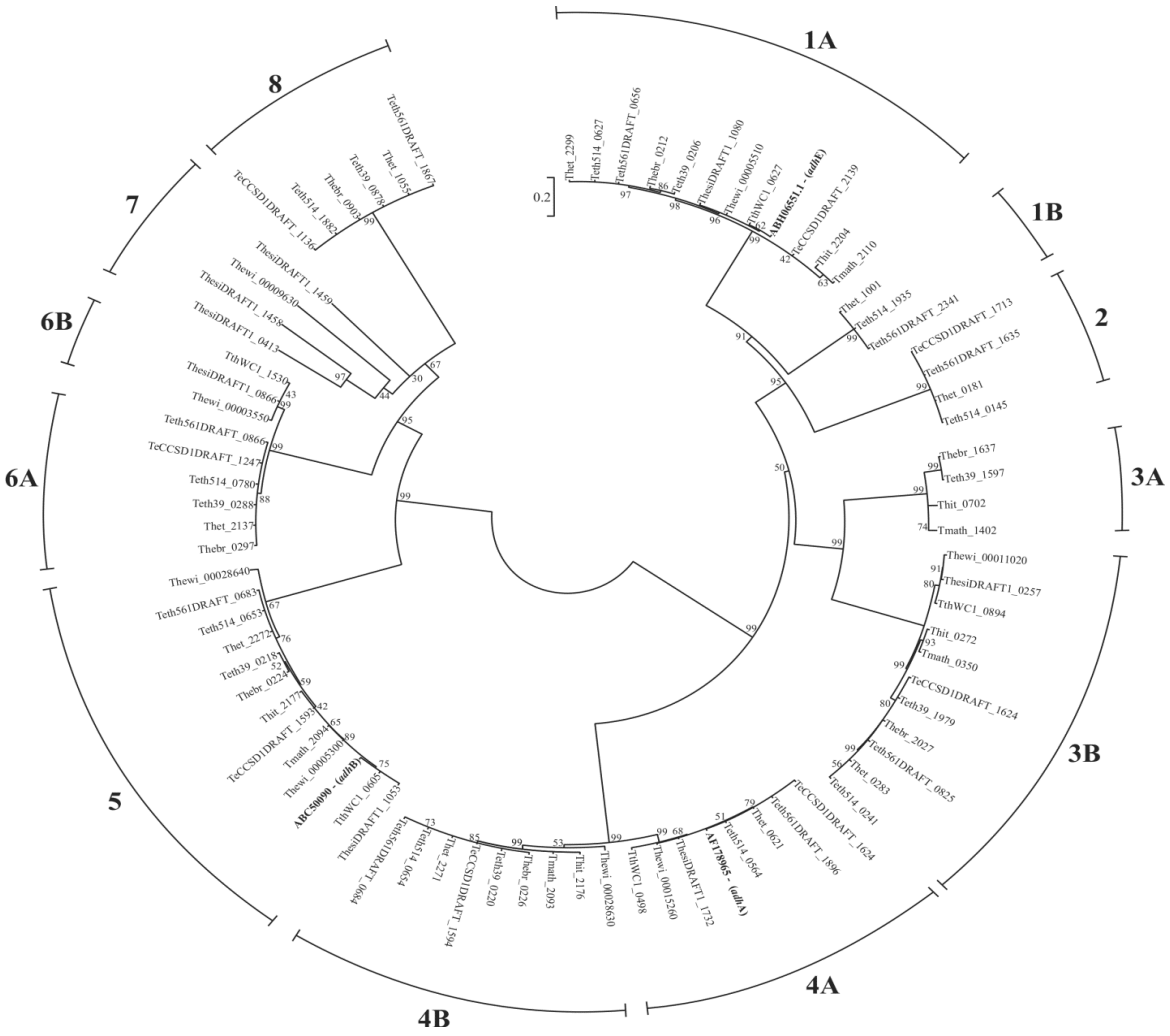


Fig. 4.3[12]. Phylogenetic analysis of all annotated alcohol dehydrogenase genes within sequenced *Thermoanaerobacter* strains. All included sequences belong to COG1063, COG1454 or COG1979. Tmath_0755 was excluded from analysis as the annotated sequence appears to be a CDS fragment. Sequences in bold correspond to the GenBank accession numbers for functionally characterized sequences from *T. ethanolicus* JW200 (Pei *et al*., 2010b, Holt *et al*., 2000). Tree construction was as described (Chapter 4.3.4). Bootstrapping support values are indicated by their respective nodes.

[12] A higher resolution version of this Figure is accessible at: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0059362

suggests only alcohol dehydrogenase, and not aldehyde dehydrogenase, activity. Surprisingly, Clusters 3A and 3B, which belong to COG1979, group more closely to Clusters 1 and 2 (COG1454) than does Cluster 4 (also COG1454).

Given the number and diversity of *adh* genes annotated, and that expression patterns of homologous genes in different strains are distinct (Hemme *et al*., 2011), suggesting differential regulation, the current 3-gene model (Pei *et al*., 2010b) proposed for ethanol formation in *Thermoanaerobacter* spp. may not translate across all strains. This is supported by the fact that the study by Hemme *et al*. (2011) identified varied expression of 8 of the 9 *adh* annotated sequences in *Thermoanaerobacter* sp. X514 under different growth conditions. However, given that multiple *Thermoanaerobacter* spp. have been reported to grow on sugar alcohols (Zeikus *et al*., 1980; Cook *et al*. 1996; Kozianowski *et al*., 1997, Larsen *et al*., 1997; Xu *et al*., 2010), it is possible that some of these may have catabolic functions and are not involved with ethanol synthesis. Understanding ethanologenesis in *Thermoanaerobacter* spp. will ultimately require more in depth functional and expression analysis studies across multiple strains.

4.4.6.5 Hydrogen synthesis.

Hydrogen production, similar to lactate production, competes for reducing equivalents with ethanol synthesis to a greater or lesser extent in all strains of *Thermoanaerobacter* (Table 4.4), but in-depth cross-species analysis of *Thermoanaerobacter* hydrogenases has yet to be conducted.

Table 4.5. Annotated hydrogenase encoding genes[a] within sequenced *Thermoanaerobacter* spp.

		FeFe hydrogenases			Ni-Fe hydrogenase
Clade		Bifurcating[b]	Fd-linked	PAS-sensory	Ech
1	*T. brockii* subsp. *finnii* Ako-1	1491-1495	1498	0227	
	T. ethanolicus CCSD1	0391- 0395	0398	1595	
	T. pseudethanolicus 39E	1456- 1460	1463	0221	
	Thermoanaerobacter sp. X513	0793- 0797	0790	2270	
	Thermoanaerobacter sp. X514	2138- 2142	2145	0655	
	Thermoanaerobacter sp. X561	0433- 0437	0440	0685	
2	*T. italicus* Ab9	0826- 0830	0823	2175	1612- 1617
	T. mathranii subsp. *mathranii* A3	0865- 0869	0862	2092	1603- 1608
			1048		
3	*T. siderophilus* SR4	1833- 1837	1830		1402- 1407
	T. thermohydrosulfuricus WC1	1780- 1784	1787		1092- 1097
	T. wiegelii Rt8.B1	00017870-00017920	00017850	00028620	00009040-00009090
			00028740	00005270	

[a]Only numerical suffixes of locus tags are provided. Corresponding prefixes for each strain are identified in Table A.3.1.

[b]Bifurcating activity proposed based on orthology to identified sequences (Schut & Adams, 2009). See Chapter 4.4.6.5.1 for details.

4.4.6.5.1 (Fe-Fe) hydrogenases.

Gene sequences homologous to the cytosolic NADH-dependent Fe-only enzyme characterized in *Cal. subterraneus* subsp. *tengcongensis* MB4 (Soboh *et al*., 2004) are conserved in all strains of the *Thermoanaerobacter* genus (Table 4.5). Sequence analysis of the hydrogenase conserved domains suggests that these hydrogenases are most similar to the heterotrimeric A1 group exhibiting a TR(M3) modular structure described by Calusinska *et al*. (2010). These gene products are thought to be NAD-dependent due to the presence of NADH-binding domains in the accessory subunits. However, in *Cal. subterraneus* subsp. *tengcongensis* MB4, these same genes, as well as the orthologs in *T. pseudethanolicus* 39E have recently been proposed to function as potential bifurcating (Fe-Fe) hydrogenases (Schut & Adams, 2009), which couple the thermodynamically unfavourable oxidation of NADH to H_2 production through utilization of the exergonic oxidation of Fd_{red}.

Upstream of the genes in the A1 grouping in all *Thermoanaerobacter* strains is a histidine kinase protein as well as another putative hydrogenase gene, which shows similar domain architecture to group D hydrogenases. This is consistent with the genomic organization described for bifurcating hydrogenases (Calusinska *et al*., 2010), which is also observed in *Cal. subterraneus* subsp. *tengcongensis* MB4 (Soboh *et al*., 2004).

PAS-domain (pfam00989) containing sensory hydrogenases are also conserved throughout the genus with the exception of *T. siderophilus* SR4 and *T. thermohydrosulfuricus* WC1 (Table 4.5). Sensory hydrogenases have been reported to be linked with histidine kinase based signal transduction mechanisms and could play a role

in regulating cellular redox levels (Taylor *et al.*, 1999; Calusinska *et al*., 2010). *T. wiegelii* Rt8.B1 and *T. mathranii* subsp. *mathranii* A3 both contain additional (Fe-Fe)-hydrogenases showing modular structures similar to the B1 or B3 monomeric hydrogenases described by Calusinska *et al*. (2010).

4.4.6.5.2 (Ni-Fe) hydrogenases.

(Ni-Fe) hydrogenase encoding genes can be identified for strains in Clade 2 and Clade 3, but not in Clade 1 (Table 4.5). Clade 2 and 3 strains both have the conserved 6-gene cluster coding for a membrane-bound, cation transporting Fd-consuming energy conserving hydrogenase (Ech) directly followed by the 6-gene *hypABFCDE* gene cluster responsible for assembly of the (Ni-Fe) center. Fd_{red} generated via POR in pyruvate catabolism is thought to provide the electrons needed for H_2 evolution using the Ech.

Fd-dependence has been shown for the orthologous gene sequences in *Cal. subterraneus* subsp *tengcongensis* MB4 (Soboh *et al*., 2004). Furthermore, cell extracts of *Cal. subterraneus* subsp. *tengcongensis* MB4 were shown to favour H_2-evolution over H_2-consumption and Ech was additionally proposed to play a role in "proton respiration" (Soboh *et al*., 2004). However, the physiological role of Ech has not been determined for any *Clostridia* and the exact nature of the exported cation (H^+ of Na^+) has not yet been determined. It is interesting to note that the absence of Ech encoding homologs (Table 4.5) in *T. pseudethanolicus* 39E (Clade 1) correlates with the lowest reported molar H_2 yield (and highest molar ethanol yield) of strains with data available (Table 4.4). This is perhaps indicative of an important physiological role for Ech in Clade 2 and/or Clade 3 strains. Apart from Ech, no other (Ni-Fe) hydrogenases were identified.

4.4.7 Energy metabolism

Energy metabolism is interconnected with carbon and electron flux and governs many of the physiological processes involved with lignocellulosic ethanol production (Fig. 4.1). The principal forms of metabolic energy in bacteria include ion motive force, ATP and/or in some cases, pyrophosphate (PPi). Understanding inter-strain differences in energy metabolism may help provide insight into the mechanisms governing observed physiological differences (Table 4.4).

4.4.7.1 Transmembrane ion gradient generating/consuming reactions.

Transmembrane ion gradients can be used to drive endergonic reactions such as solute transport, including carbohydrate transport (Fig. 4.1, Table 4.2), and ATP synthesis. The mechanisms of balancing ion motive force with the other energy currencies are poorly characterized in *Thermoanaerobacter* spp., but the present *in silico* analysis indicates that the potential energy conserving mechanisms within the genus show significant intra-clade conservation (Fig. 4.4).

Strains of Clade 3 contain a 13-gene cluster (Figure 4.4) similar in genomic structure to the *mbx* genes of *Pyrococcus furiosus*. These gene products are proposed to encode a complex with Fd_{red}: $NAD(P)^+$ oxidoreductase activity (Schut *et al*., 2007), with energy released via the oxidation of Fd_{red} being conserved through the translocation of a cation. In *P. furiosus*, the *mbx* gene cluster has been proposed to play a role in the reduction of elemental sulfur, where it transfers electrons from Fd_{red} to NAD(P)H, which is subsequently oxidized via a NAD(P)H elemental sulfur oxidoreductase (Schut *et al*.,

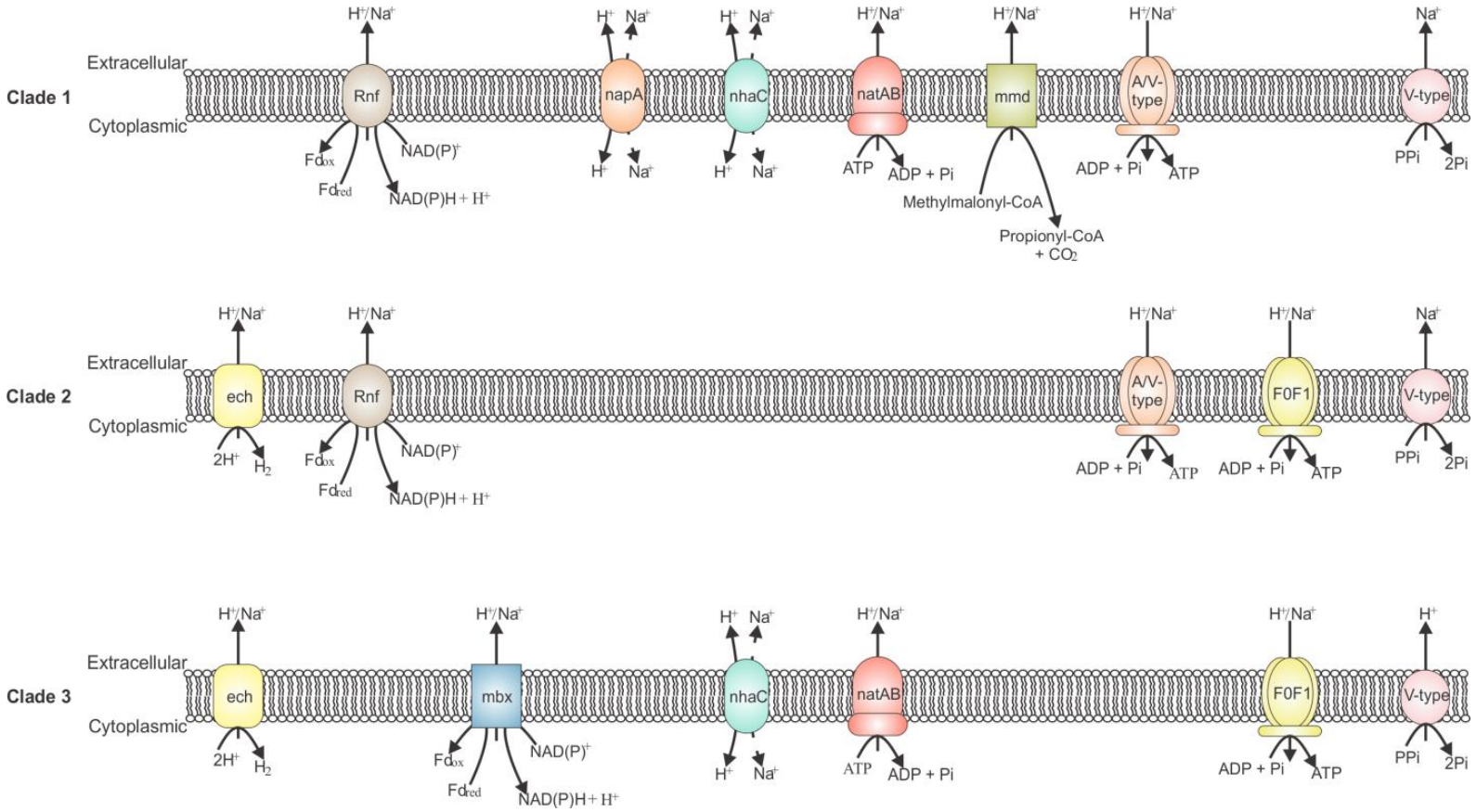


Fig. 4.4. Transmembrane ion gradient generating and consuming reactions involved with *Thermoanaerobacter* cellular energetics. With the exception of the natAB complex (Table A.3.5), all enzyme complexes show intra-clade conservation. Cation specificity is not inferred unless specifically discussed within Chapter 5. Dashed lines associated with napA and nhaC antiporters indicate a counter-directional flow of Na^+ ions in relation to H^+.

2007). However, the role in sulfur metabolism for *P. furiosus* has been inferred based upon microarray data indicating increased gene expression of the *mbx* gene cluster in response to the addition of elemental sulfur to the growth medium. Assuming the proposed Fd: $NAD(P)^+$ oxidoreductase activity of mbx is also observed in the Clade 3 strains, there is no evidence yet that suggests it is connected with sulfur reduction in these *Thermoanaerobacter* strains.

Clade 1 strains contain apparent remnants of the *mbx* gene cluster (Table A.3.5) as not all 13-genes could be identified. Five genes, the *mbxMJKLN* cluster, are not found immediately following orthologs of the *mbxABCDFGHH*' gene cluster as is observed in Clade 3. Clade 1 strains do therefore not appear to contain the genes necessary for a functional MBX complex. No orthologs were identified in the Clade 2 strains.

A functionally analogous Fd: $NAD(P)^+$ oxidoreductase system, the ion-translocating Rnf (NFO) complex, is present in all Clade 1 and Clade 2 strains (Fig. 4.4, Table A.3.5). The genomic organization is identical to what is reported for *Acetobacterium woodii* and multiple subunits of the *T. pseudethanolicus* 39E complex have been reported to be genetically similar to the partially characterized protein complex in *Ac. woodii* (Biegel *et al*., 2009; Biegel *et al*., 2011). In *Ac. woodii*, the complex is thought to translocate Na^+ ions, yet definitive proof has not yet been determined.

Na^+ energetics may play a prominent role in multiple *Thermoanaerobacter* strains. Of the five classes of recognized primary Na^+ pumps (Mulkidjanian *et al*., 2008), four of the classes are observed within the genus (Table A.3.5). One class, comprised of Na^+-translocating decarboxylation reactions, is limited to strains belonging to Clade 1. Hemme *et al*. (2011) identify both a methylmalonyl-CoA decarboxylase and an

oxaloacetate decarboxylase in the genomes of *T. pseudethanolicus* 39E and *Thermoanaerobacter* sp. X514. However, the present analysis is unable to find locus tags supportive of a membrane-associated oxaloacetate decarboxylase complex.

Annotation of the methylmalonyl-CoA decarboxylase encoding genes (Fig. 4.4, Table A.3.5) is supported by the presence of genes annotated to encode methylmalonyl-CoA mutase and methylmalonyl-CoA epimerase immediately adjacent to the methylmalonyl-CoA decarboxylase annotated gene sequences in all Clade 1 genomes. In all sequenced *Thermoanaerobacter* strains, genes annotated as oxaloacetate decarboxylase subunits (α, β, γ) can be identified, though they are never co-localized into a single gene cluster, as is observed with other *Clostridia* (van de Werken *et al*., 2008). Therefore, a functional membrane associated oxaloacetate decarboxylase complex may not exist in any of the *Thermoanaerobacter* strains.

Genes homologous to the *natAB* genes described in *Bacillus subtilis* (Cheng *et al*., 1997) are identified in multiple *Thermoanaerobacter* strains, but do not show intra-clade conservation (Table A.3.5). A third primary Na^+ pump includes the Rnf complex. The fourth and final class of potential Na^+ pumps identified in *Thermoanaerobacter* are the V-type inorganic pyrophosphatases (Table A.3.5). All V-type pyrophosphatases identified in Clades 1 and 2 appear to be orthologous to each other. Genes in Clade 3 though, are significantly different than the V-type pyrophosphatases identified in Clades 1 and 2. Key residues, as determined by Luoto *et al.* (2011), were identified for all annotated V-type pyrophosphatase genes via sequence alignments. The sequences present in Clades 1 and 2 share key residues identical to reported K^+-dependent, Na^+-

exporting V-type pyrophosphatases, while Clade 3 sequences share identical residues with K^+-independent, H^+-exporting versions (Table A.3.6).

Given that the cation specificity for any of the above mentioned ion-translocating processes discussed have not yet been determined experimentally for any *Thermoanaerobacter* strain, it is impossible to predict their role. Cook (2000) reports inhibited growth by *T. wiegelii* Rt8.B1 in the presence of the Na^+ ionophore monensin, suggesting the importance of maintaining a transmembrane Na^+ gradient for cell viability. However, genomic analysis identifies that a H^+ motive force is also expected in *T. wiegelii* Rt8.B1 (e.g. via a V-type pyrophosphatase). Both gradients may exist, and cellular demands are balanced by the use of H^+/Na^+ antiporters. Cation exchange antiporters, showing homology to the NhaC family, can be identified in all strains except Clade 2 (Figure 4.4). Clade 1 also contains genes homologous to NapA type antiporters. Surprisingly, analysis of the annotated transport systems for Clade 2 strains does not reveal any cation exchangers. Therefore, if Clade 2 strains generate both H^+ and Na^+ ion-motive forces, it is unclear how they balance these forces.

4.4.7.2 ATP and pyrophosphate (PPi) as energy currencies.

Synthesis of ATP as an energy currency, which is closely linked with carbohydrate and pyruvate metabolism (Fig. 4.1), can occur via glycolysis and through the production of acetate via acetate kinase in all *Thermoanaerobacter* strains. According to its annotation, both Clade 1 and Clade 3 genomes also encode an ATP-linked (in contrast to GTP-linked) PEP carboxykinase based on the enzyme commission number assigned to the annotated sequences (EC: 4.1.1.49) (Aich & Delbaere, 2007).

This bidirectional enzyme could either carboxylate PEP yielding 1 ATP + oxaloacetate or decarboxylate oxaloacetate at the expense of ATP. The flux models for *T. pseudethanolicus* 39E and *Thermoanaerobacter* sp. X514 (Hemme *et al*., 2011) suggest that, in these two strains, oxaloacetate decarboxylation occurs, but it is unknown if this translates universally throughout the genus.

ATP synthesis via ATP synthase genes can occur through two distinct means within *Thermoanaerobacter* spp. Strains of Clade 1 contain a 9-gene cluster designated as an A/V-type ATP synthase (Fig. 4.3, Table A.3.5), while Clade 3 strains contain the F_0F_1-ATP synthase in a conserved 8-gene cluster. Strains of Clade 2 contain both the AV-type and the F_0F_1-ATP synthase gene clusters in identical genomic organization as Clade 1 or Clade 3, respectively.

PPi as an alternative energy carrier to ATP has been observed during exponential phase for the phylogenetically related strains *Moorella thermoacetica* (Heinonen & Drake, 1988) and *Caldicellulosiruptor saccharolyticus* DSM 8903 (Bielen *et al*., 2010). Many of the key genomic elements to potentially generate and utilize PPi as a central energy carrier are similarly identified within all *Thermoanaerobacter* spp. Pyruvate kinase, which is identified in all *Thermoanaerobacter* strains, is typically considered to be responsible for the conversion of phosphoenolpyruvate to pyruvate. However, two distinct pyruvate phosphate dikinase (PPDK) genes (566-567 and 877 amino acids) are also universally conserved throughout the genus (Table 4.4). The longer of these genes shows >77.7% amino acid identity with the annotated PPDK gene (Csac_1955) from *Ca. saccharolyticus* DSM 8903, whose genome also encodes a pyruvate kinase (van de Werken *et al*., 2008). In *Ca. saccharolyticus* DSM 8903, the PPDK has been proposed to

function in a catabolic role during exponential growth whereby the conversion of PEP to pyruvate is coupled to the conversion of AMP + PPi to ATP + Pi (Bielen *et al*., 2010). Its presence in *Thermoanaerobacter* spp. suggests that PPi may also play a role in cellular energetics.

Additionally, sequence alignments (Fig. A.3.3) of the annotated 6-phosphofructokinase (PFK) genes (Table A.3.7) identifies that Clade 1 and Clade 3 strains possess one copy of PFK that contains the conserved Asp_{104} + Lys_{124} residues (*Escherichia coli* numbering) associated with PPi-dependence (Bapteste *et al*., 2003) and also found in *Ca. saccharolyticus* DSM 8903 (Bielen *et al*., 2010). Energy may additionally be conserved as an ion-motive force through the use of a membrane linked cation-translocating V-type pyrophosphatase (Chapter 4.4.7.1). Strains reported to use PPi as an energy currency during exponential growth are expected to have relatively high intracellular PPi/ATP ratios and low cytosolic PPiase activity. While, the intracellular ATP concentrations of exponential *T. wiegelii* Rt8.B1 cells (Cook *et al*., 2000) are reportedly higher than those observed for *Ca. saccharolyticus* DSM 8903 (Bielen *et al*., 2010), PPi levels and PPiase activity has not yet been investigated for any strain of the *Thermoanaerobacter* genus.

4.5 Conclusions

The analyses presented here have identified inter-strain differences at the genomic level within the *Thermoanaerobacter* genus that may account for differences in the industrial application of these bacteria in a CBP system. Based on genome content, Clade 2 strains seem most well suited to biomass hydrolysis and utilization, though these

strains have not yet been reported to have as high of ethanol yields as some Clade 1 strains do (Table 4.4). Conversely, despite the ethanologenic capabilities of Clade 1 strains, their genomes encode the fewest extracellular CAZymes of all strains within the genus (Table 4.1), which may potentially limit their hydrolytic capabilities. The genomes of Clade 3 strains show intermediate hydrolytic and substrate utilization capabilities, but also represent the most divergent lineage of the genus (Fig. 4.2) and may have yet unexamined potential.

The use of a specific strain for development of a universal approach to lignocellulosic ethanol production may represent an idealistic concept. Rather, strain selection may be specific to a single CBP system and dependent on the nature of the feedstock. Efficient conversion of arabinose to ethanol is of little value for bioenergy feedstocks such as eucalyptus, which contains comparatively low amounts of arabinan (0.3%), in contrast to switchgrass (3.0%) (Carroll & Somerville, 2009). Alternatively, extracellular xylan hydrolysis may not be an essential component in the microbial conversion of feedstocks such as softwoods, which have a low xylan and high glucomannan hemicellulose content (Gregg & Saddler, 1996; Saha, 2003). This is particularly true given that no extracellular glucomannanases were identified in any strain of *Thermoanaerobacter* (Table 4.1).

Development of biocatalysts with desired physiological characteristics using a strain with diverse, rather than specialized capabilities may be advantageous for constructing a robust and dynamic CBP system. For example, the construction of a single mutant (Δ*ldh*) in *T. mathranii* BG1, a platform organism of BioGasol (http://www.biogasol.com) (Mikkelsen & Ahring, 2007), has shown to improve ethanol

yields and still maintain substrate utilization capabilities similar to *T. mathranii* subsp. *mathranii* A3 (Clade 2). Alternatively, strategies that broaden the capabilities of a relatively specialized strain have also shown to be successful. The cloning and expression of a functional endoglucanase into *Thermoanaerobacter* sp. X514 (Lin *et al*., 2010), a comparatively good ethanol producer, has improved that strain's hydrolytic capabilities.

The purpose of this chapter was to evaluate genomic differences within members of the *Thermoanaerobacter* genus, which may influence strain suitability in a CBP co-culture system. Also, correlating genome content with the reported physiologies is a first step to help shape molecular engineering strategies for strain improvement. It is important to consider though that the analysis presented here is of the genomic potential of sequenced *Thermoanaerobacter* strains and is not of the observed phenotypes. Future experiments, such as expression profiling studies and enzymatic characterization, which can supplement the data presented here, will help to improve our understanding of the extent that the genomic potential is realized within these strains. Experiments targeted towards improving the hydrolysis of raw lignocellulosic biomass, understanding carbon transport and simultaneous utilization of mono-, oligo-, and polysaccharides, evaluating the genomic and regulatory basis for the observed differences in end-product synthesis ratios and improving the correlation between energy metabolism and end-product yields will all help develop *Thermoanaerobacter* spp. into more efficient CBP microorganisms.

This study is focused on components associated with lignocellulosic biofuel production and does not investigate the genomics associated with other phenotypes reported within the genus such as peptide and amino acid oxidation (Faudon *et al*., 1995),

metal-reduction (Slobokin *et al*., 1997; Roh *et al*., 2002; Gavrilov *et al*., 2003) or sulfur reduction (Lee *et al*., 1993). The potential impact that these phenotypes may have on biofuel production is not yet known. For example, vitamin B_{12} biosynthesis, associated with co-factor metabolism, has recently been shown to play an important role in improving observed ethanol yields (He *et al*., 2011; Hemme *et al*., 2011) in select *Thermoanaerobacter* strains. Therefore, we cannot discount the possibility that additional components of cellular physiology may similarly influence the lignocellulosic ethanol production capabilities of these strains. However, this work does identify key genomic criteria pertinent to strain evaluation for the development of a *C. thermocellum*-*Thermoanaerobacter* sp. co-culture and represents the most comprehensive comparative genomic analysis of the genus to date. Furthermore, comparative genomic analyses such as these can be useful in identifying important physiological questions to address through experimentation not only for *Thermoanaerobacter* spp., but also in other organisms of interest for lignocellulosic ethanol production through CBP.

4.6 Authors' contributions

Experiments were conceived and designed by Tobin J. Verbeke, Xiangli Zhang, Vic Spicer, Brian Fristensky, David B. Levin and Richard Sparling. Genome assembly was performed by Xiangli Zhang, while proteogenomic analyses were conducted by Tobin J. Verbeke, Vic Spicer and Oleg V. Krokhin. Genome annotation and comparative analyses was done by Tobin J. Verbeke. *De novo* CAZyme analyses were performed by Bernard Henrissat. The original manuscript, upon which this chapter is based, was

authored by Tobin J. Verbeke within input from Xiangli Zhang, Bernard Henrissat, David B. Levin and Richard Sparling.

4.7 Acknowledgements

We would like to thank John A. Wilkins of the Manitoba Centre for Proteomics and Systems Biology. The use of his facilities and equipment made our proteogenomic analyses possible. This work was supported by funds provided by a Genome Canada grant titled ‘‘Microbial Genomics for Biofuels and Co-Products from Biorefining Processes’’, the Natural Sciences and Engineering Research Council Strategic Grant (STPGP 365076) and by the University of Manitoba.

Chapter 5. Metabolic and label-free quantitative proteomic analysis of *Thermoanaerobacter thermohydrosulfuricus* WC1 on single and mixed substrates[13]

5.1 Abstract

Thermoanaerobacter spp. have long been considered suitable *Clostridium thermocellum* co-culture partners for improving lignocellulosic biofuel production through consolidated bioprocessing. Despite this, studies using global expression strategies to better understand carbon utilization and biofuel producing pathways have been limited to only a few strains thus far. To better characterize the overall physiology and refine carbon utilization pathways of a recently isolated, xylanolytic *Thermoanaerobacter* strain, *Thermoanaerobacter thermohydrosulfuricus* WC1, label-free quantitative proteomic analysis was combined with metabolic profiling. SWATH-MS[14] proteomic analysis quantified 832 proteins in each of six proteomes generated from mid-exponential phase cells grown on xylose, cellobiose, or a mixture of both. Despite encoding genes consistent with a carbon catabolite repression network observed in other Gram-positive organisms, simultaneous consumption of both substrates was observed. Lactate was the major end-product of fermentation observed across all conditions despite the high expression of gene products involved with ethanol and/or acetate synthesis, suggesting carbon flux in this strain may be subject to metabolite-based (allosteric)

[13] A modified version of this chapter has been submitted for publication. Contributing authors: **Verbeke TJ, Spicer V, Krokhin OV, Zhang X, Schellenberg JJ, Fristensky B, Wilkins JA, Levin DB, Sparling R**.

[14] SWATH-MS is a data independent acquisition method used in proteomics whereby all peptide parent ions are archived in sequentially cycled blocks, or *swaths*, allowing for retrospective extraction of peptide/protein intensities based on previously observed or computationally derived transitions (see Gillet *et al*., 2012).

regulation or is constrained by metabolic bottlenecks. The analyses presented here improve our understanding of *T. thermohydrosulfuricus* WC1 metabolism and identify important physiological limitations to be addressed in its development as a biotechnologically relevant strain in ethanologenic designer co-cultures.

5.2 Introduction

The use of designer co-cultures is an increasingly investigated strategy for achieving improved biofuel yields and conversion efficiencies from lignocellulosic biomass through consolidated bioprocessing (CBP) (He *et al*., 2011; Kridelbaugh *et al*., 2013; Zuroff *et al*., 2013). In a CBP platform, which involves concomitant enzyme production, biomass hydrolysis and biofuel production (Lynd *et al*., 2005), an ideal consortium would achieve: i) efficient and complete biomass hydrolysis; ii) simultaneous, rather than sequential, utilization of cellulose and hemicellulose constituent saccharides; and iii) industrially relevant biofuel yields. The selection of microorganisms is therefore an important component in achieving a successful designer consortium.

The extensive suite of lignocellulose degrading enzymes encoded by *Clostridium thermocellum* has made it an attractive candidate for CBP platforms (Zhang & Lynd, 2005; Raman *et al*., 2009; Rydzak *et al*., 2012). However, its inability to grow and produce biofuels from hemicellulose constituent saccharides, most notably pentoses (Ng *et al*., 1977; Demain *et al*., 2005), has often provided a rationale for the identification and investigation of suitable co-culture partners. Previous *C. thermocellum* co-cultures with bacteria possessing more diverse substrate utilization capabilities have resulted in

improved rates of biomass degradation and biofuel yield (Liu *et al*., 2008; He *et al*., 2011; Hemme *et al*., 2011; Kridelbaugh *et al*., 2013).

In such co-cultures, the hydrolysis of lignocellulosic biomass would generate a pool of mixed sugars available for fermentation. While the constituent co-culture members may have the potential to utilize many of the hydrolysis products, distinct preferences for certain carbon sources, at the exclusion of others, may exist. This preferential utilization by many bacteria, known as carbon catabolite repression (CCR), has been well documented in *Firmicutes* (Brückner & Titgemeyer, 2002; Deutscher, 2008), including strains of interest for lignocellulosic biofuel production such as *Clostridium cellulolyticum* (Abdou *et al*., 2008) and *Thermoanaerobacterium saccharolyticum* (Tsakraklides *et al*., 2012). Additionally, strains of the genus *Thermoanaerobacter* have been shown to exhibit CCR under some mixed sugar conditions (Hyun *et al*., 1985; Jones *et al*., 2002a), while showing no evidence of CCR under other conditions (Erbeznik *et al*., 1998a; Jones *et al*., 2002b; Lin *et al*., 2011).

Thermoanaerobacter thermohydrosulfuricus WC1, an isolate from woodchip compost, has recently been characterized (Chapter 2). Phylogenetic (Chapter 3) and genomic analyses (Chapter 4) of the strain have shown that it belongs to a divergent and lesser characterized lineage (Clade 3) of the *Thermoanaerobacter* genus. Its ability to hydrolyze and grow on xylan, which distinguishes *T. thermohydrosulfuricus* WC1 from many other *Thermoanaerobacter* strains (Chapter 4, Appendix A.5), coupled with its ability to utilize most hemicellulose constituent saccharides, suggest that *T. thermohydrosulfuricus* WC1 may be an effective *C. thermocellum* co-culture partner. However, the simultaneous utilization of lignocellulose constitutive saccharides such as

cellobiose (from cellulose hydrolysis) and xylose (from xylan or hemicellulose hydrolysis) by this organism is not yet understood. Additionally, molar ethanol yields of *T. thermohydrosulfuricus* WC1 in comparison to other *Thermoanaerobacter* spp. is comparatively low (Chapter 4 – Table 4.4), ultimately reducing its industrial relevance. Identifying approaches to address these limitations is hampered by a lack of understanding of *T. thermohydrosulfuricus* WC1 physiology.

To date, only a few studies have investigated *Thermoanaerobacter* physiology and gene expression profiles at a global level (Hemme *et al*., 2011; Lin *et al*., 2011), and these studies have focused on strains belonging to a lineage distinct from *T. thermohydrosulfuricus* WC1. As such, metabolic and quantitative proteomic analyses were conducted to achieve two major objectives.

First, *C. thermocellum* mediated hydrolysis of cellulose is reported to yield primarily higher order cellodextrins with little concomitant glucose (Zhang & Lynd, 2005). As *T. thermohydrosulfuricus* WC1 encodes two distinct annotated cellobiose specific phosphotransferase system (PTS) type transporters, and a connection between PTS-mediated sugar transport and CCR is well established in *Firmicutes* (Brückner & Titgemeyer, 2002; Warner & Lolkema, 2003), the ability for *T. thermohydrosulfuricus* WC1 to simultaneously use cellobiose and xylose, two important lignocellulose constituent saccharides, was investigated. Secondly, proteomic analyses was performed under single and multiple substrate conditions as a means of better characterizing carbon flux in *T. thermohydrosulfuricus* WC1 through correlation of the proteome with the observed physiology.

5. 3 Materials and methods

5.3.1 Bacteria and culture conditions

Independent glycerol stocks of *T. thermohydrosulfuricus* WC1 were revived for each experimental replicate performed in this study. Cultures were grown at 60^{o}C in ATCC 1191 medium as previously described (Chapter 2.3.2) with the following exceptions: i) the concentration of yeast extract was reduced from 2 g/L to 0.67 g/L to reduce the potential for growth on yeast extract components and ii) gassing: degassing butyl rubber stoppered bottles to make the medium anoxic were shortened to three cycles (1 min gassing: 3 min degassing) with a final gas cycle of 100% nitrogen (Daniels *et al*., 1986). Anoxic, filter-sterilized sugar solutions were added to the medium post-autoclaving so that medium contained (final concentrations) 5 g/L xylose, 5 g/L cellobiose or 5 g/L xylose plus 5 g/L cellobiose (sugar mix). An equal volume of anoxic, filter sterilized water was used for the no substrate conditions. *T. thermohydrosulfuricus* WC1 has since been deposited in the Deutsche Sammlung von Mikroorganismen und Zellkulturen culture collection as DSM 26960.

5.3.2 Genome analysis

Analysis of the *T. thermohydrosulfuricus* WC1 annotated genome was performed as previously described (Chapter 4.3.3) using the Joint Genome Institute's Integrated Microbial Genome's Expert-Review (IMG-ER) online tool (Markowitz *et al*., 2009). The genome is publicly available in GenBank under accession number AMYG00000000.

An *ad hoc* perl script was developed to identify putative catabolite responsive elements (*cre*) in the *T. thermohydrosulfuricus* WC1 genome using the reported

degenerate consensus sequences from *Bacillus subtilis* (Miwa *et al*., 2000), *Lactobacillus casei* (Monedero *et al*., 1997) and *Clostridium difficile* (Antunes *et al*., 2012) as queries.

5.3.3 Growth and metabolic analyses

5.3.3.1 Growth

Revived glycerol stocks were serially passaged three times at 60^{o}C on 5 g/L xylose every 24 hours prior to inoculating (10% vol/vol) 10 ml tubes containing 5 g/L xylose, 5 g/L cellobiose, 5 g/L xylose plus 5 g/L cellobiose or no substrate. The optical density (OD_{600}) of inoculated cultures was routinely measured until an increase in OD_{600} was no longer observed. Reported doubling times are averages of three independent experiments with each experiment containing three biological replicates.

5.3.3.2 Metabolic analyses

Ten ml cultures, prepared as described, were grown to target optical densities based on the growth profiles observed. For all conditions (excluding the no substrate control) three tubes per experiment were harvested immediately after inoculation (T=0) or upon reaching a target OD_{600} of 0.10, 0.20, 0.40, 0.70, 0.80 or 0.85. Cultures were then analyzed for end-product and protein/biomass synthesis as well as residual sugar concentrations.

Gas analysis (H_2 and CO_2) was conducted by injecting triplicate 1 ml samples per tube into a Varian 490 Micro-GC gas chromatograph (Agilent Technologies, CA) using nitrogen as a carrier gas. Gas solubility was accounted for in determining concentrations by measuring barometric pressure, tube pressure and temperature as described (Sander *et*

al., 1999). Additionally, the bicarbonate fraction was determined as part of the total CO_2 calculation (Darrett & Grisham, 1995). Acetate, lactate, xylose and cellobiose concentrations were determined via high-performance liquid chromatography (Waters Corp., Milford, MA) equipped with a refractive index detector (Model 2414) and an ion exclusion column (Aminex HPX-87H; Bio-Rad Laboratories, CA) using sulfuric acid (5 mM) at a flow rate of 0.6 ml/min as the mobile phase. Ethanol and protein measurement assays were performed as described (Chapter 2.3.8).

5.3.4 Proteomic analyses

5.3.4.1 Growth

Cultures (10 ml) were prepared and grown on xylose, cellobiose, or xylose plus cellobiose from two independent, glycerol stock cultures of *T. thermohydrosulfuricus* WC1 (designated as Culture A or Culture B). Upon reaching a target OD_{600} of 0.40, representing mid-exponential phase, cultures were centrifuged at 4,700 × *g* for 10 min and the resulting pellet washed three times in 600 µl phosphate buffered saline (8 g/l NaCl, 0.2 g/l KCl, 1.44 g/l Na_2HPO_4, 0.24 g/l KH_2PO_4, pH = 7.4). Cell pellets derived from the same inoculum (Culture A or Culture B) and grown under the same conditions were combined (10 × 10 ml culture) and treated as a single sample. Thus, six samples (3 conditions × 2 glycerol stock cultures) were used for proteomic analysis. Samples were frozen at -80^{o}C until protein extraction could be performed.

5.3.4.2 Protein extraction

A modified version of the filter aided sample preparation (FASP) protocol (Wiśniewski *et al*., 2009) was used for protein extraction. Cell pellets were resuspended in 1 ml lysis buffer (4% (wt/vol) sodium dodecyl sulfate, 100 mM Tris-HCl, 0.1 M dithiothreitol, pH 7.6) and heated at 95^{o}C for 5 minutes. The samples were continuously sonicated at an output of 3.5 for two minutes using a Branson Sonifier 450 (Branson Ultrasonics Corporation, Danbury, CT) and subsequently centrifuged at 16,000 × *g* for 20 min. The resulting supernatant was transferred to an Amicon Ultra-15 10K filter device (Millipore, Billerica, MA) and washed three times in 12 ml urea solution (8M urea in 0.1 M Tris-HCl, pH 8.5). Each wash step was centrifuged at 4,000 × *g* for at least 10 minutes until the final volume remaining in the filter tube was less than 1 ml. Two ml of iodoacetamide solution (50 mM iodoacetamide in urea solution) was added to the filter device and incubated at room temperature in the dark for 20 min. After centrifugation, the filter membrane was washed twice more with an additional 12 ml of urea solution. A 50 µl aliquot was taken from the filter unit and analyzed using the BCA protein assay kit (Pierce Chemical Co., Rockford, IL) to estimate the total protein content of the sample. The filter membrane was washed twice with 12 ml of 50 mM ammonium bicarbonate in water and the remaining protein was trypsin digested for 18 hours at room temperature (trypsin to protein ratio of 1 µg: 100 µg). The filter unit was transferred to a new collection tube, spun at 4,000 × *g* for 10 min and the filtrate retained for downstream analysis. The membrane was washed with 1 ml of 0.5 M NaCl and the resulting filtrate combined with the corresponding previous filtrate and stored at -80^{o}C.

5.3.4.3 Peptide purification and mass spectrometry (MS) analysis

Peptide concentrations in the combined filtrate were measured using a Nanodrop (Nanodrop Technologies, Wilmington, DE). Acidified aliquots of combined filtrate containing ~150 μg of the digest were purified by loading onto a 1 × 100 mm C18 column (5 μm Luna C18(2); Phenomenex, Torrance, CA) and eluted using 80% acetonitrile (vol/vol). Purified aliquots were lyophilized and re-dissolved in buffer A (0.1% formic acid in water) for subsequent LC-MS analysis.

A splitless nano-flow 2D LC Ultra system (Eksigent, Dublin, CA) with 10 μL sample injection via a 300 μm × 5 mm PepMap100 trap-column (Thermo Fisher Scientific; Rockford, IL) and a 100 μm × 200 mm analytical column packed with 5 μm Luna C18(2) were used for all LC separations. Both eluents A (water) and B (acetonitrile) contained 0.1% formic acid as the ion-pairing modifier. Tryptic digests were separated using a 0.33% gradient (0.5-36% acetonitrile over 107 minutes), corresponding to 2 hours of LC-MS instrument time. Quantities of 1.0 μg and 0.5 μg of the *T. thermohydrosulfuricus* WC1 digest were used for injection into IDA (information dependent acquisition) and SWATH-MS (Gillet *et al*., 2012) analyses, respectively. Aside from the differences in the amount of injected digest, the identical chromatography settings on the same mass spectrometer permitted the direct utilization of IDA-based identification data to drive the subsequent SWATH-MS analyses.

A TripleTOF5600 mass spectrometer (Applied Biosystems, Foster City, CA) was used for both IDA acquisition and SWATH-MS analyses. Each scan in the standard MS/MS IDA included a 250 ms survey of MS spectra (m/z 300-1500) and up to 20 MS/MS measurements on the most intense parent ions (300 counts/sec threshold, +2 - +4

charge state, m/z 100-1500 mass range for MS/MS, 100 ms each). Previously targeted parent ions were excluded from repetitive MS/MS acquisition for 12 sec (50 mDa mass tolerance). Raw spectra files were converted into Mascot Generic File (MGF) format for peptide/protein identification. Each cycle in SWATH-MS acquisition mode consisted of a 250 ms survey MS spectra followed by 34 MS/MS spectra in 25 Da wide parent blocks (100 ms each) giving a total of 3.65 seconds per cycle.

5.3.4.4 Data processing and analyses

Peptides were identified from the observed MS/MS spectra using an in-house GPU peptide search engine (McQueen *et al*., 2012) from a single missed cleavage tryptic database derived from the *T. thermohydrosulfuricus* WC1 genome annotation with only the fixed post-translational modification of the carbamidomethlyation of cysteine residues (+57.021 Da) applied. The GPU search engine settings were 20 ppm on parent mass ions and a 0.2 Da window on fragment ions respectively and peptides with expectation log(e) ≤ -1 were reported. Each identified peptide in the six IDA runs was aligned against its corresponding collision-induced dissociation fragment spectra to form a fragment ion library with each hypothetical transition (b or y ions for M/Z where Z = +1, Z = +2) computed from the peptide sequence and integrated over a 20 ppm window of the target spectrum. A strategy was then devised to combine these six libraries into a single reference database, which could potentially allow for peptides observed in only a single IDA run to be detected in the potentially deeper SWATH-MS analysis.

For each fragment ion library the most intense identification of every non-redundant peptide sequence was selected. Across the collection of IDA runs, the

retention times of these non-redundant peptides were averaged, with the identified peptides with the greatest signal providing the fragments. This non-redundant "averaged" collection was then formatted to a tab delimited collection of Q1/Q3 transitions for processing by PeakView software (AB Sciex, Farmingham, MA). This ion library contained 111,856 transitions spanning 9,110 peptides belonging to 1,313 proteins. PeakView transition XIC extraction within detection windows of ± 5 minutes and 20 ppm were applied to all six SWATH-MS runs and the resulting peptide-level intensity report contained 6,264 entries spanning 1,171 proteins. Proteins identified with only a single member peptide were excluded from the analysis.

The expression measurement for each protein was total ion current (TIC), measured as the sum of the signal intensities for each protein's observed member peptides, and expressed on a $\log_2$ scale. Normalized TIC values (nTIC) of the raw data for each protein were computed by dividing the observed TIC by the $\log_2$ value of its mass in kilodaltons. An additional filtering step of removing proteins with $\log_2$ nTIC values having a difference in magnitude ≥2 across biological replicates (Culture A and Culture B) was applied yielding a final dataset of 832 proteins quantified across all six conditions.

Relative expression levels between growth conditions were calculated for each protein using a simple transformation, which combines difference measurements into a unified expression (W) based on protein level Z-scores and is measured in units of standard deviation. First, expression values for all populations are normalized to a mean of zero and a standard deviation of one. For each comparison, four difference measurements were determined. These difference measurements incorporate the intra-

replicate variability ($R0 = B_X - A_X$, $R1 = B_Y - A_Y$) as well as the inter-condition variability ($Z0 = A_X - A_Y$; $Z1 = B_X - B_Y$) where A and B represent different glycerol stock cultures and X and Y represent different substrates. Vectors were computed as the distance from the origin to the mapped points ((R0, R1) or (Z0, Z1)). The difference between the magnitudes was scaled by the ratio of their respective population widths yielding W with the sign being computed via the angle subtended from the x-axis to (Z0, Z1). Angles between 315^o and 135^o are positive, or otherwise negative. Finally, W was normalized to W_{net} and for the present work is represented as: i) $W0_{net}$ = xylose grown cells – cellobiose grown cells; ii) $W1_{net}$ = xylose grown cells – sugar mix grown cells; and iii) $W2_{net}$ = cellobiose grown cells – sugar mix grown cells. Therefore, in $W0_{net}$ (for example), for any protein, positive values represent higher nTIC values observed for growth on xylose in comparison to cellobiose, while negative values represent the opposite.

Global expression trends were analyzed using an in-house analysis platform, which groups proteins/genes into membership in "higher order variables (HOVs)" extracted from the IMG-ER "export gene information" function (Markowitz *et al*., 2009) for an individual genome. These include: i) METACYC pathways (Caspi *et al*., 2008); ii) enzyme class (EC) numbers (IUBMB, 1992); iii) clusters of orthologous group (COG) classes (Tatusov *et al*., 2000); or iv) KEGG modules (Kanehisa *et al*., 2008). Coarse global expression trends associated with an HOV are defined as an asymmetry (up-regulated or down-regulated) in the expression of proteins in that HOV, relative to the expression profiles of the overall population. For this study, HOV analysis has been limited to COG categories and a W_{net} value of ±1 standard deviation (representing the

outermost 32% of the population) has been used as an initial guide to investigate population regulation asymmetries.

5.4 Results

5.4.1 Genomic analysis of CCR network genes

All genes needed to encode for a potential carbon catabolite control protein A (ccpA)-dependent CCR network in *T. thermohydrosulfuricus* WC1 were identified. The identified homologous protein sequences include: i) a histidine-containing protein (HPr) (TthWC1_1711); ii) a catabolite repression HPr-like protein (CrH) (TthWC1_2012); and iii) a HPr kinase/phosphotase (TthWC1_1297). The His_{15} and Ser_{46} residues of HPr in *Bacillus subtilis,* which activate HPr upon phosphorylation for either PTS-mediated transport or CCR network signalling respectively (Ye *et al*., 1994; Galinier *et al*., 1999; Viana *et al*., 2002), were conserved in the *T. thermohydrosulfuricus* WC1 annotated protein sequence (Fig. A.4.1). Additionally, sequence alignments (Fig. A.4.2) of the CrH protein confirms a conserved Ser_{46} residue suitable for HPr kinase dependent phosphorylation, as well as the absence of the His_{15} residue, as described in *B. subtilis* (Galinier *et al*., 1997). Nine annotated proteins homologous to the *lacI*-family transcriptional regulator protein ccpA were also identified (based on assignment into COG1609), making it difficult to determine which homolog, if any, may function as ccpA *in vivo*.

The genome was also searched for catabolite responsive elements (*cre*), which may serve as putative ccpA binding sites for repressing transcription. Twenty-three sites homologous to the *B. subtilis* sequence and one site homologous to the *C. difficile*

sequence were identified (Table A.4.1). The putative *cre* sequences were not associated with any genes expected to be involved in xylose catabolism with the exception of one sequence upstream of a gene cluster that encodes a putative ribose transporter ((TthWC1_1369-1371; TthWC1_1373), which may be involved with xylose uptake in some *Thermoanaerobacter* strains (Chapter 5.4.6). While sequences were found both internal to coding sequence(s) (CDS), as well as in intergenic regions, 25% of the identified sequences were also found to be on a strand opposite to a CDS and are unlikely to represent *cre* sequences.

5.4.2 Growth and metabolic analyses

Based on the genomic analysis of a potential CCR network in *T. thermohydrosulfuricus* WC1, the substrate utilization profiles of cultures grown on xylose, cellobiose and xylose plus cellobiose were evaluated. Cells transferred from a xylose grown inoculum consistently lagged when transferred to medium containing cellobiose only (Fig. 5.1). A lag was not observed when transferred to xylose or xylose plus cellobiose containing medium. All cultures, with the exception of the no substrate control, grew to a similar maximum OD_{600} (~0.85). The fastest doubling time (2.79 ± 0.11 h/generation) was observed for cells grown on cellobiose only, while growth on xylose (3.53 ± 0.25 hr/generation) and xylose plus cellobiose (3.25 ± 0.16 hr/generation) was slower.

To account for differences in growth rates, metabolic analysis across conditions was normalized by using cultures harvested at pre-determined target OD_{600} values. The measured OD_{600} values from the cultures showed a mean deviation from the target OD_{600}

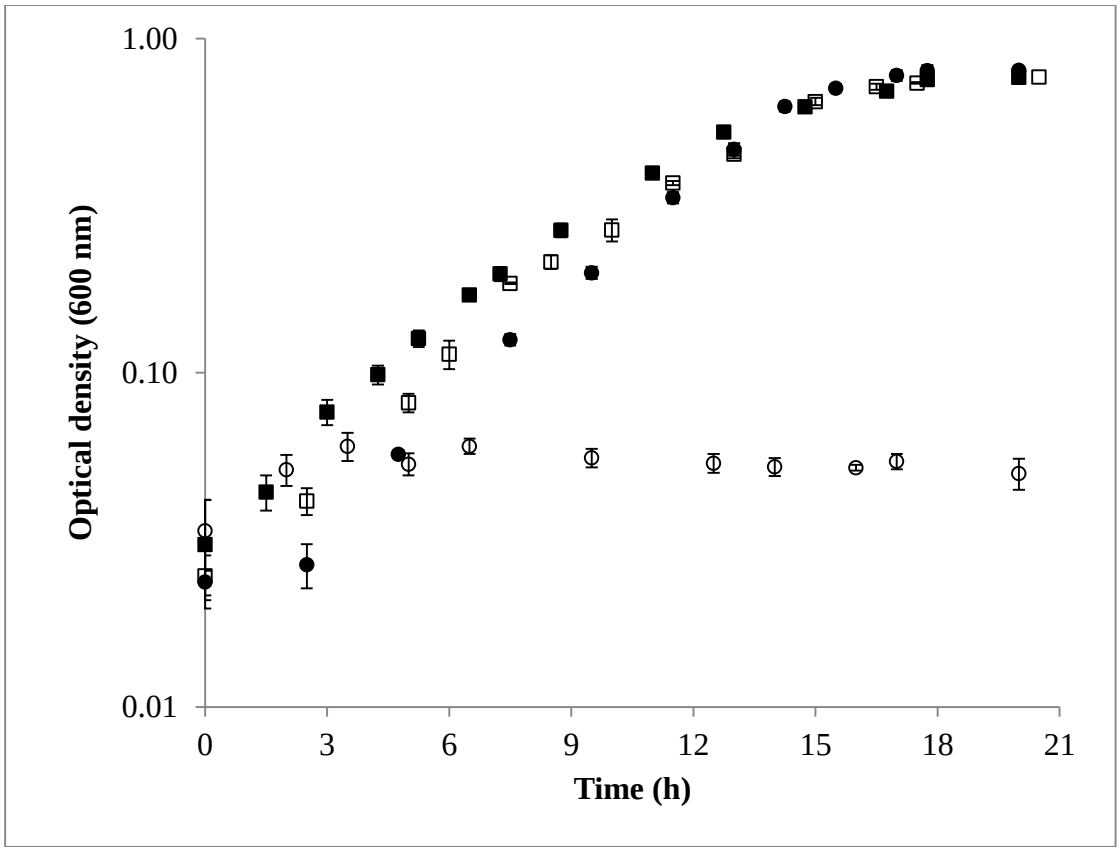


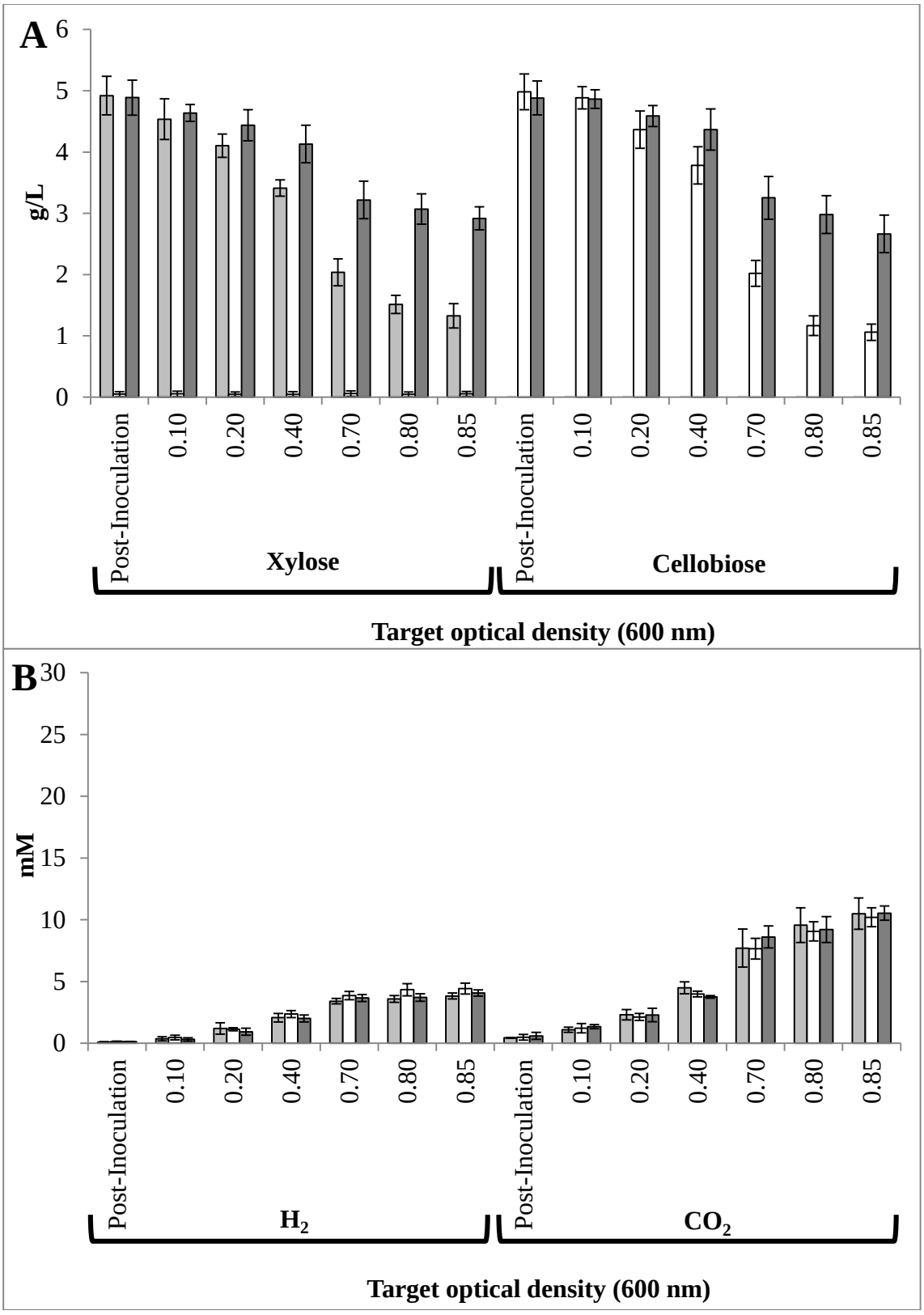
Fig. 5.1. Typical growth curve of *T. thermohydrosulfuricus* WC1 under single substrate, mixed substrate, or no substrate conditions tested. Symbols are as follows: 5 g/L xylose (black squares), 5 g/L cellobiose (black circles), 5 g/L xylose plus 5 g/L cellobiose (white squares) or with no added sugar (white circle). Reported values represent the average of three biological replicates from a single experiment. Error bars represent the standard deviation between replicates.

values of 1.54%, with a maximum deviation of 5.5% (Fig. A.4.3). The differences between observed OD_{600} values and the target optical densities were therefore considered negligible.

Growth on all substrates was in carbon excess conditions for *T. thermohydrosulfuricus* WC1 (Fig. 5.2A). Residual xylose and cellobiose concentrations were higher for the mixed substrate condition throughout growth in comparison to the single substrate conditions. This was likely due to the simultaneous utilization of both xylose and cellobiose in the mixed substrate condition suggesting a lack of CCR-type repression with this sugar combination. Further, analyses of the specific substrate consumption rates (Table 5.1) identified that xylose was consumed at a faster rate than cellobiose under single sugar conditions. This trend was also observed in the mixed substrate conditions despite an overall decrease in the consumption rates of each specific sugar. End-product and biomass analyses revealed that most of the major products formed from sugar consumption were accounted for (Fig A.4.4; Table A.4.2) and were detected throughout growth, with lactate being the most abundant end-product.

5.4.3 Proteomic analysis and global expression trends

Six two-hour IDA runs were conducted on *T. thermohydrosulfuricus* WC1 yielding 205,309 MS/MS spectra (Table A.4.3). Over half of the MS/MS spectra yielded peptide identifications and the ratio of non-redundant peptides to overall peptides was approximately 1: 3. SWATH-MS analyses allowed for the quantification of 892 proteins under all three growth conditions, which was reduced to 832 after filtering data with large differences between biological replicates. The dynamic range of observed $\log_2$ nTIC


A
g/L
0
1
2
3
4
5
6
Post-Inoculation
0.10
0.20
0.40
0.70
0.80
0.85
Xylose
Cellobiose
Target optical density (600 nm)
B
mM
0
5
10
15
20
25
30
H2
CO2
Target optical density (600 nm)

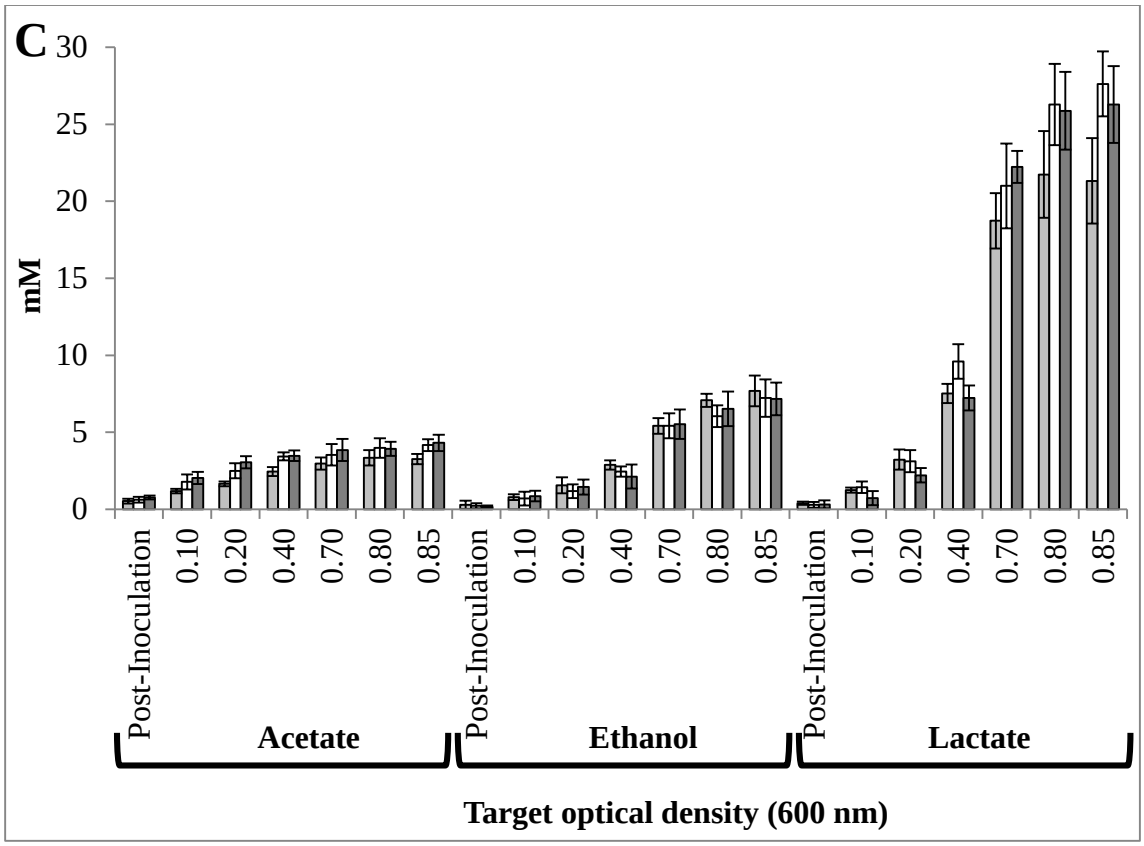


Fig. 5.2. Substrate consumption and metabolite production of *T. thermohydrosulfuricus* WC1 on single substrate or mixed substrate conditions tested. Consumption or end-product formation on 5 g/L xylose (light gray), 5 g/L cellobiose (white) or 5 g/L xylose plus 5 g/L cellobiose (dark gray) at target optical densities. Post-inoculation refers to cultures sacrificed immediately after inoculation from xylose grown parent cultures. (A) Residual sugar concentrations. (B) H_2 and CO_2 production. (C) Acetate, ethanol and lactate production.

Table 5.1. Specific substrate utilization rates of *T. thermohydrosulfuricus* WC1.

	Average[a] substrate consumption rates[b] (mmol consumed per gram of cellular protein per hour)	
Growth substrate	Xylose consumption	Cellobiose consumption
Xylose	29.03 ± 2.41	NC[c]
Cellobiose	NC	14.61 ± 0.97
Sugar mix	17.35 ± 3.55	8.20 ± 0.63

[a]Consumption rates determined independently for each experimental replicate. Averages represent the calculated values from experiments done in triplicate.

[b]Rates calculated as the slope of semi-logarithmic plots for each substrate consumed (Islam *et al*., 2006). In all cases, the R^2 values exceeded 0.93.

[c]Not calculable as substrate not consumed under condition tested (Fig. 5.2A).

values was >17. Good correlation was observed between biological replicates (Fig. A.4.5) as linear regression analysis of the replicates showed a slope >0.96 and R^2 values >0.97. The cross-condition correlations were also high (Fig. A.4.6) suggesting relatively stable gene expression patterns under all conditions tested. Proteomes derived from xylose grown vs. cellobiose grown cells were the most different (Fig. A.4.6A), while proteomes from cellobiose vs. sugar mix grown cells were the most similar (Fig. A.4.6C), suggesting that cellobiose may have a broader impact on global expression profiles than does xylose.

To determine if substrate specific global expression profiles emerged, the observed proteins were grouped according to their respective annotated COG class and W_{net} values found in the outermost 32% of each population were used to identify regulatory responses of pathways as a whole. The most pronounced changes, in terms of number of proteins affected, were observed with proteins in COG class G – carbohydrate metabolism and transport (Table 5.2). Genes involved in energy production and conversion (COG class C) were up-regulated on xylose alone in comparison to cellobiose alone or the sugar mix. Conversely, proteins in COG class E – amino acid transport and metabolism, and class J – translation, were up-regulated in cellobiose containing cultures (as a single substrate or in the sugar mix) in comparison to growth on xylose alone. This trend is consistent with the increased growth rates observed on cellobiose or the sugar mix in comparison to growth on xylose alone. The fewest proteins up- or down-regulated were observed when comparing the proteomes from the cellobiose grown culture against the sugar mix culture, with most of the changes occurring in COG class G.

Table 5.2. Number of proteins in specific COG classes up- or down-regulated[a] under specific growth conditions.

		Xylose vs. cellobiose		Xylose vs. sugar mix		Cellobiose vs. sugar mix	
COG class[b]	Description	Up	Down	Up	Down	Up	Down
C	Energy production and conversion	13	5	18	3	6	4
D	Cell cycle control and mitosis	0	0	1	2	0	2
E	Amino acid metabolism and transport	9	18	10	17	8	3
F	Nucleotide metabolism and transport	3	6	3	6	0	2
G	Carbohydrate metabolism and transport	20	11	17	10	15	13
H	Coenzyme metabolism	0	0	6	3	5	3
I	Lipid metabolism	2	3	2	1	0	0
J	Translation	2	8	1	10	3	0
K	Transcription	0	0	5	0	2	1
L	Replication and repair	0	0	0	0	2	4
M	Cell wall/membrane/envelope biogenesis	6	1	7	1	0	4
O	Post-translational modification, protein turnover, chaperone functions	1	2	1	2	1	2
P	Inorganic ion transport and metabolism	1	2	0	2	4	3
Q	Secondary metabolites biosynthesis, transport and catabolism	0	0	0	0	0	2
R	General function prediction only	8	7	7	8	0	0
S	Function unknown	6	1	8	3	1	2
T	Signal transduction	0	3	1	3	0	2
Total		71	67	87	71	47	47

[a]Up- or down-regulation limited to proteins having a relative Z-score expression ratio (W_{net}) outside (±) one standard deviation from the population mean.

[b]Limited to COG classes with at least one protein up- or down-regulated

5.4.5 Proteomic analysis of CCR network genes

Proteins of a potential ccpA-dependent CCR network (HPr, CrH, HPr kinase/phosphatase) were not observed under any conditions. In addition, only three of the nine annotated *ccpA* homologs (TthWC1_0680, TthWC1_2179 and TthWC1_2451) were observed in the proteome (Table A.4.4). Thus, given that these proteins were not detected, in conjunction with the simultaneous utilization of xylose and cellobiose (Fig. 5.2; Table 5.1), it suggests that if a ccpA-dependent CCR network exists in *T. thermohydrosulfuricus* WC1, that cellobiose consumption does not exert a repressive effect preventing xylose catabolism, nor vice versa.

5.4.6 Cellobiose and xylose transport and hydrolysis

Comparison of the average nTIC values between subunits of both annotated cellobiose-specific PTS-type transporters (Fig. 5.3A, Table A.4.4) suggest similar expression of both gene clusters when grown on cellobiose alone. However, the large negative $W0_{net}$ and $W1_{net}$ values observed for the TthWC1_0920-0922 gene cluster in comparison to the TthWC1_2146-2148 gene cluster suggests that expression of the former is more responsive to the presence of cellobiose. Both PTS-type transporters are co-localized in the genome with transcriptional anti-terminator proteins containing PTS-regulatory domains (pfam00874), which may regulate their expression, similar to what has been proposed for PTS-transporters in *Thermoanaerobacter* sp. X514 (Lin *et al*., 2011). The TthWC1_0920-0922 gene cluster is also co-localized with glycoside hydrolase (GH) family 1 and family 4 proteins, which may be involved with cellobiose hydrolysis (see below), while the TthWC1_2146-2148 gene cluster has no obvious genes

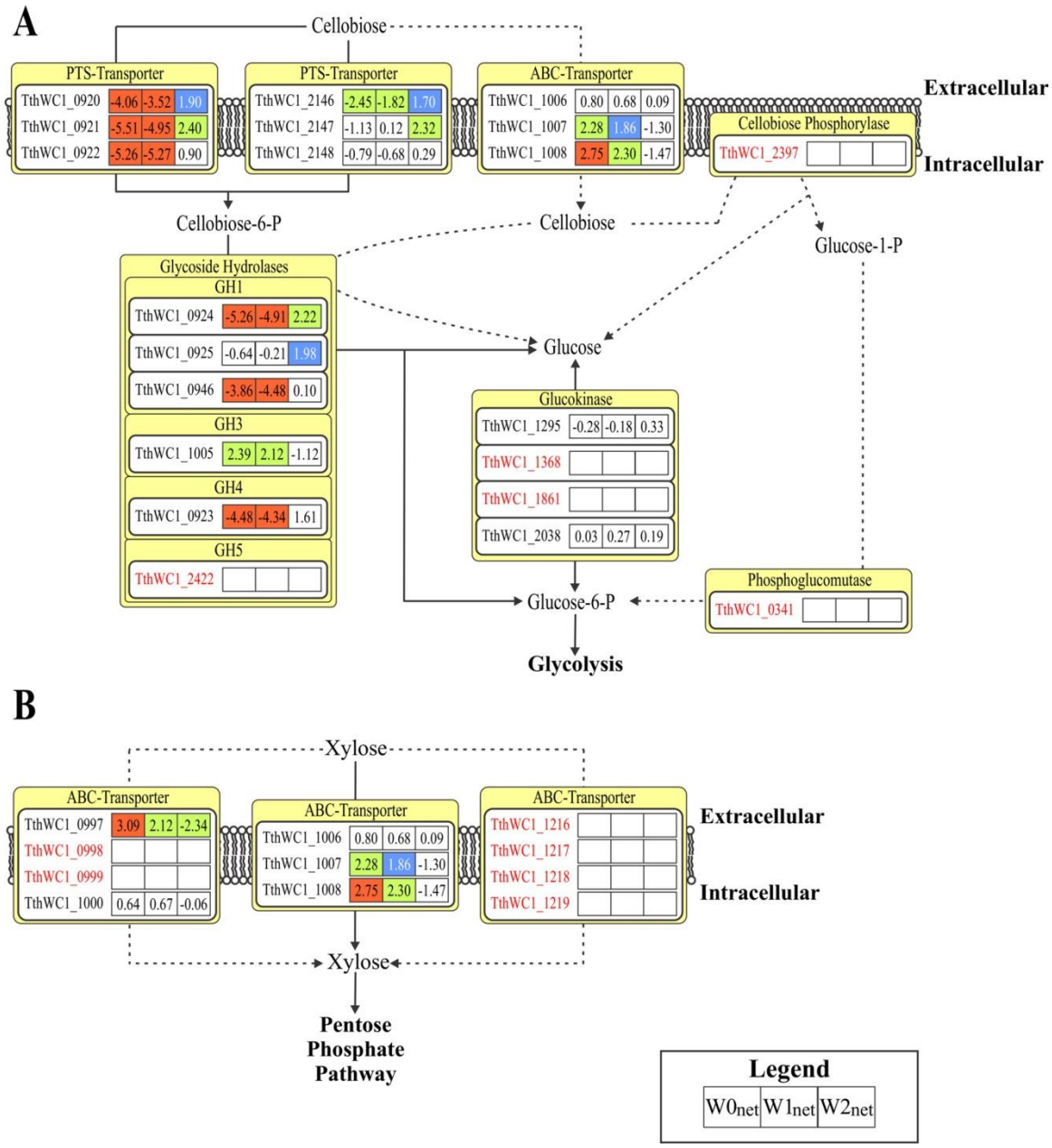


Fig. 5.3. Normalized relative Z-score expression ratios (W_{net}) of select *T. thermohydrosulfuricus* WC1 sugar transport and hydrolysis mechanisms. Transport and hydrolysis mechanisms have been limited to cellobiose (A) or xylose (B) and reflect potential modes of entry into the cell and hydrolysis for entry into either the glycolytic (cellobiose) or pentose phosphate (xylose) pathways for subsequent catabolism. Locus

Fig. 5.3 cont.

tags in black were observed in the proteomes, while those in red were not. Solid lines represent potential routes when transport is through phosphotransferase system (PTS)-mediated mechanisms, while dashed lines represent ABC-mediated transport. $W0_{net}$, $W1_{net}$ and $W2_{net}$ are as defined (Chapter 5.3.4.4). The four-colour scheme represents the populations of W_{net}, where red boxes are the outermost 1% ($|W_{net}| \geq 2.58$), green boxes are the outermost 5% ($|W_{net}| \geq 1.96$), blue boxes are the outermost 10% ($|W_{net}| \geq 1.65$), and white boxes are the innermost 90%.

involved with cellobiose hydrolysis proximally located.

The genome also encodes a gene cluster (TthWC1_1006-1008) orthologous to the proposed cellobiose ABC-type importer (*cglF-cglG*) in *Thermoanaerobacter brockii* subsp. *brockii* HTD4 (Breves *et al*., 1997). Its expression though, correlates with the presence of xylose rather than cellobiose (Fig. 5.3).

Multiple cytosolic GH enzymes potentially capable of cellobiose hydrolysis, in addition to a membrane bound cellobiose phosphorylase, are annotated in *T. thermohydrosulfuricus* WC1. Three genes (TthWC1_0923; TthWC1_0924; TthWC1_0925) are immediately adjacent to one of the cellobiose-specific PTS-type transporters. Only two of the genes (TthWC1_0923; TthWC1_0924) follow a similar expression profile to the cellobiose responsive PTS-type transporter (Fig. 5.3). An additional GH1 enzyme, TthWC1_0946, was also observed at a higher expression level in the presence of cellobiose, though its average nTIC value is less than TthWC1_0923 or TthWC1_0924 by a $\log_2$ difference ≥1.58 under cellobiose only conditions (Table A.4.4). Of the three GH enzymes up-regulated in response to cellobiose, TthWC1_0924 is the most abundant in both the cellobiose only and the sugar mix conditions. Neither the cellobiose phosphorylase, nor the annotated phosphoglucomutase needed to convert glucose-1-P to glucose-6-P, were detected in the proteomes supporting the proposal that cellobiose is transported via PTS-mediated mechanisms and not ABC-dependent mechanisms in *T. thermohydrosulfuricus* WC1 (Fig. 5.3A).

Expression of the TthWC1_1006-1008 transporter correlates with growth on xylose (Fig. 5.3B). Its function as a presumed xylose transporter is further supported by its genomic context, which includes an annotated extracellular endo-xylanase (GH10

enzyme) (TthWC1_1010), an acetyl-xylan esterase (TthWC1_1011) and an extracellular GH52 beta-xylosidase (TthWC1_1012). Proteins in the region spanning TthWC1_1005-1023, which includes the xylose isomerase (TthWC1_1022) and xylulose kinase (TthWC1_1021), are more highly expressed when grown on xylose or the sugar mix compared to growth on cellobiose alone (Table A.4.4). No ATPase subunit is found associated with the annotated ABC-type transporter proteins and complex formation may therefore rely on an independently transcribed ATPase subunit.

A second xylose specific ABC-transporter (TthWC1_1216-1219), which includes an ATPase subunit, was not observed in the proteomes. An ABC-transporter (TthWC1_0997-1000), annotated to be involved with ribose transport, is also encoded near the xylose responsive gene cluster (TthWC1_1005-1023). Hemme *et al*. (2011) suggest that, in strains lacking obvious xylose specific ABC-transporters, such as *Thermoanaerobacter pseudethanolicus* 39E, this orthologous ribose transporter may have xylose transport capabilities. However, it was also identified that this gene cluster was not up-regulated in response to growth on xylose by *T. pseudethanolicus* 39E. In *T. thermohydrosulfuricus* WC1, only two of the four annotated subunits were observed in the proteomes, and only expression of one subunit, the periplasmic component (TthWC1_0997), strongly correlated with the presence of xylose.

5.4.7 Central carbon metabolism

Xylose catabolism in *T. thermohydrosulfuricus* WC1 likely proceeds through the non-oxidative pentose phosphate pathway (Fig. 5.4) as no phosphoketolase enzyme, present in some *Firmicutes* (Tanaka *et al*., 2002; Liu *et al*., 2012), is identified within the

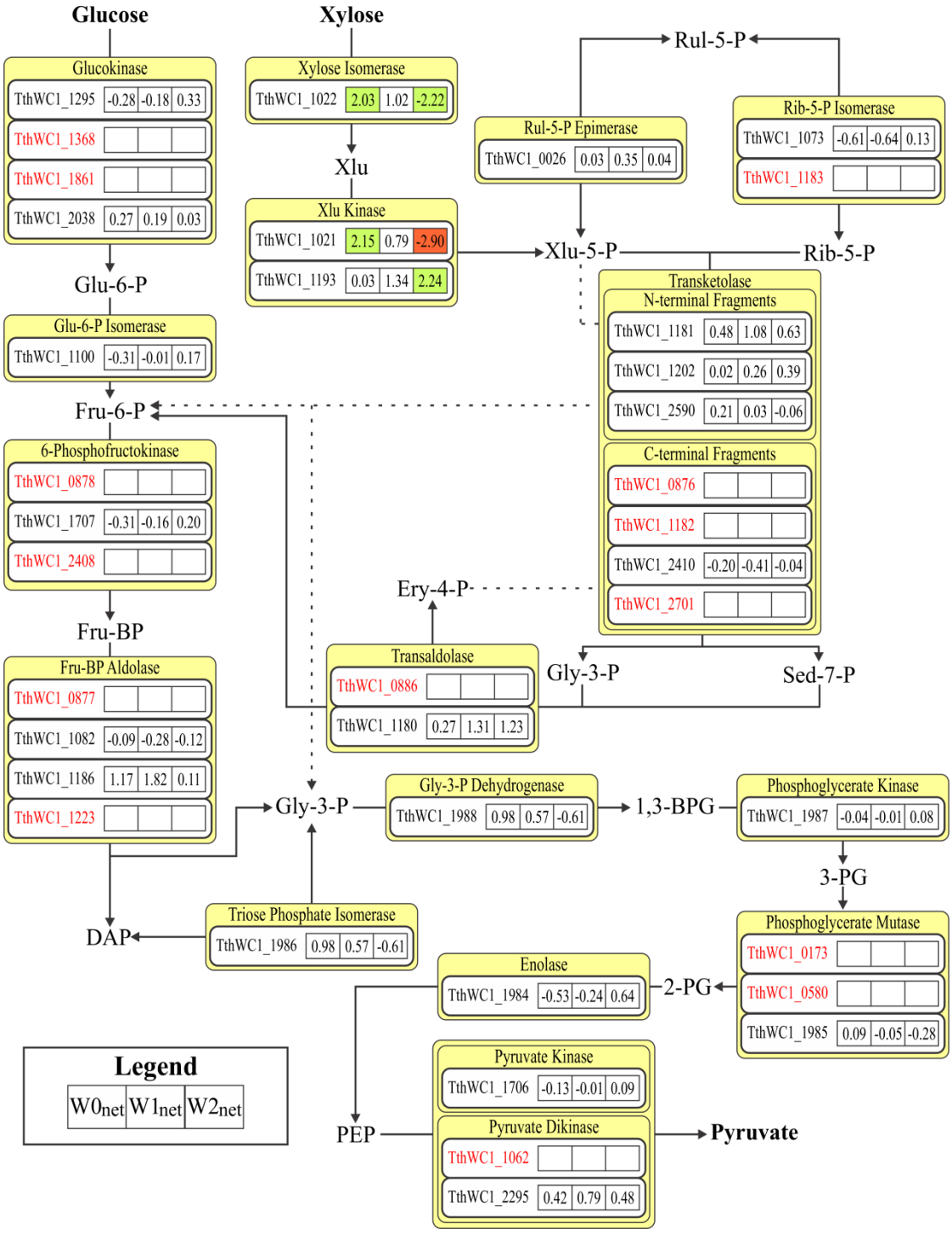


Fig. 5.4. Normalized relative Z-score expression ratios (W_{net}) of the glycolytic and pentose phosphate pathways in *T. thermohydrosulfuricus* WC1. Dashed lines represent the second transformation catalyzed by transketolase enzymes as part of the pentose

Fig. 5.4 cont.

phosphate pathway. Colour scheme as identified in Fig. 5.3. $W0_{net}$, $W1_{net}$ and $W2_{net}$ are as defined (Chapter 5.3.4.4). Abbreviations: Glu-6-P = glucose-6-phosphate; Fru-6-P = fructose-6-phosphate; Fru-BP = fructose-1,6-bisphosphate; Xlu = xylulose; Xlu-5-P = xylulose-5-phosphate; Rul-5-P = ribulose-5-phosphate; Rib-5-P = ribose-5-phosphate; Sed-7-P = sedoheptulose-7-phosphate; Ery-4-P = erythrose-4-P; DAP = dihydroxy acetone phosphate; Gly-3-P = glyceraldehyde-3-phosphate; 1,3-BPG = 1,3-bisphosphoglycerate; 3-PG = 3-phosphoglycerate; 2-PG = 2-phosphoglycerate; PEP = phosphoenolpyruvate.

genome. While the xylose isomerase and xylulose kinase are up-regulated under xylose or sugar mix conditions, a second annotated xylulose kinase (TthWC1_1193) shows the highest expression under cellobiose conditions (Table A.4.4), making it unlikely to be significantly involved in xylose catabolism. Relatively uniform expression across growth conditions was observed for genes involved in the conversion of xylulose-5-P to ribose-5-P (Fig. 5.4).

Identification of the transketolase encoding gene(s), which would catalyze the formation of sedoheptulose-7-P and glyceraldehyde-3-P from ribose-5-P and xylulose-5-P, remains ambiguous in *T. thermohydrosulfuricus* WC1. Functionally characterized transketolase enzymes are homodimers (Sprenger, 1993; Park *et al*., 2006). These thiamine diphosphate-dependent enzymes have an N-terminal diphosphate binding domain, a central-pyrimidine binding domain and a C-terminal domain of unknown function (Kochetov & Sevostyanova, 2005). In *T. thermohydrosulfuricus* WC1, as in all sequenced *Thermoanaerobacter* strains (Fig. A.4.7), the N-terminal domain is annotated to be encoded on a separate CDS from the protein encoding the middle- and C-terminal domains. Similar observations have also been reported for other sequenced *Thermoanaerobacterales* (van de Werken *et al*., 2008). Of the only co-localized transketolase genes in the *T. thermohydrosulfuricus* WC1 genome, only the N-terminal domain encoding protein (TthWC1_1181) was observed in the proteomes (Fig. 5.4), while the adjacent protein (TthWC1_1182) (encoding the middle and C-terminal domains) was not. Transketolase activity has been reported in the related strain, *Thermoanaerobacter wiegelii* Rt8.B1 (Cook *et al*., 1994b), and is also expected in *T.*

thermohydrosulfuricus WC1 as part of xylose catabolism, but the present data are unable to resolve which gene(s) may encode a functional transketolase in this strain.

Two divergent (24% amino acid similarity; 178 residue difference in length) transaldolase genes are annotated, though only one was observed in the proteomes. The transaldolase observed in the proteomes is only one of two proteins observed in the transaldolase (TthWC1_1180), transketolase (TthWC1_1181-1182) and ribose-5-P isomerase (TthWC1_1183) gene cluster predicted to be involved with xylose catabolism (Fig. 5.4).

Similar expression between growth conditions was also observed for proteins found in the glycolytic pathway (Fig. 5.4). The proteins corresponding to the five gene cluster responsible for the conversion of dihydroxyacetone phosphate (DAP) to phosphoenolpyruvate (PEP) were all expressed, though average nTIC $\log_2$ difference ≥3.97 was observed between the most abundant (TthWC1_1988; glyceraldehyde-3-phosphate dehydrogenase) and least abundant (TthWC1_1985; phosphoglycerate mutase) proteins under each condition (Table A.4.4).

Only one 6-phosphofructokinase, proposed to be ATP-, rather than pyrophosphate (PPi)-dependent (Chapter 4, Bapteste *et al*., 2003), was observed (Fig. 5.4). This suggests that *T. thermohydrosulfuricus* WC1 does not use PPi as an alternative energy currency to ATP during exponential growth, as has been described in other thermophilic *Firmicutes* (Bielen *et al*., 2010). Expression analysis of other proteins though suggests PPi may play a role during growth. Specifically, pyruvate phosphate dikinase, as an alternative to pyruvate kinase, may catalyze the conversion of phosphoenolpyruvate to pyruvate as part of glycolysis using AMP + PPi. While both the kinase and the dikinase

were observed in the proteome, the dikinase (TthWC1_2295) was in the top thirty most abundant enzymes and had an average nTIC $\log_2$ difference ≥1.96 than the pyruvate kinase across all growth conditions (Table A.4.4). Additionally, the membrane bound V-type pyrophosphatase (TthWC1_1224), which conserves energy through the hydrolysis of PPi, was more abundant than the observed cytosolic pyrophosphatase (TthWC1_0544) by an nTIC average $\log_2$ difference ≥0.92.

5.4.8 End-product and co-factor metabolism

Lactate was the major end-product observed (Fig. 5.2C) and is most likely catalyzed by a highly expressed lactate dehydrogenase (*ldh*) (TthWC1_1229) (Fig. 5.5A, Table A.4.4). Alternative to lactate production, pyruvate may also be decarboxylated to acetyl-CoA yielding CO_2 and reduced ferredoxin (Fd_{red}) via pyruvate ferredoxin oxidoreductase (POR). Two multi-subunit, as well as two single subunit PORs homologous to the characterized single subunit POR in *Moorella thermoacetica* (Furdui & Ragsdale, 2002; Pierce *et al*., 2008), are identified in the genome. Synthesis of all subunits forming the putative multi-subunit PORs was not observed, while both the single subunit PORs were observed. Expression analysis identified that TthWC1_1927 was observed at an average nTIC $\log_2$ difference ≥3.47 higher than TthWC1_1529 in each condition (Table A.4.4) and is likely the primary POR utilized.

Acetate production may proceed primarily via the co-localized phosphotransacetylase and acetate kinase genes (Fig. 5.5A). Gene products annotated as phosphotransbutyrylase (TthWC1_2222; TthWC1_2230) and butyrate kinase (TthWC1_2228; TthWC1_2231), whose specificity is not known and which may also

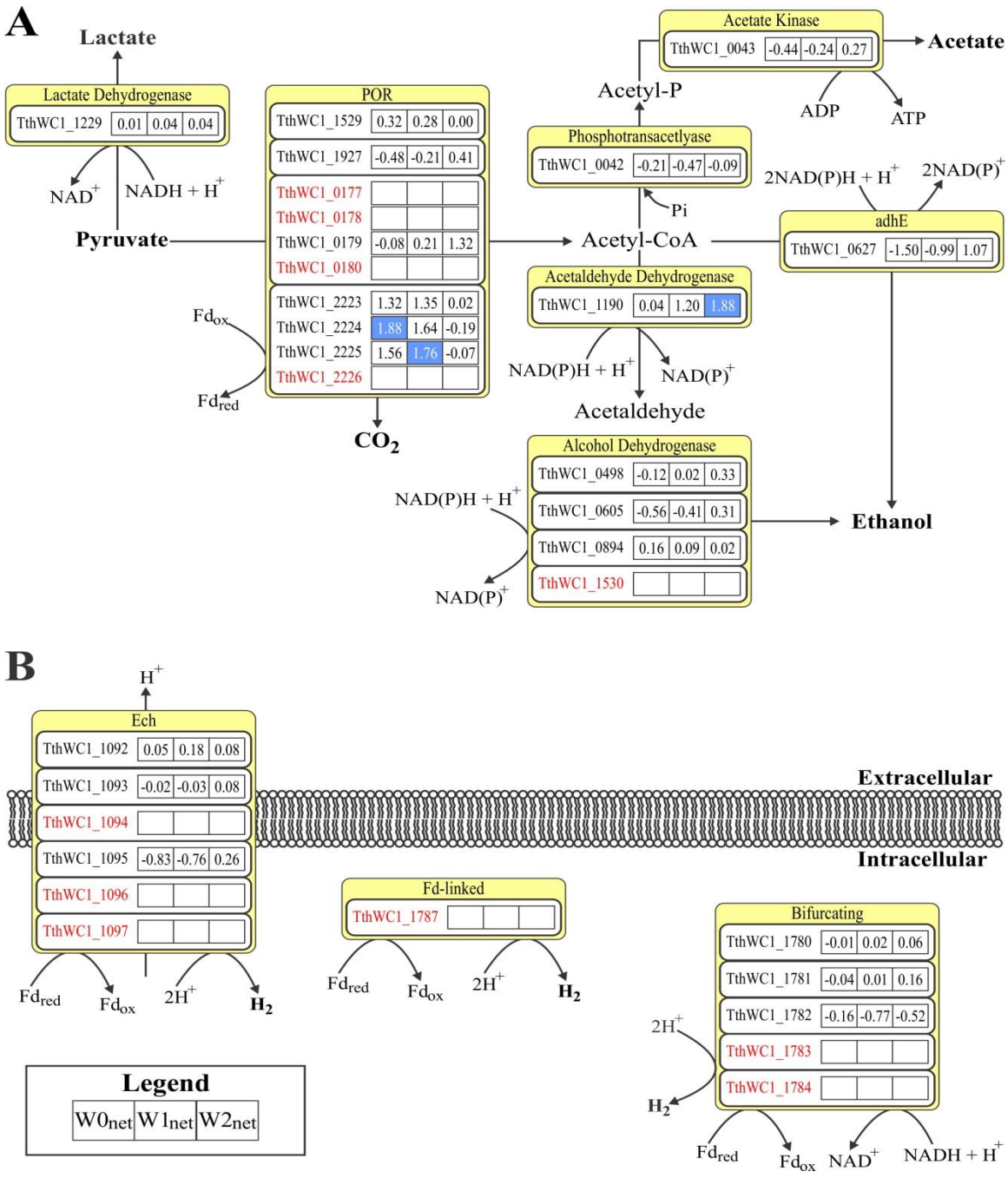


Fig. 5.5. Normalized relative Z-score expression ratios (W_{net}) of the end-product producing reactions in *T. thermohydrosulfuricus* WC1. (A) Acetate, ethanol, lactate and CO_2 producing reactions. (B) Hydrogenases. Colour scheme as identified in Fig. 5.3. $W0_{net}$, $W1_{net}$ and $W2_{net}$ are as defined (Chapter 5.3.4.4). Abbreviations: Acetyl-P = acetyl-phosphate.

contribute to acetate production, were also observed in the proteomes, but at much lower levels (Table A.4.4).

Ethanol production from acetyl-CoA in *T. thermohydrosulfuricus* WC1 may occur directly through a bifunctional acetaldehyde: alcohol dehydrogenase (AdhE) or via a standalone aldehyde dehydrogenase and independent alcohol dehydrogenase (ADH) (Fig. 5.5A). The standalone aldehyde dehydrogenase (TthWC1_1190), which exists in only a few *Thermoanaerobacter* spp. (Chapter 4), was observed in the proteome, but was not highly expressed (Table A.4.4). AdhE, on the other hand, was found in high abundance in all proteomes and may represent the principal route through which acetaldehyde is formed. Additionally, the protein (TthWC1_0498) encoded by the gene orthologous to *adhA* in *Thermoanaerobacter ethanolicus* JW200 (Bryant *et al*., 1988; Holt *et al*., 2000) was found in even higher abundance and was in the top five most abundant proteins across all growth conditions. Two other ADH proteins were also observed in the proteomes, while a final annotated ADH encoding gene (TthWC1_1530) was not.

Three distinct hydrogenases, including an energy-conserving hydrogenase (Ech), a putative bifurcating hydrogenase, and a Fd-linked hydrogenase, are present in *T. thermohydrosulfuricus* WC1 (Fig. 5.5B). The first, the Ech-hydrogenase, had only three of the six annotated subunits observed in the proteomes. While it is possible that the remaining three proteins are synthesized and were just not observed, only a single protein, TthWC1_1090, encoded within the adjacent *hypA-hypF* gene cluster, known to be responsible for hydrogenase maturation and the assembly of the NiFe centre in NiFe hydrogenases (Blokesch *et al*., 2002), was also observed (Table A.4.4) suggesting low, if any, expression of Ech.

Only three subunits composing a potential bifurcating hydrogenase were identified in the proteomes, though ambiguity exists in terms of what subunits are needed to form a functional complex. The gene cluster is orthologous to those found in other *Thermoanaerobacter* spp. (Chapter 4) including those previously proposed to form a bifurcating hydrogenase (Schut & Adams, 2009) with the exception that TthWC1_1784 is annotated as a pseudogene. Additionally, TthWC1_1783 is annotated as a histidine kinase and its role is unknown. Average nTIC analysis identifies that the TthWC1_1780-1781 cluster is highly expressed (top 35 most abundant proteins) across all conditions, though the ferredoxin encoding gene (TthWC1_1782) is not as highly expressed (Table A.4.4). A putative single subunit, Fd-linked hydrogenase (TthWC1_1787) was also not observed in the proteomes.

The *adh* gene products in some *Thermoanaerobacter* spp. have been shown to utilize both NADH and/or NADPH as cofactors (Bryant *et al*., 1988; Burdette *et al*., 1996; Holt *et al*., 2000). While NADH is typically considered a by-product of glycolysis, NADPH production in *Thermoanaerobacter* spp. is less well understood. *T. thermohydrosulfuricus* WC1, like other strains of *Thermoanaerobacter* (Lamed & Zeikus, 1980b; Feng *et al*., 2009), has no apparent 6-P-gluconolactonase potentially preventing NADPH production through the oxidative portion of the pentose phosphate pathway. The genome does encode an annotated NADP-specific isocitrate dehydrogenase (TthWC1_0990) though, which is considered NAD(P)-specific (Doležal *et al*., 2004), and is also observed in the proteomes (Table A.4.4).

Additionally, a two gene cluster orthologous to the *nfnAB* genes in *Clostridium kluyveri* (Wang *et al*., 2010) and *M. thermoacetica* (Huang *et al*., 2012), which couples

the reduction of $NADP^+$ via Fd_{red} with the reduction of $NADP^+$ via NADH, is annotated (TthWC1_0606-0607). Only one protein (TthWC1_0606) was observed in the proteomes.

5.4.9 Energy conservation

T. thermohydrosulfuricus WC1 has many mechanisms to conserve energy as an ion-motive force or as ATP/PPi (Chapter 4.4.7). Two potential ion-gradient generating reactions, catalyzed by Ech or a H^+-translocating V-type pyrophosphatase, have been discussed (above) with only the synthesis of the V-type pyrophosphatase confirmed. No gene products in the ion-translocating *mbx* gene cluster (TthWC1_0717-TthWC1_0729) were observed, while the gene product homologous to *nhaC*, encoding a potential H^+/Na^+ antiporter was also not observed. Both products of the two gene cluster (TthWC1_1955-1956) homologous to the *natAB* ATPase in *B. subtilis* (Cheng *et al*., 1997) were observed (Table A.4.4), while all but subunits "a" (TthWC1_0477) and "c" (TthWC1_0476) of the F_0F_1-ATPase were also observed. An annotated ATP-linked (EC: 4.1.1.49) (Aich & Delbaere, 2007) PEP carboxykinase (TthWC1_1173) was also not observed.

The genes products associated with ion-motive force or ATP/PPi generation were not the same products in COG class C found to be up-regulated under xylose only conditions (Table 5.2; Table A.4.4). Of note, in these up-regulated class C genes is a gene cluster (TthWC1_2220-2231), which encodes gene products putatively associated with amino acid metabolism including an annotated indolepyruvate ferredoxin oxidoreductase, a multi-subunit pyruvate/2-ketoisovalerate ferredoxin oxidoreductase and a leucine dehydrogenase.

5.5 Discussion

A lack of CCR allowing for the simultaneous utilization of lignocellulose relevant saccharides is a desirable phenotype for CBP microorganisms. Despite encoding genes for a potential CCR network, *T. thermohydrosulfuricus* WC1 was able to utilize both xylose and cellobiose, two important lignocellulose constituent saccharides, simultaneously, suggesting a lack of CCR with respect to this combination of substrates.

Independent of observing sugar utilization preferences, understanding the nature of substrate-linked gene regulation is important to improving our understanding of *T. thermohydrosulfuricus* WC1 physiology. The HOV analysis presented identified that growth on xylose alone lead to an up-regulation in genes involved with energy production (COG C) consistent with reports for *Thermoanaerobacter* sp. X514 (Lin *et al*., 2011). Despite this, growth of *T. thermohydrosulfuricus* WC1 under xylose only conditions corresponded to the slowest growth rate and the presence of xylose seemed to even slow growth under the mixed substrate conditions (above). These differences may be partly attributable to differences in potential energy yield from both the transport and hydrolysis of each substrate.

The proteomic results suggest xylose transport is principally through ABC-type mechanisms as has been reported in other *Thermoanaerobacter* spp. (Erbeznik *et al*., 1998b; Jones *et al*., 2002b; Erbeznik *et al*., 2004; Lin *et al*., 2011) and would consume one mole of ATP per mole of xylose transported. Conversely, transport of the disaccharide cellobiose seemingly occurs via PTS-mediated mechanisms. Its transport would also consume one mole ATP equivalent/mole cellobiose. In this case, PEP would phosphorylate the transported cellobiose molecule, rather than ADP (forming ATP) as it

would in the conversion of PEP to pyruvate. Thus, the energy expenditure is equivalent for transporting one mole of either substrate.

However, assuming PTS-mediated transport of cellobiose occurs, its hydrolysis would yield 1 glucose plus 1 glucose-6-P (Fig. 5.3) and the theoretical ATP yield from converting 1 mole of cellobiose to pyruvate would be 5 moles ATP. The conversion of xylose to pyruvate through the pentose phosphate pathway would only yield 1.67 moles ATP/mole xylose providing less energy for growth. If true for *T. thermohydrosulfuricus* WC1, these differences could potentially account for the differences in the observed doubling times. This is further supported by the observed differences in substrate consumption rates (Table 5.1). While xylose was consumed at a faster rate than cellobiose under both the single substrate and the mixed sugar conditions, the calculated rate was not sufficiently fast enough to potentially mitigate differences in the energetic potential from each substrate. Similar findings were observed by Lacis and Lawford (1991) whom report lower ATP yields for *Thermoanaerobacter ethanolicus* grown on xylose in contrast to glucose, which correlated with slower growth of *T. ethanolicus* on xylose under batch conditions.

To compensate for growth on a poorer energy source such as xylose, *T. ethanolicus* simultaneously consumed yeast extract as an additional energy source during growth, but would not do so during growth on glucose (Lacis & Lawford, 1991). Like many other *Thermoanaerobacter* spp. (Faudon *et al*., 1995), *T. thermohydrosulfuricus* WC1 is capable of growing on yeast extract (Chapter 2), though the reduced concentrations of yeast extract and absence of thiosulfate (Faudon *et al*., 1995; Fardeau *et al*., 1997) in the medium used makes it unlikely to act as a major carbon or energy source

here (Fig. 5.1). It is plausible however, that the genes up-regulated during growth on xylose (COG class C), notably the indolepyruvate ferredoxin oxidoreductase, 2-ketoisovalerate ferredoxin oxidoreductase and leucine dehydrogenase, which are known to be involved with peptide fermentation (Mai & Adams, 1994; Heider *et al*., 1996), may be a mechanism by *T. thermohydrosulfuricus* WC1 to maximize energy when grown on poorer energy sources. It should be noted though, that the specificity of these gene products is not yet known and in some cases, such as with the annotated indolepyruvate ferredoxin oxidoreductase or the 2-ketoisovalerate ferredoxin oxidoreductase, the gene products may have POR activity and not be involved with amino acid fermentation.

T. thermohydrosulfuricus WC1 produced comparable amounts of end-products across growth conditions with lactate being the major end-product formed (Fig. 5.2). This was surprising given that previous experiments with *T. thermohydrosulfuricus* WC1 using lower substrate and higher yeast extract concentrations (2 g/L cellobiose, 2 g/L yeast extract) in the same medium found that acetate and ethanol were produced in comparable amounts with lactate being a minor end-product (Chapter 2). Similar observations have been reported with *T. ethanolicus* JW200 as increasing glucose concentrations have led to a similar disproportionate increase in lactate production (Lacis & Lawford, 1991).

Given the high expression of all end-product forming enzymes (Table A.4.4) and that *adh* gene products, in some cases, were even more abundant than the LDH, the formation of lactate as the major end-product does not correlate with a lack of expression of the other gene products. Therefore the basis for lactate formation may be related to metabolic flux or metabolite-based activity control (allosteric regulation).

In *Thermoanaerobacter brockii* subsp. *brockii* HTD4, high lactate formation has been associated with high intracellular concentrations of fructose-1,6-bisphosphate (FBP) (Ben-Bassat *et al*., 1981), while the LDH enzymes of *T. pseudethanolicus* 39E (Lovitt *et al*., 1988) as well as *T. ethanolicus* JW200 (Zhou & Shao, 2010) have been shown to be allosterically activated by FBP. If also true in *T. thermohydrosulfuricus* WC1, intracellular FBP levels must always be sufficiently high to activate lactate formation under the conditions tested as lactate was produced throughout all stages of growth. The disproportionate increase in lactate production by mid-exponential phase (Fig. 5.2) may represent an excess accumulation of FBP at this stage of growth potentially caused by a bottleneck in its consumption during glycolysis by FBP aldolase, allowing for increased activation of LDH.

Alternatively, high lactate production may occur through a limitation in the ability of *T. thermohydrosulfuricus* WC1 to produce alternative end-products. One possible bottleneck could be in the decarboxylation of pyruvate through POR, as has been reported in other saccharolytic *Clostridia* (Guedon *et al*., 1999; Bahl *et al*., 1986), leading to the disposal of carbon, and reducing equivalents in the form of NADH, by producing lactate. Slow consumption of acetyl-CoA generated from POR mediated catalysis could also exist. This would require that the phosphotransacetylase and the enzyme(s) catalyzing acetaldehyde formation have a much lower affinity for acetyl-CoA than LDH does for pyruvate, or function at a much slower catalytic rate.

Ethanol production may also be limited by a lack of suitable reducing equivalents. In *T. ethanolicus* JW200, the LDH enzyme has been characterized as being NADH-dependent and no activity was observed when NADPH was used as an alternative co-

factor (Zhou & Shao, 2010). Conversely, the AdhA protein in *T. ethanolicus* JW200, whose ortholog in *T. thermohydrosulfuricus* WC1 was the most abundantly expressed *adh* gene product (Table A.4.4), is considered to be primarily NADPH-dependent (Bryant *et al*., 1988; Holt *et al*., 2000). If similar specificities are also present in *T. thermohydrosulfuricus* WC1, and assuming that glycolysis principally generates NADH rather than NADPH, ethanol formation via AdhA may be limited by a lack of available NADPH. While the *adhE* gene was also highly expressed in *T. thermohydrosulfuricus* WC1, the study by Pei *et al*. (2011) with *T. ethanolicus* JW200 has suggested that AdhE may be principally involved with acetaldehyde, rather than alcohol, formation.

Based on flux analyses conducted in other *Thermoanaerobacter* spp. (Hemme *et al*., 2011), carbon flux through an NADP-specific isocitrate dehydrogenase seems unlikely to provide significant levels of NADPH for ethanol production via AdhA. As the *mbx* gene cluster in *T. thermohydrosulfuricus* WC1 (Chapter 4), whose NAD/NADP specificity is not known, was not observed in the proteomes (Table A.4.4), it also unlikely contributes significantly to cellular NADPH levels. Therefore, the only obvious mechanism for NADPH production may be the *nfnAB* gene cluster, though not all subunits were observed in the proteome. Thus, the mechanisms of NADPH synthesis in *T. thermohydrosulfuricus* WC1 are not yet understood.

5.6 Conclusions

The present study has identified that *T. thermohydrosulfuricus* WC1 has the ability to co-ferment cellobiose and xylose with no overt repression in the utilization of either sugar. Additionally, proteomic analyses of the strain grown on single and multiple

substrates has been conducted. However, we have not been able to fully resolve pathways of carbon flux, as enzyme specificity is currently limited to insights provided by genome annotation and detailed knowledge of enzyme activity has not yet been conducted. Further, the current proteomic analysis is limited to proteins observed with high confidence through MS/MS analyses and does not reflect the entirety of the gene products synthesized by the bacterium. A major finding of the analysis though, is that *T. thermohydrosulfuricus* WC1 end-product synthesis patterns seem to be more connected to metabolic flux or metabolic regulation, rather than to a lack of gene expression for ethanol forming enzymes.

The low ethanol yields observed in *T. thermohydrosulfuricus* WC1 is a significant issue in its development as a CBP-relevant, biofuel producing organism. Further refinement of metabolic flux may help identify mechanisms and strategies to potentially eliminate bottlenecks currently limiting ethanol production. The current work can help serve as a baseline for such studies and, as it represents the first known proteomic investigation into a *Thermoanaerobacter* strain, may help shed insight into the physiology of other strains within the genus.

5.7 Authors' contributions

Experiments were conceived and designed by Tobin J. Verbeke, Oleg V. Krokhin, Brian Fristensky, John A. Wilkins, David B. Levin and Richard Sparling. Growth studies, metabolic analyses and proteomic sample prep was conducted by Tobin J. Verbeke. Custom perl scripts were developed by Xiangli Zhang. Mass spec analyses was conducted by Oleg V. Krokhin and processing of the raw proteomic data was done

by Vic Spicer. Analyses and calculations of the proteomic data was done by Tobin J. Verbeke. The manuscript, upon which this chapter is based, was authored by Tobin J. Verbeke with input from Vic Spicer, Oleg V. Krokhin, John J. Schellenberg, John A. Wilkins, David B. Levin and Richard Sparling.

5.8 Acknowledgements

This work was supported by funds provided by Genome Canada for the project titled "Microbial Genomics for Biofuels and Co-Products from Biorefining Processes ($MGCB^2$)", the Province of Manitoba through funds provided by the Department of Innovation, Energy, and Mines "Manitoba Research Innovation Fund" (MRIF), the Natural Sciences and Engineering Research Council Strategic Grant (STPGP 365076), and the University of Manitoba.

We also thank Dmitry Shamshurin and Kirill Levin at the Manitoba Centre for Proteomics and Systems Biology for technical assistance and guidance in preparation of samples for MS/MS analysis.

Chapter 6. Thesis conclusions and future perspectives

6.1 Introduction

Developing second generation biofuel technologies into a viable industry is desirable for environmental sustainability and long-term energy security. However, its progression into industry is currently fraught with numerous challenges, not the least of which includes the selection and development of suitable platform microorganisms. This aspect is particularly important in CBP platforms, which requires that selected strains be proficient in numerous physiological processes related to lignocellulose conversion and biofuel production. For rapid development into industry, selecting innately capable strains to use in CBP is inherently logical, yet the potential of this approach has not yet been realized due to physiological limitations associated with currently investigated microbes.

Addressing such limitations through a combined applied ecological/physiological approach has been the goal of this thesis. As shown throughout, the value of bioprospecting lies not only in its potential for novel organism discovery, but it also requires that current developing platform strains be re-evaluated. Furthermore, as novel isolates are unlikely to continually possess CBP-relevant phenotypes that are "advantageous" or "exceed" the desirable qualities of current investigated strains, the value of exploring the diversity of biofuel producing microbes may lie in the physiological insights gained. Advancements in the understanding of diverse physiologies may continue the development of current platform strains potentially leading to industry-ready microbes.

6.2 Thesis findings and conclusions

6.2.1 Strain isolation, identification and taxonomic implications within the genus *Thermoanaerobacter*

Efforts to explore the microbial diversity present in decaying woodchip compost identified a collection of isolates belonging to only two genera, *Thermoanaerobacter* and *Thermoanaerobacterium* (Chapter 2). While it is unlikely that these isolates encompass the entire breadth of diversity present in the compost, previous studies have identified the CBP-relevance of strains in these genera (Lynd *et al*., 2008; Shaw *et al*., 2010; Hemme *et al*., 2011) and thus provided a rationale for their continued investigation.

Assigning taxonomic identities to the (now recognized) twelve *Thermoanaerobacter* isolates, for sole purposes of identifying close phylogenetic relatives, proved to be an initial challenge. Inferring phylogenies using conventional 16S rRNA sequence analyses was complicated by the presence of IVS in numerous 16S rRNA encoding sequences within the genus (Chapter 2). As IVS contribute to unusually high inter-operonic *rrn* variability within *Thermoanaerobacter* spp. and in related strains (Acinas *et al*., 2004b; Chapter 2), the reliability of 16S as a phylogenetic signature gene for *Thermoanaerobacter* spp. became questionable.

MLST approaches, most notably a 3-gene model developed by Zeigler (2003), which had been developed based on recommendations to find alternatives to DNA-DNA hybridization studies (Stackebrandt *et al*., 2002), offered potential as an alternative to 16S-based typing for evaluation of the novel isolates. Concurrent with the application of Zeigler's 3-gene model to the isolates, a single gene-model for predicting genomic

relatedness, based on the *cpn60* UT gene sequence, was also developed (Chapter 3). This 1-gene model was built using diverse bacterial lineages consistent with Zeigler (2003), and was further validated using the recognized "gold-standard" for delineating bacterial species (Richter & Rosselló-Móra, 2009). It was shown to offer a technically simple and reliable approach for rapidly inferring phylogenetic relatedness amongst diverse microorganisms. Since its development, the model itself, or justification for the use of *cpn60* based on this model, have been effectively employed to differentiate between strains of *Brachyspira* (Rubin *et al*., 2013), *Pseudomonas putida* (Sharma *et al*., 2012), *Mesorhizobium* (Moukoumi *et al*., 2013) and *Gardnerella vaginalis* (Jayaprakash *et al*., 2012).

Application of both the 3-gene model and the 1-gene model to isolates WC1-WC12 supported the assignment of the isolates to the species, *Thermoanaerobacter thermohydrosulfuricus* (Chapter 3). Further, both models suggested that the 12 novel isolates, as well as the 11 reference *Thermoanaerobacter* spp. used, may actually comprise only three independent species based on a species delineation cut-off of 70% predicted shared genome sequence identity. The coherence of three distinct species/clades was further supported by comparisons of the whole-genome functional annotation profiles observed amongst sequenced *Thermoanaerobacter* spp. (Fig. 4.2; Fig. A.3.1; Fig. A.3.2).

However, as the intent of this thesis was never to provide a comprehensive taxonomic classification of the *Thermoanaerobacter* genus, formal proposals for the reclassification of species within the genus were not pursued. This decision was additionally justified by the realization that numerous reclassifications of multiple *Thermoanaerobacter* spp. (Lee

et al., 1993; Rainey *et al*., 1993; Collins *et al*., 1994; Cayol *et al*., 1995; Fardeau *et al*., 2004; Onyenwoke *et al*., 2007) have made deciphering the related scientific literature a challenge (Appendix A.8). Further, as significant physiological differences can be identified amongst subspecies of a single lineage (Cayol *et al*., 1995; Chapter 2), maintaining current taxonomic designations, rather than combining multiple strains into a single species, is preferable for the continued delineation of *Thermoanaerobacter* strains.

6.2.2 The CBP-potential of the genus *Thermoanaerobacter*

The evaluation of the CBP-relevant physiology of *Thermoanaerobacter* spp. presented in this thesis has largely been limited to only a few processes (e.g. biomass hydrolysis, substrate utilization, biofuel production) overtly related to second-generation biofuel production (see Chapter 1.3; Fig. 4.1). The basis for these limitations largely stems from the fact that the value in understanding these processes are well established, yet the actual physiology within the genus is poorly understood and can vary widely. Further, differences in these physiological processes are reported to have practical implications in the implementation of certain *Thermoanaerobacter* strains into designed CBP-platforms (Hemme *et al*., 2011).

For example, none of the novel isolates were cellulolytic (Chapter 2), as is consistent with reports for most strains within the genus (Wiegel, 2009). Therefore, the use of most *Thermoanaerobacter* spp. in CBP-technologies would ultimately require external means of lignocellulose hydrolysis (e.g. addition of free enzymes) or more practically, through co-culturing with a cellulolytic thermophile. As recent co-culturing efforts have shown (Kridelbaugh *et al*., 2013), consortia whereby both members possess

hydrolytic capabilities (cellulolytic and/or hemicellulolytic) are desirable for CBP development. The analyses presented in Chapter 4 has identified that hydrolytic capabilities are not evenly dispersed throughout the *Thermoanaerobacter* genus. In fact, only three strains, *T. italicus* Ab9 (Clade 2[15]), *T. mathranii* susbp. *mathranii* A3 (Clade 2) and a novel isolate arising from work done in this thesis, *T. thermohydrosulfuricus* WC1 (Clade 3), from the *Thermoanaerobacter* spp. examined have the ability to hydrolyze and grow on polymeric xylan. This ability was proposed to be attributable to the presence of extracellular GH10 endo-xylanases in the genomes of these strains (Chapter 4; Appendix A.5).

The ability to ferment diverse saccharides, another desirable CBP-relevant phenotype, also varies widely amongst genus members (Chapters, 2, 4), but has often provided part of the rationale for investigating the CBP-potential of *Thermoanaerobacter* spp. (Taylor *et al*., 2009; Hemme *et al*., 2011; Lin *et al*., 2013). As the most diverse potential for substrate utilization is also identified in *T. italicus* Ab9 and *T. mathranii* susbp. *mathranii* A3 (Chapter 4), as well as *T. thermohydrosulfuricus* WC1 (Chapters 2, 4), this once again distinguishes these strains as potentially being "advantageous" for CBP-biofuel production over other *Thermoanaerobacter* spp. evaluated. However, despite possessing diverse substrate utilization potential, the simultaneous utilization of lignocellulose relevant saccharides has only been reported in Clade 1 strains (Lin *et al*., 2011) or Clade 3 strains (Cook & Morgan, 1994a; Chapter 5). While the simultaneous utilization of lignocellulose relevant saccharides has been limited to combinations of xylose plus glucose or xylose plus cellobiose only, this is still a highly desirable

[15] See Fig. 4.2 for Clade designations of select strains

characteristic of these strains given the relative abundance of these saccharides in many types of lignocellulosic biomass (Carroll & Somerville, 2009).

Finally, the highest yields of biofuels, which, for *Thermoanaerobacter* spp. has principally focused on ethanol production (Ben-Bassat *et al*., 1981; Lovitt *et al*., 1988; Burdette *et al*., 1996; He *et al*., 2009; Lin *et al*., 2011), are found in Clade 1 strains (Chapter 4). The novel *Thermoanaerobacter* isolates identified (WC1-WC12) have lower molar ethanol yields than both Clade 1 or Clade 2 strains (Chapters 2, 4) under the conditions tested, and at elevated sugar concentrations *T. thermohydrosulfuricus* WC1 produce lactate as the major end-product (Chapter 5).

The analyses presented in this thesis in unable to identify a novel isolate or reference *Thermoanaerobacter* strain that is an obviously "advantageous" strain for development into a platform organism. Rather, in all of these strains evaluated, both desirable and undesirable CBP-relevant phenotypes are evident. Therefore, while efforts directed towards the continued development of any *Thermoanaerobacter* strain can build upon the data/analyses presented here, the choice of strain for development remains with individual researchers.

6.2.3 Core metabolism of *T. thermohydrosulfuricus* WC1

Given its diverse substrate utilization potential and its relative ethanol production capabilities, isolate *T. thermohydrosulfuricus* WC1 was selected from the twelve novel *Thermoanaerobacter* isolates (Chapter 2) for deeper investigation into its core metabolism. As the nature of the fermentative end-products in *T. thermohydrosulfuricus* WC1 is consistent with many other *Thermoanaerobacter* spp., yet the ratios can vary

significantly (Chapters 2, 4), understanding its core metabolism may help identify approaches to improve its ethanol production and ultimately, its industrial relevance. To investigate this, links between observed metabolism and genome content (Chapter 4), as well as proteomic profiling (Chapter 5) were established.

Genomic and proteomic analyses of *T. thermohydrosulfuricus* WC1 support the proposal that lignocellulose constituent hexose sugars or pentose saccharides are either consumed, or feed into, the EMP and pentose phosphate pathways, respectively. Further, relatively stable protein abundance levels were observed independent of growth substrate (Chapter 5). The same is also true for gene products involved in pyruvate catabolism; however, not all annotated genes whose products may be involved with pyruvate catabolism were observed. As such, the highly expressed gene products observed are therefore considered to be rational targets for genetic engineering strategies to manipulate carbon flux in this strain and improve biofuel yields.

However, knockdown/knockout strategies that target specific genes to improve biofuel (ethanol) yield in *T. thermohydrosulfuricus* WC1 may also have unintended physiological implications. For example, while lactate was the predominant end-product observed under the conditions tested, abundance levels of specific alcohol dehydrogenases (*adhA/adhE*) were also high suggesting end-product formation in *T. thermohydrosulfuricus* WC1 is more influenced by enzyme activity/regulation than it is by protein abundance (Chapter 5). Deletion of the *ldh* would therefore be an obvious target to reduce carbon flux towards lactate formation. However, as lactate production coincides with the consumption of reducing equivalents (NAD(P)H), deletion of the *ldh* may ultimately generate a redox imbalance. This balance may be restored though if

ethanol production and lactate production utilize the same reducing equivalent (NADH or NADPH). If true, this ultimately raises the question that if both lactate and ethanol formation are dependent on the same reducing equivalent, why is lactate principally formed under the conditions tested, and not ethanol?

The explanation for this may simply be differences in the native catalytic activity of the *T. thermohydrosulfuricus* WC1 LDH-enzyme in comparison to other strains. Specifically, the LDH of *T. thermohydrosulfuricus* WC1 may have a higher turnover number (high k_{cat}) or show increased binding affinity for pyruvate (low K_m) than the LDH enzymes in other strains. Or, if LDH is allosterically activated by fructose-1,6,-bisphosphate (FBP) in *T. thermohydrosulfuricus* WC1, as it is in other *Thermoanaerobacter* spp. (Ben-Bassat *et al*., 1981; Bryant, 1991; Zhou & Shao, 2010), the differences may be in the intracellular FBP levels. It is also important to consider that lactate formation is thermodynamically favourable when compared to ethanol production (Carere *et al*., 2012); yet, this is also true for all *Thermoanaerobacter* spp. and would not explain comparatively low lactate formation in those strains.

With these potential explanations in mind, deletion of *ldh* may lead to increased ethanol production under these same conditions. However, much of this discussion is predicated on the basis that lactate formation and ethanol production in *T. thermohydrosulfuricus* WC1 are reliant on the same reducing equivalent, or show low specificity for either NADH or NADPH. If this is not the case, the physiological implications associated with *ldh* deletion may have unintended and detrimental effects to cell viability, yet this remains to be determined.

The work presented in this thesis has helped to refine the routes of carbon and electron flux in *T. thermohydrosulfuricus* WC1, but has not been able to fully resolve them. Further, based on comparative genomic analyses (Chapter 4), significant conservation of potential pyruvate catabolism related genes are observed throughout the genus, and several of the observations noted in *T. thermohydrosulfuricus* WC1 (Chapter 5) may also help resolve routes of carbon flux routes in other strains.

6.2.4 General conclusions

The bioprospecting efforts presented were unable to identify a novel organism that was natively "industry-ready." Such expectations would be unrealistic though as the physiological demands of industrial strains rarely evolve naturally, particularly given the communal nature of microbes (Chapter 2). Rather, the work presented has developed whereby the discovery of novel isolates provided a rationale to continue the evaluation of, and in many cases re-visit, the genus *Thermoanaerobacter*. The relevance of *Thermoanaerobacter* spp. for second-generation biofuel production through CBP is widely accepted (Lynd *et al*., 2008; Taylor *et al*., 2009; Olson *et al*., 2012), yet numerous strains with great potential for CBP have received little attention since their initial isolation.

The progression of the work presented in this thesis has intended to, while remaining principally rooted in a single isolate, explore the CBP-relevant potential of diverse strains within the genus *Thermoanaerobacter*. The recognition of diverse lineages, or clades (Chapters 3, 4), has helped refine phylogenetic relatedness between strains and also identify the divergent, as well as the conserved, physiological potential

found amongst genus members. While this work has principally focused on CBP-relevant processes such as biomass hydrolysis, substrate utilization and biofuel production, the insights into these processes gained from evaluating diverse strains may not only help advance the development of *Thermoanaerobacter* spp. for biofuel-production, but potentially also be applicable to strains outside of the genus.

6.4 Future perspectives

From a biotechnological viewpoint, platform organism development remains one of the largest issues in achieving industrially viable lignocellulosic biofuels. This is evident by the fact that while numerous organisms show potential for development (Chapter 1), only a few, such as *Clostridium phytofermentans* (http://www.qteros.com) or *T. mathranii* BG1 (Mikkelsen & Ahring, 2007) (http://www.biogasol.com), have progressed past lab-scale experimentation.

Continuing the development of any *Thermoanaerobacter* sp. for CBP-biofuel production requires that numerous physiological limitations be resolved. Firstly, the lignocellulose hydrolytic abilities, or lack thereof, found throughout the genus significantly impede the implementation of any strain in mono-culture in a CBP system. Genetic engineering targeted towards improving the hydrolytic capabilities of a strain, such as has been attempted with *Thermoanaerobacter* sp. X514 (Lin *et al*., 2010), does offer a potential approach to improve upon this limitation. However, to engineer the hydrolytic capabilities of any *Thermoanaerobacter* sp. comparable to those innately found in other developing platform organisms (e.g. *C. thermocellum*) would require tremendous engineering efforts. While this aspect alone should not be viewed as a

limitation, this approach is largely counter-intuitive to the logic of selecting natively capable strains for development.

A more practical approach would be to co-culture a *Thermoanaerobacter* strain with a cellulolytic organism. While such co-cultures have previously shown to display desirable synergies (Ng *et al*., 1981; Mori *et al*., 1990; He *et al*., 2011; Hemme *et al*., 2011), a major challenges lies in the continued development of the co-culture itself. Specifically, the complexities of identifying and resolving physiological limitations in any of these strains are further compounded by a lack of understanding of the potential interactions that may exist between the co-culture's constituent members. Comparative expression profiling and examining culture/population dynamics may provide insights into these processes, allowing for physiological limitations to be identified, yet such efforts have not yet been reported in CBP-relevant co-cultures.

Improvement of overall biofuel yields to approach theoretical maxima in these strains remains an additional challenge. While manipulating growth conditions or fermentation broths can influence biofuel yields in some *Thermoanaerobacter* spp. (see Table 4.4), genetic engineering to redirect carbon flux in specific strains may offer a more permanent solution for strain development. Further, genetic engineering may afford flexibility in CBP-platform design for use with developed strains making them less susceptible to changes in environmental conditions. Despite the comparative increased amenability to genetic manipulation reported for *Thermoanaerobacter* spp. in comparison to other developing platform strains (Shaw *et al*., 2010), only a few attempts to engineer carbon flux in any strain of the genus have been reported (Mikkelsen & Ahring, 2007;

Yao & Mikkelsen, 2010). Genetic engineering is an underexplored area of research in *Thermoanaerobacter* spp.

However, as illustrated throughout Chapter 1, modulating central metabolism can have unintended physiological effects given the global connectivity of many metabolites. As proposed in Chapter 5, at least in the case of *T. thermohydrosulfuricus* WC1 under the conditions tested, the role of co-factors/metabolites may play an important role in determining end-product distribution patterns and must be taken into consideration when designing engineering efforts. In *Thermoanaerobacter* spp., the role of co-factors in ethanol synthesis (Bryant *et al*., 1988; Holt *et al*., 2000; Burdette & Zeikus, 1994; Pei *et al*., 2010b), and to a lesser extent lactate synthesis (Bryant, 1991; Zhou & Shao, 2010), have been previously investigated, yet global profiles have not yet emerged. Technologies such as metabolomics have not yet been applied to *Thermoanaerobacter* spp., yet insights gained from such strategies may significantly advance our understanding of the mechanism(s) controlling end-product distribution.

An important physiological aspect not yet discussed in this thesis is the tolerance of *Thermoanaerobacter* spp. to accumulating levels of their own fermentation end-products. The sensitivity of many ethanologenic strains to the toxic effects of ethanol is often viewed as a limitation for strain development (Taylor *et al*., 2009; Weber *et al*., 2010; Lin & Jian, 2013). Many thermophiles cannot grow at ethanol concentrations above 2%, while CBP-ethanol platforms require tolerance of concentrations exceeding 5% (Georgieva *et al*., 2007b). Tolerance levels exceeding 5% have been reported for *Thermoanaerobacter* spp. and have been attributed to random mutagenesis (Lovitt *et al*., 1988; Burdette *et al*., 2002), directed evolution of strains (Lin *et al*., 2013), or in some

cases, occur naturally (Georgieva *et al*., 2007a, b). While the insights provided by these studies are valuable, they often unfortunately focus on strains distinct from those having undergone significant physiological characterization. Continued development of many currently investigated *Thermoanaerobacter* spp. would additionally benefit from experiments of this type.

Of the numerous challenges facing the development of *Thermoanaerobacter* spp., as well as other potential platform organisms, for lignocellulosic biofuel production, one of the most significant may be the time needed for the development of an "industry-ready" strain relative to the demand for specific biofuels. The basis for this statement is founded in the argument that selecting microorganisms possessing innate CBP-relevant phenotypes is an inherently logical approach to develop second-generation biofuel platforms. As such, for most *Thermoanaerobacter* spp. the native fermentative pathways yield only two potential fuels, H_2 or ethanol. Given that other developing platform organisms natively produce much higher H_2 yields than *Thermoanaerobacter* spp. (Table 1.5), pursuing biological H_2 production using those microorganisms is likely advantageous. Therefore, the role of *Thermoanaerobacter* spp. for CBP-based biofuel production to meet the increasing fuel needs of the transportation sector is centred on the ethanologenesis capabilities of these strains.

Current government mandates by the United States and the European Union (Antizar-Ladislao *et al*., 2008; Hertel *et al*., 2010; Hsu *et al*., 2010; Roberts & Schlenker, 2010), and to a lesser extent the Canadian government (Mabee & Saddler, 2010; Li *et al*., 2012), to increase fuel ethanol production are a particularly powerful driving force behind research into lignocellulosic ethanol production; including research into

Thermoanaerobacter spp. Despite this, the long-term sustainability and viability of ethanol as an alternative fuel are uncertain (Hamelinck *et al*., 2005; Tyner & Viteria, 2010). Issues such as the low energy density of ethanol in comparison to gasoline and other fuels (Durre, 2007; Balat & Balat, 2009), limitations in the extent it can be blended with gasoline (Tyner & Viteria, 2010) and the incompatibility of ethanol with current transportation infrastructure due to its corrosive and hygroscopic nature (Lee *et al*., 2008) have already prompted efforts to attain alternative second-generation "drop-in[16]" biofuels such as butanol. The concurrent development of non-ethanol second-generation fuels, in conjunction with the expiration of current government mandates (e.g. in the year 2022 for current U.S. mandates; Hsu *et al*., 2010), may lead efforts to produce ethanol as a fuel additive to be reduced, if not completely abandoned, in the future. However, historically speaking, the demand for fuel ethanol has undergone numerous high- and low-periods (Rosillo-Calle & Walter, 2006; Solomon *et al*., 2007; Songstad *et al*., 2009) and the current pressures driving research into lignocellulosic ethanol production may not only persist, but could also potentially increase. Therefore, a potential also exists whereby lignocellulosic biofuel production using *Thermoanaerobacter* spp. will continue to be driven as it is currently.

Independent of the fuel type demanded, a biotechnological potential exists in *Thermoanaerobacter* spp. that has not yet been observed in many other thermophilic organisms of interest for CBP-biofuel production. As mentioned, the relative amenability to genetic manipulation of *Thermoanaerobacter* spp. in comparison to other thermophilic *Firmicutes* (Shaw *et al*., 2010), should be regarded as a major technologically

[16] Drop-in biofuels refer to alternative fuels that can be transported and consumed using existing infrastructure and technologies (for petroleum consumption) without the need for modification (Savage, 2009).

advantageous phenotype. This is particularly true given that the innate CBP-relevant capacities of any strain, no matter how "advantageous" in comparison to other organisms, most likely requires significant development prior to becoming "industry-ready." As such, a potential exists to engineer native or non-native metabolic pathways in a *Thermoanaerobacter* host strain leading to the production of non-ethanol biofuels or other valuable products.

Precedence for such an approach has been well established in mesophilic host organisms (Steen *et al*., 2008; Li *et al*., 2011; Peralta-Yahya *et al*., 2012; Chapter 1.2.2), yet no analogous thermophilic host has yet been developed. As such, the CBP-relevant advantageous physiological or bioengineering aspects associated with thermophily (Chapter 1.3.1) cannot be exploited in mesophilic strains. While it would be presumptuous to assume that any *Thermoanaerobacter* sp. have a biotechnological potential comparable to those demonstrated by organisms like *E. coli*, *B. subtilis* or *S. cerevisiae*, further research into *Thermoanaerobacter* spp. for these reasons may hold great potential. Such efforts have recently been undertaken in close phylogenetic relatives of *Thermoanaerobacter* for purposes of producing alternative fuels or solvents including acetone, isopropanol, n-propanol or propanediol (Shaw *et al*., 2011b; Mcbride *et al*., 2012).

No single, universal strategy has yet emerged for achieving industrially viable lignocellulosic biofuel production. Specifically, the type of biofuel, nature of the feedstock, process design, and particularly, the platform organism(s) used for catalyzing the conversion process, all remain debatable. While it is unlikely that a single, universally applicable strategy, or microorganism, exist, it is important to realize that

consistent amongst all current platform organisms is a well developed understanding of their native physiology. In many cases, these insights have developed not just through dedicated efforts into a single strain, but have evolved through the examination of the physiologies found in diverse microbes. As such, investigations into *Thermoanaerobacter* spp. will continue to play a valuable role in the development of platform microorganisms for second-generation biofuel production.

Appendix

A.1 Supplemental material for Chapter 2: Isolates of *Thermoanaerobacter thermohydrosulfuricus* from decaying wood compost display genetic and phenotypic microdiversity

A.1.1 Supplemental methods

Table A.1.1. Nucleotide accession numbers for sequences used in Fig. 2.1 and Fig. 2.2.

	16S	*cpn*60
Strain	GenBank accession number	GenBank accession number
Caldanaerobacter subterraneus subsp. *subterraneus* SEBR 7858	AF195797	NA[a]
Caldanaerobacter subterraneus subsp. *tengcongensis* MB4	AF209708	NC_003869
Caldanaerobacter subterraneus subsp. *yonsiensis* KB1	AF212925	NA
Clostridium thermocellum ATCC 27405	L09173	NC_009012
Thermoanaerobacter acetoethylicus ATCC 33265	L09163	NA
Thermoanaerobacter brockii subsp. *brockii* HTD4	L09165	HM623909
Thermoanaerobacter brockii subsp. *finnii* Ako-1	L09166	NZ_ACQZ01000003
Thermoanaerobacter brockii subsp. *lactiethylicus* SEBR 5268	U14330	NA
Thermoanaerobacter ethanolicus JW200	L09162	NZ_ACXY01000003
Thermoanaerobacter italicus Ab9	AJ250846	NZ_ACVH01000076
Thermoanaerobacter kivui DSM 2030	L09160	NA
Thermoanaerobacter mathranii subsp. *mathranii* A3	Y11279	DQ439966
Thermoanaerobacter pseudethanolicus 39E	L09164	NC_010321
Thermoanaerobacter siderophilus SR4	AF120479	NA

Table A.1.1 cont.

Strain	16S	*cpn60*
Thermoanaerobacter sp. Rt8.G4	NA	U56021
Thermoanaerobacter sp. strain X513	AF542520	ACPF01000040
Thermoanaerobacter sp. strain X514	AF542517	NC_010320
Thermoanaerobacter sp. strain X561	AF542518	ACXP01000037
Thermoanaerobacter sulfurophilus L-64	NR_026458	NA
Thermoanaerobacter thermocopriae ATCC 51646	L09167	AY691308
Thermoanaerobacter thermohydrosulfuricus DSM 567	L09161	HM623910
Thermoanaerobacter wiegelii Rt8.B1	NR_029301	CP002991
Thermoanaerobacterium thermosaccharolyticum DSM 571	M59119	NZ_ACVG01000042
Thermoanaerobacter thermohydrosulfuricus WC1	HM585213	HM623896
Thermoanaerobacter thermohydrosulfuricus WC2	HM585214	HM623897
Thermoanaerobacter thermohydrosulfuricus WC3	HM585215	HM623898
Thermoanaerobacter thermohydrosulfuricus WC4	HM585216	HM623899
Thermoanaerobacter thermohydrosulfuricus WC5	HM585217	HM623900
Thermoanaerobacter thermohydrosulfuricus WC6	HM585218	HM623901
Thermoanaerobacter thermohydrosulfuricus WC7	HM585219	HM623902
Thermoanaerobacter thermohydrosulfuricus WC8	HM585220	HM623903
Thermoanaerobacter thermohydrosulfuricus WC9	HM585221	HM623904
Thermoanaerobacter thermohydrosulfuricus WC10	HM585222	HM623905
Thermoanaerobacter thermohydrosulfuricus WC11	HM585223	HM623906
Thermoanaerobacter thermohydrosulfuricus WC12	HM585224	HM623907
Thermoanaerobacterium thermosaccharolyticum WC13	HM585225	HM623908

[a]NA = Sequence data not available at time of analysis.

A.1.2 Supplemental results

Table A.1.2. Fermentation of 30 test carbon substrates by isolates WC1-WC12, *Th. thermosaccharolyticum* WC13, *T. brockii* subsp. *brockii* HTD4, *T. pseudethanolicus* 39E and *T. thermohydrosulfuricus* DSM 567.

	WC isolates															
Substrate	1	2	3	4	5	6	7	8	9	10	11	12	13	T.pse	T.bro	T.the
Acetate	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Arabinose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Casamino acids	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cellobiose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Citrate	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Esculin	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Fructose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Galactose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Galacturonic acid	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Glucose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Isoleucine	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Lactate	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Lactose	+	-	+	+	-	-	-	-	+	-	+	+	+	+	+	-
Maltose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Mannitol	+	-	+	+	-	-	-	-	+	-	+	-	-	+	+	+
Mannose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Melibiose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
Pyruvate	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
Raffinose	+	+	-	-	+	+	+	-	+	+	+	+	+	+	+	-
Rhamnose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ribose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Salicin	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Sorbitol	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
Sorbose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Succinate	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sucrose	+	-	-	-	-	-	-	-	-	-	+	-	+	+	+	+
Tartaric acid	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Tryptone	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Xylose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Yeast extract	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Symbols denote observable growth (+) or no observable growth (-) on a particular substrate. Abbreviations: T.pse = *T. pseudethanolicus* 39E; T.bro = *T. brockii* subsp. *brockii* HTD4; T.the = *T. thermohydrosulfuricus* DSM 567.

Table A.1.3. Sequence identity score matrices of aligned IVS sequences as found in the 16S rRNA gene sequences of *T. pseudethanolicus* 39E, *T. thermohydrosulfuricus* DSM 567 and isolates WC1-WC12.

	T. pseudethanolicus 39E	*T. thermohydrosulfuricus* DSM 567	Isolate WC1
IVS 1			
T. pseudethanolicus 39E[a]	1.000		
T. thermohydrosulfuricus DSM 567	0.973	1.000	
Isolate WC1[b]	1.000	0.973	1.000
IVS 2			
T. pseudethanolicus 39E	1.000		
T. thermohydrosulfuricus DSM 567	1.000	1.000	
Isolate WC1	1.000	1.000	1.000
IVS 3			
T. pseudethanolicus 39E	1.000		
T. thermohydrosulfuricus DSM 567	1.000	1.000	
Isolate WC1	1.000	1.000	1.000
IVS 4			
T. pseudethanolicus 39E	1.000		
T. thermohydrosulfuricus DSM 567	0.915	1.000	
Isolate WC1	1.000	0.915	1.000

[a]The 16S rRNA gene sequence used for *T. pseudethanolicus* 39E is found at position 2,265,749 - 2,267,517 within the published genome and has 4 IVS regions.

[b]Isolate WC1 used to represent all WC isolates. IVS region is identical amongst WC isolates.

Note: The 16S rRNA gene sequence for *T. brockii* subsp. *brockii* HTD4 contains no identifiable IVS

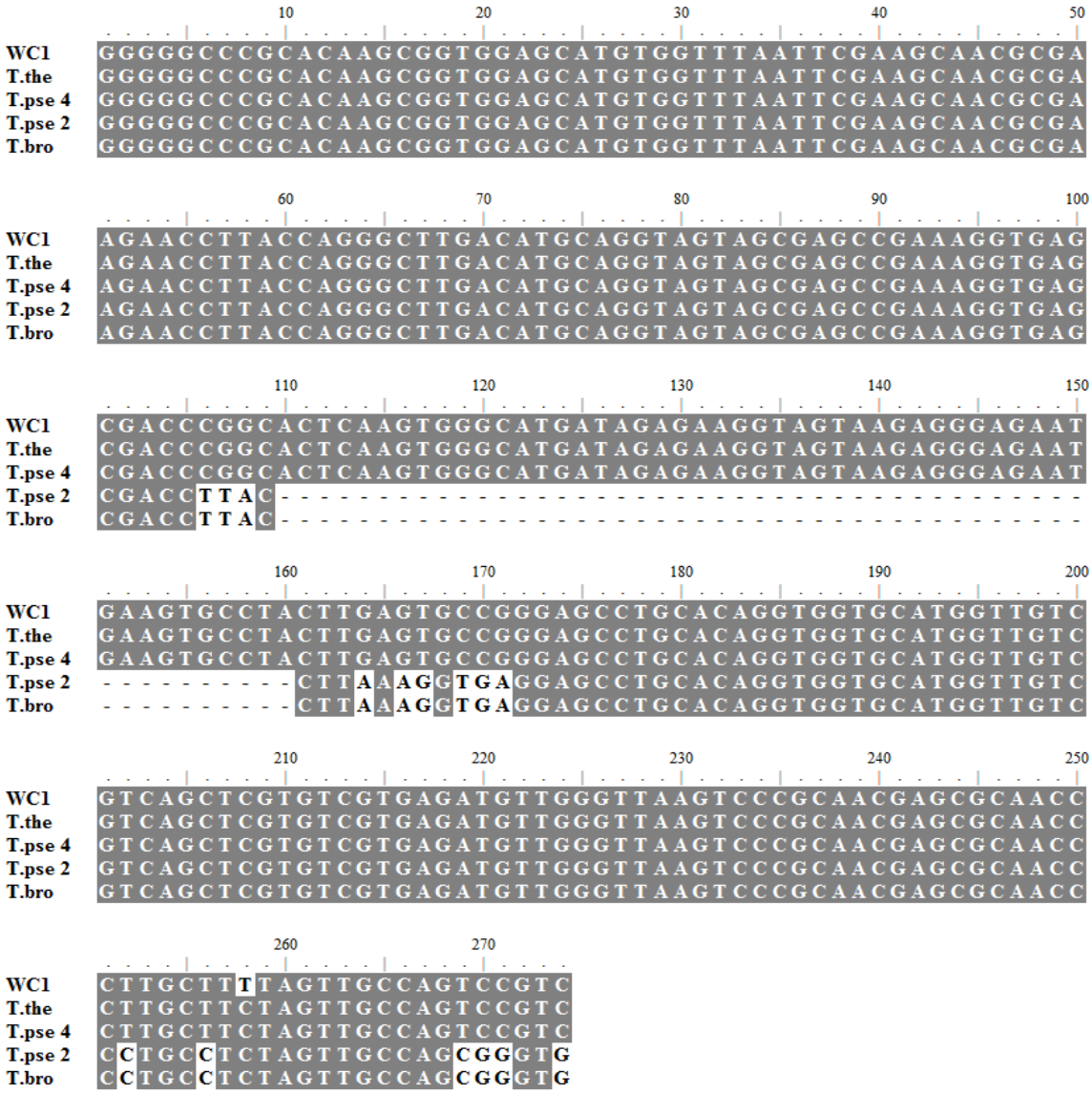


Fig. A.1.1. Representative sequence alignment of a single IVS found in isolates WC1-WC12, *T. thermohydrosulfuricus* DSM 567 and *T. pseudethanolicus* 39E. T.pse 4 refers to the gene found at position 2,265,749 - 2,267,517 containing 4 IVS, while T.pse 2 refers to the gene found at position 1,373,190 – 1,374, 716 containing no IVS. Shading identifies identical nucleotides (grey) or non-identical nucleotides (white). Abbreviations: T.the = *T. thermohydrosulfuricus* DSM 567; T.pse = *T. pseudethanolicus* 39E; T. bro = *T. brockii* subsp. *brockii* HTD4.

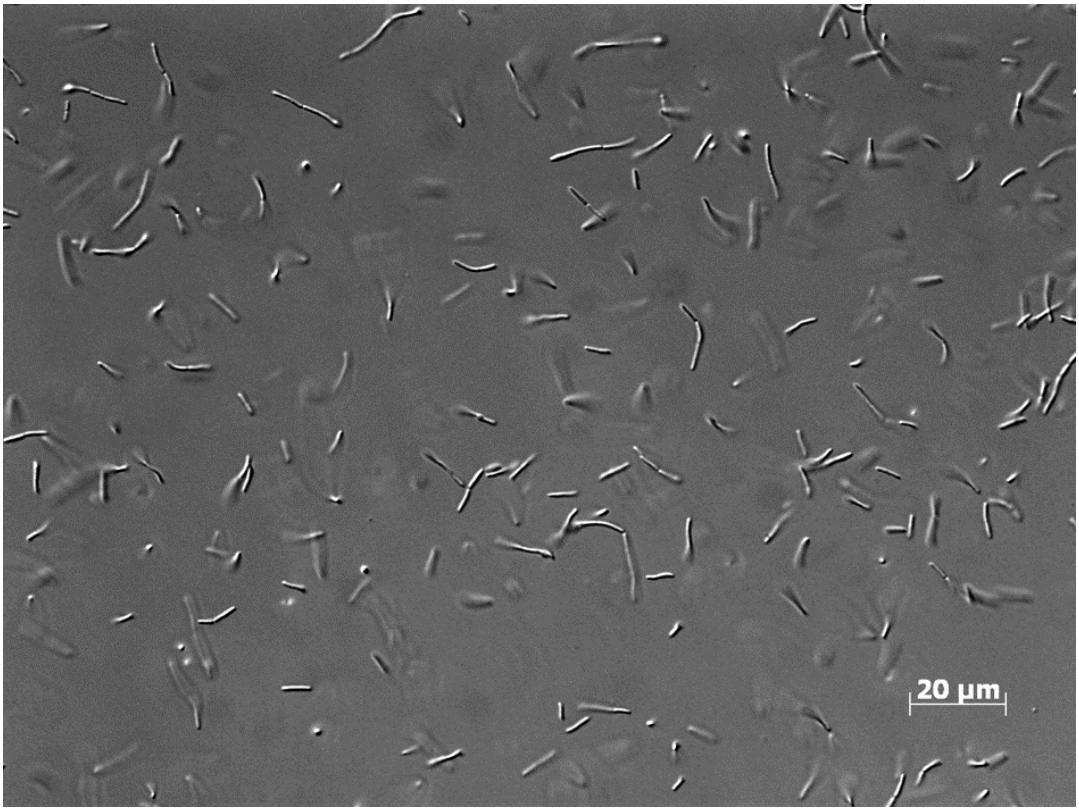


Fig. A.1.2. Representative example of cell morphology and arrangement. Image of WC6 taken at 1000× magnification as described (Chapter 2.3.4).

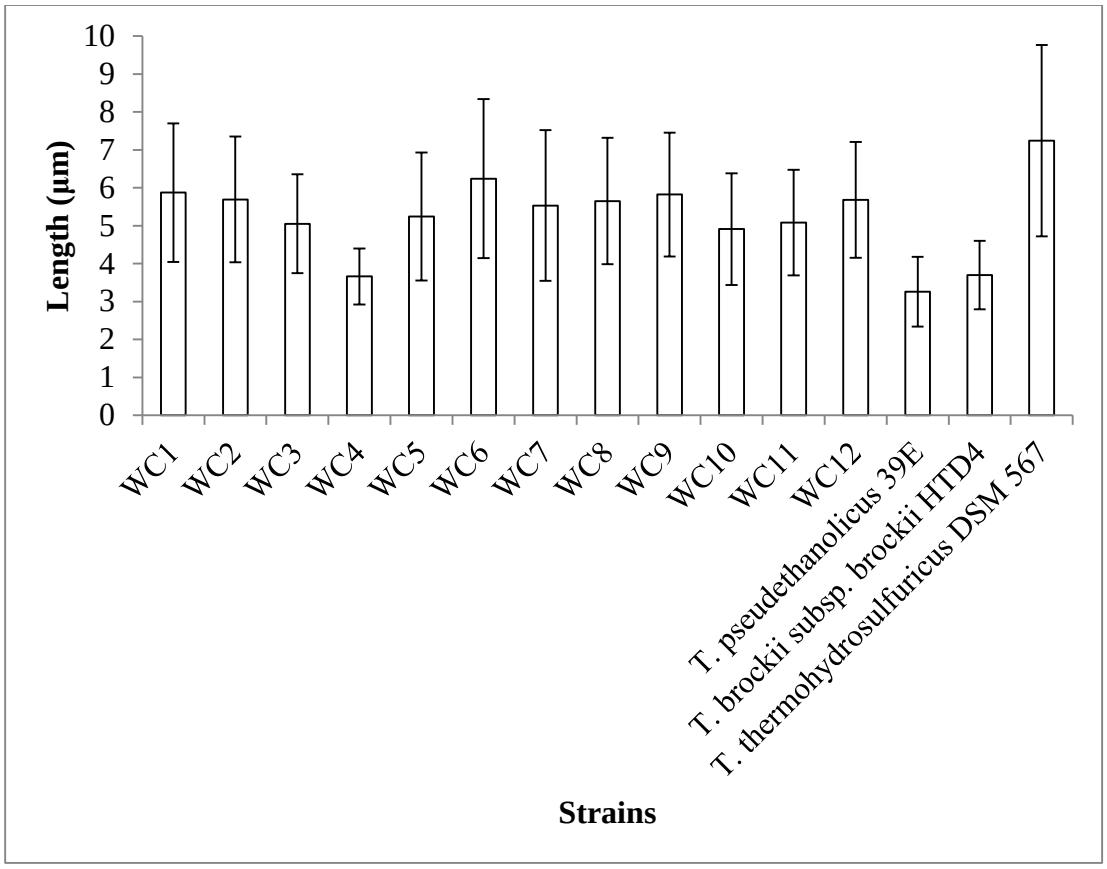


Fig. A.1.3. Average cell lengths of cultures grown to an $OD_{600} = 0.75 \pm 0.03$ as described in Chapter 2.3.4.

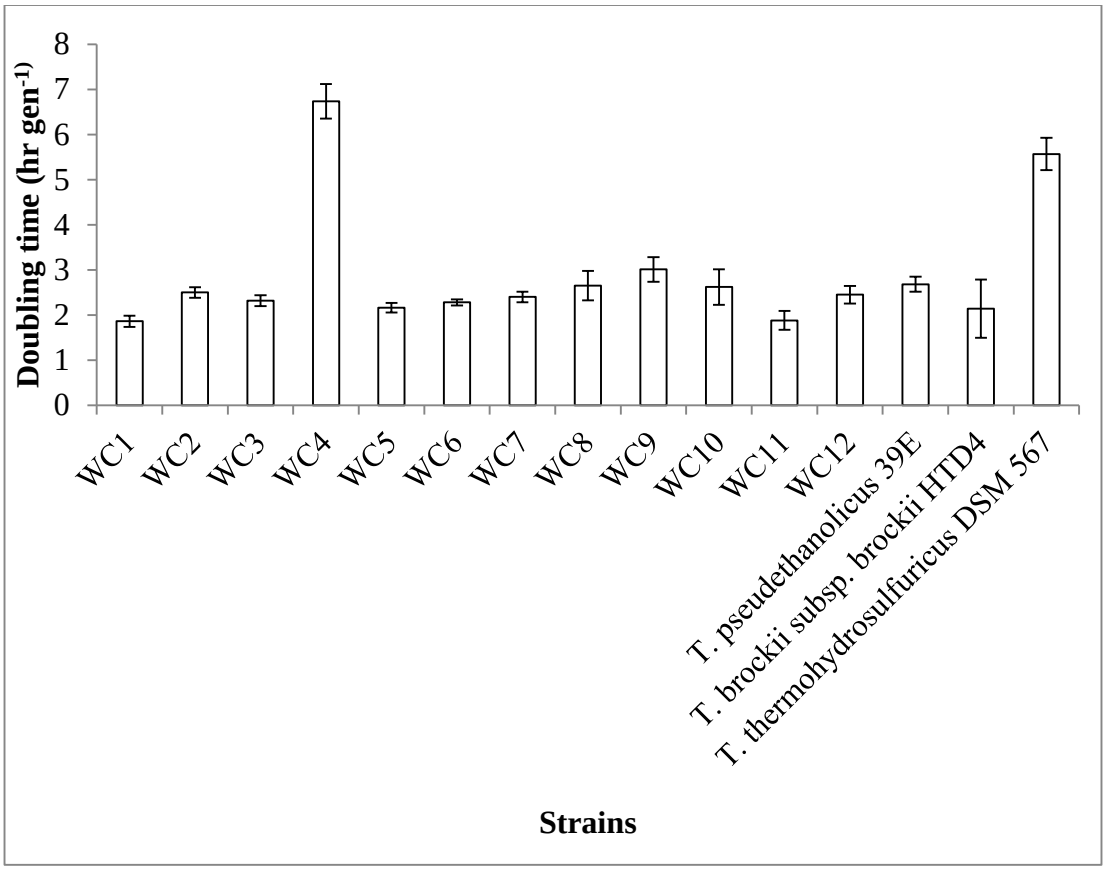


Fig. A.1.4. Average calculated doubling time of cultures grown to an OD_{600} = 0.75 ± 0.03.

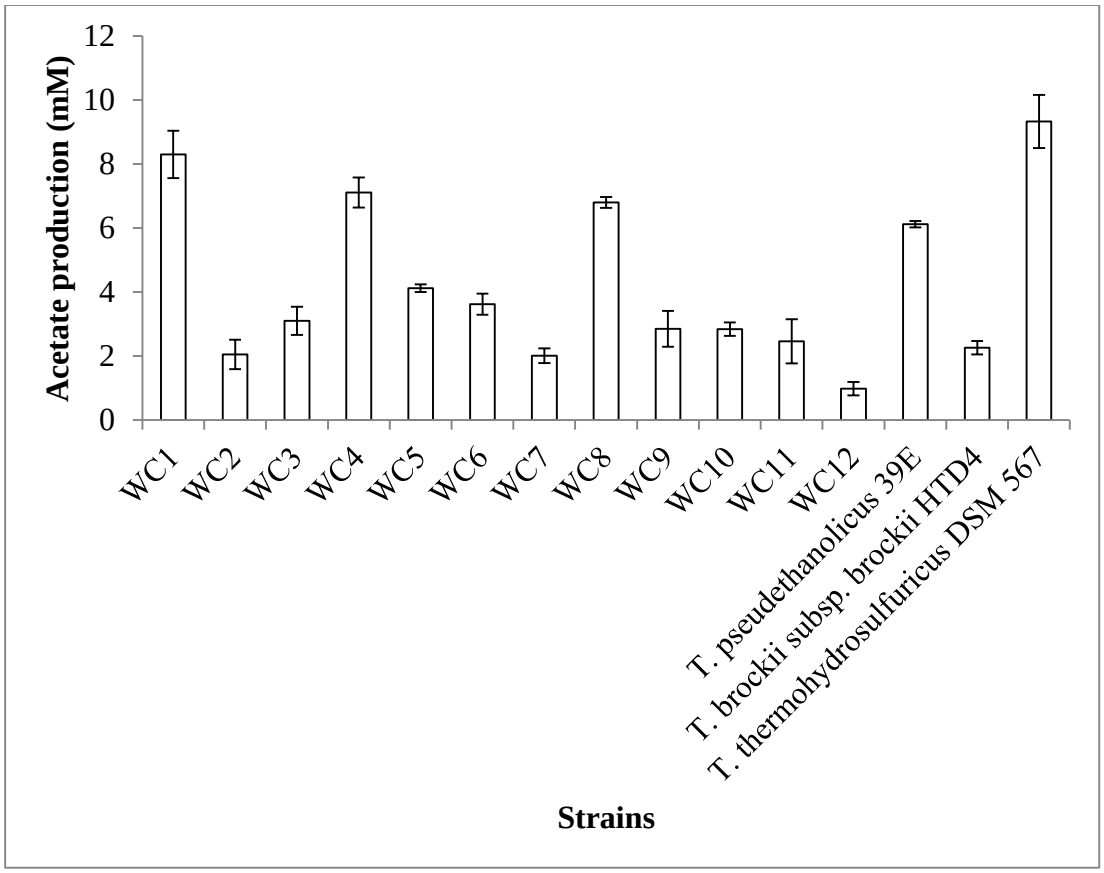


Fig. A.1.5. Average acetate production of cultures grown to an OD_{600} = 0.75 ± 0.03.

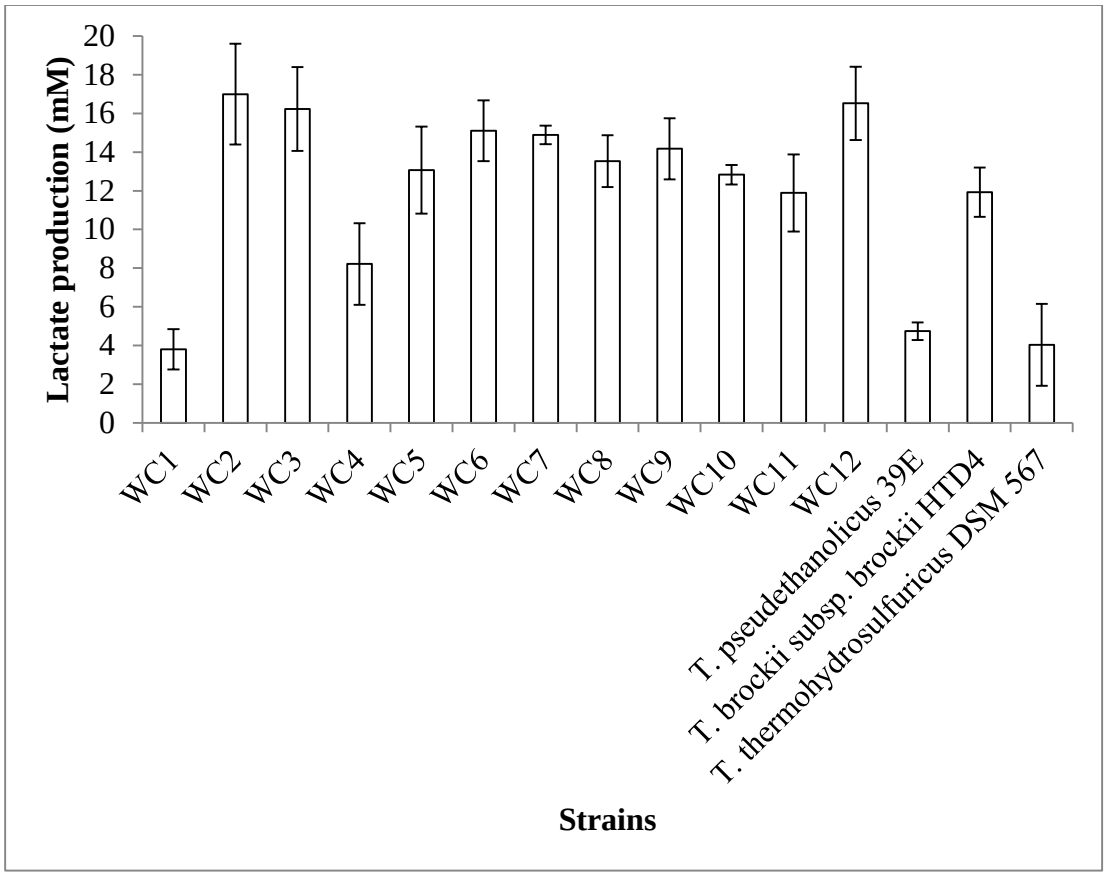


Fig. A.1.6. Average lactate production of cultures grown to an OD_{600} = 0.75 ± 0.03.

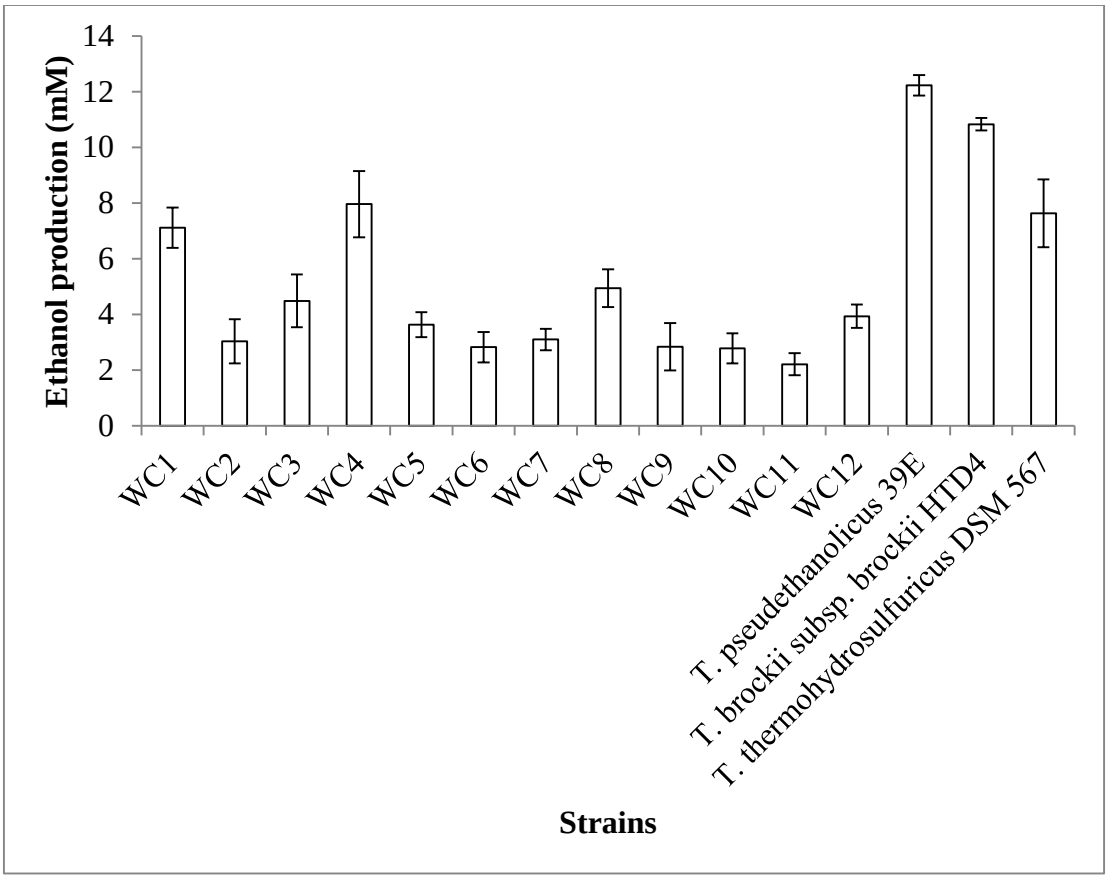


Fig. A.1.7. Average ethanol production of cultures grown to an OD_{600} = 0.75 ± 0.03.

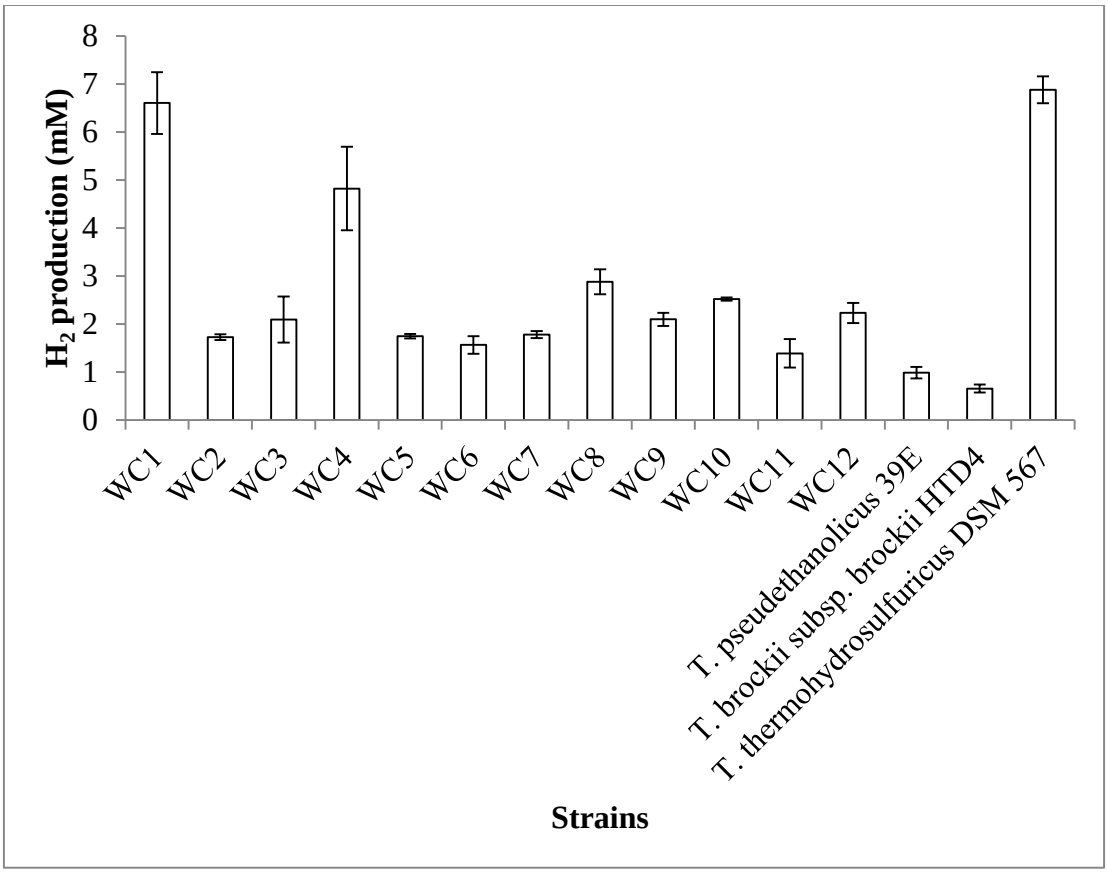


Fig. A.1.8. Average H_2 production of cultures grown to an $OD_{600} = 0.75 \pm 0.03$.

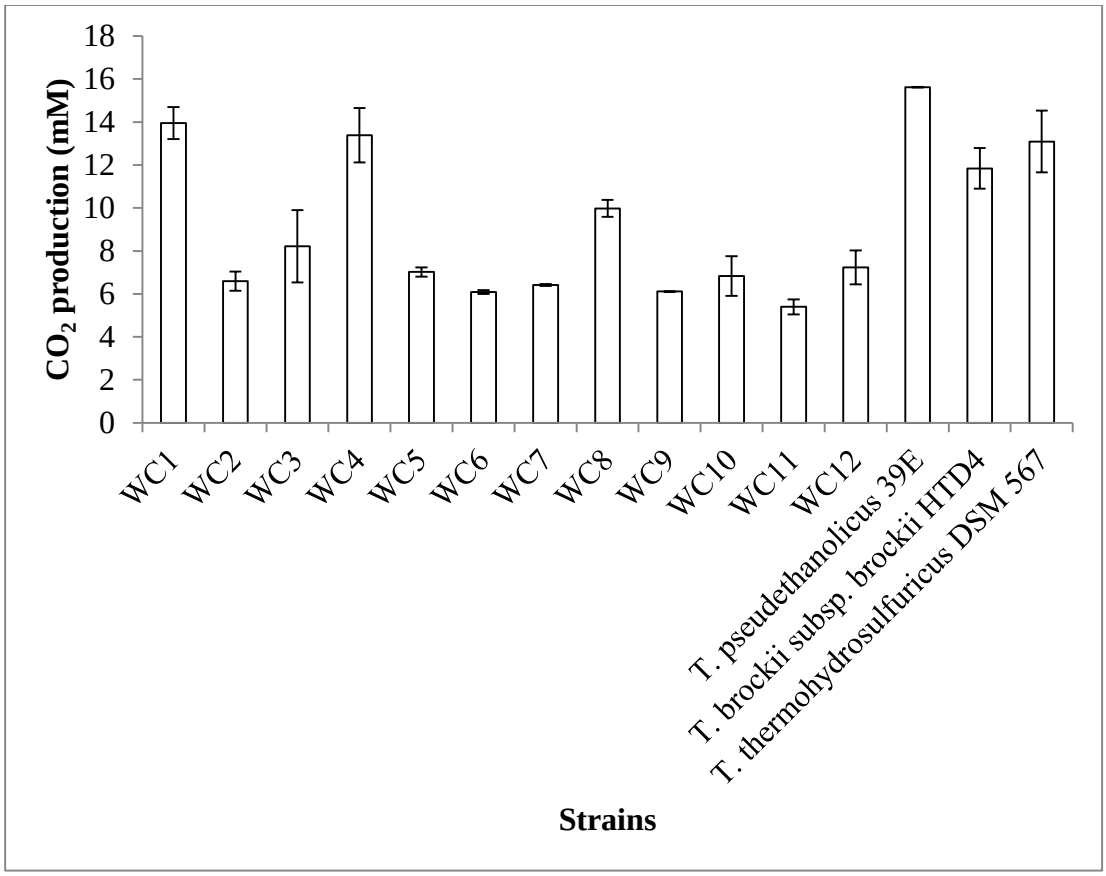


Fig. A.1.9. Average CO_2 production of cultures grown to an OD_{600} = 0.75 ± 0.03.

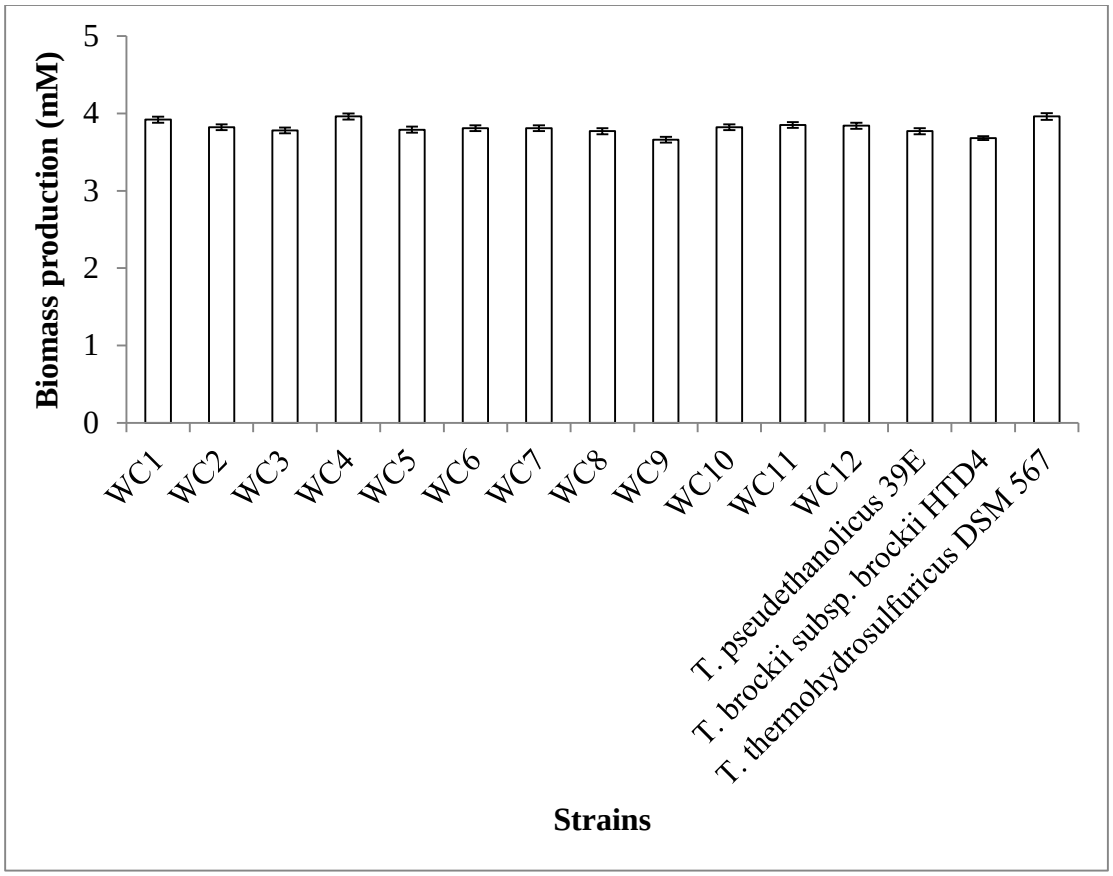


Fig. A.1.10. Average biomass production of cultures grown to an OD_{600} = 0.75 ± 0.03.

A.2 Supplemental material for Chapter 3: Predicting relatedness of bacterial genomes using the chaperonin-60 universal target (*cpn60* UT): Application to *Thermoanaerobacter* species.

A.2.1 Supplemental methods

A.2.1.1 Primer design

Genomic data for *T. pseudethanolicus* 39E, *T. brockii* subsp. *finnii* Ako-1, *T. ethanolicus* CCSD1, *Thermoanaerobacter* sp. X513, *Thermoanaerobacter* sp. X514, *Thermoanaerobacter* sp. X561, *T. italicus* Ab9 and *T. mathranii* subsp. *mathranii* A3 was accessed from the Joint Genome Institute's Integrated Microbial Genomes database (Markowitz *et al*, 2012)). Sequences for *recN*, *thdF* (*trmE*) and *rpoA*, including the 100 nucleotides both upstream and downstream of each gene were determined and aligned using BioEdit v.7.0.9.0 (Hall, 1999). Conserved regions were identified and primers (Table A.2.1) were designed using the publicly accessible Primer3 software (Untergrasser *et al*., 2012) such that the entire gene region would be accessible for sequencing. Additional sequencing primers, internal to the gene sequence, were designed for use with *recN*.

Table A.2.1. Amplification and sequencing primers for *Thermoanaerobacter* isolates and reference strains.

Target	Primer	Sequence (5' - 3')	Use
recN	recNEF	TTCTGGGGGTTTTGTTATGC	Amplification/sequencing
	recNER	TCCGATAGTAAAAGGAAGTGAGA	Amplification/sequencing
	recNIF2	GAAAAGGCGAATTTGAGTGTAG	Sequencing
	recNIR2	AAGGCTTCAAAGGTTCTCCTC	Sequencing
rpoA	rpoAF	AGGAGGGTTCACAGTGTTTGA	Amplification/sequencing
	rpoAR	GAAGGACGGCCTAATTTTCTG	Amplification/sequencing
thdF	thdF	TTGGTTGGTGATGAAAATGG	Amplification/sequencing
	thdR	CCAACAACTGCAACATCGTA	Amplification/sequencing

A.2.2 Supplemental results

Table A.2.2. Gene sequence identities and prediction of genome sequence identities using the three-gene model and the c*pn60* UT one-gene model: data for all isolates and reference strains[a] .

					Prediction of genome SI using		
Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	$SI_{cpn60\ UT}$	Three-gene model[b]	One-gene model[c]	Species prediction (both models)
T. brockii subsp. *brockii* HTD4-WC1	0.969	0.967	0.873	0.893	0.681	0.512	Different
T. brockii subsp. *brockii* HTD4-WC2	0.969	0.967	0.874	0.882	0.682	0.475	Different
T. brockii subsp. *brockii* HTD4-WC3	0.966	0.969	0.875	0.873	0.684	0.444	Different
T. brockii subsp. *brockii* HTD4-WC4	0.966	0.969	0.872	0.880	0.679	0.469	Different
T. brockii subsp. *brockii* HTD4-WC5	0.968	0.969	0.876	0.884	0.686	0.481	Different
T. brockii subsp. *brockii* HTD4-WC6	0.966	0.968	0.876	0.880	0.683	0.469	Different
T. brockii subsp. *brockii* HTD4-WC7[e]	0.966	ND[d]	ND	0.871	ND	0.438	Different
T. brockii subsp. *brockii* HTD4-WC8	0.968	0.968	0.873	0.891	0.681	0.506	Different
T. brockii subsp. *brockii* HTD4-WC9	0.966	0.967	0.873	0.891	0.679	0.506	Different
T. brockii subsp. *brockii* HTD4-WC10	0.966	0.969	0.875	0.891	0.684	0.506	Different
T. brockii subsp. *brockii* HTD4-WC11[f]	ND	0.968	0.875	0.893	ND	0.512	Different
T. brockii subsp. *brockii* HTD4-WC12	0.969	0.967	0.873	0.888	0.681	0.494	Different
T. pseudethanolicus 39E-WC1	0.969	0.967	0.877	0.893	0.688	0.512	Different
T. pseudethanolicus 39E-WC2	0.969	0.967	0.878	0.882	0.690	0.475	Different
T. pseudethanolicus 39E-WC3	0.966	0.969	0.879	0.873	0.691	0.444	Different
T. pseudethanolicus 39E-WC4	0.966	0.969	0.876	0.880	0.686	0.469	Different
T. pseudethanolicus 39E-WC5	0.968	0.969	0.880	0.884	0.694	0.481	Different
T. pseudethanolicus 39E-WC6	0.964	0.968	0.880	0.880	0.691	0.469	Different
T. pseudethanolicus 39E-WC7	0.966	ND	ND	0.871	ND	0.438	Different
T. pseudethanolicus 39E-WC8	0.968	0.968	0.877	0.891	0.688	0.506	Different

Table A.2.2 cont.

Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	$SI_{cpn60\ UT}$	Three-gene model	One-gene model	Species prediction (both models)
T. pseudethanolicus 39E-WC9	0.966	0.967	0.877	0.891	0.686	0.506	Different
T. pseudethanolicus 39E-WC10	0.966	0.969	0.879	0.891	0.691	0.506	Different
T. pseudethanolicus 39E-WC11	ND	0.968	0.879	0.893	ND	0.512	Different
T. pseudethanolicus 39E-WC12	0.969	0.967	0.876	0.888	0.686	0.494	Different
T. thermohydrosulfuricus DSM 567-WC1	0.997	0.997	0.998	0.984	0.934	0.824	Same
T. thermohydrosulfuricus DSM 567-WC2	0.997	0.997	0.997	0.987	0.931	0.836	Same
T. thermohydrosulfuricus DSM 567-WC3	0.994	1.000	0.998	0.996	0.934	0.868	Same
T. thermohydrosulfuricus DSM 567-WC4	0.994	1.000	0.998	0.989	0.932	0.843	Same
T. thermohydrosulfuricus DSM 567-WC5	0.995	1.000	0.999	0.993	0.936	0.855	Same
T. thermohydrosulfuricus DSM 567-WC6	0.989	0.999	0.999	0.996	0.932	0.868	Same
T. thermohydrosulfuricus DSM 567-WC7	0.994	ND	ND	0.995	ND	0.861	Same
T. thermohydrosulfuricus DSM 567-WC8	0.995	0.999	0.998	0.978	0.934	0.805	Same
T. thermohydrosulfuricus DSM 567-WC9	0.994	0.997	0.998	0.984	0.931	0.824	Same
T. thermohydrosulfuricus DSM 567-WC10	0.994	1.000	0.998	0.986	0.932	0.830	Same
T. thermohydrosulfuricus DSM 567-WC11	ND	0.999	1.000	0.980	ND	0.811	Same
T. thermohydrosulfuricus DSM 567-WC12	0.997	0.997	0.994	0.980	0.927	0.811	Same
WC1-WC2	1.000	0.995	0.995	0.989	0.928	0.843	Same
WC1-WC3	0.997	0.997	0.998	0.980	0.934	0.811	Same
WC1-WC4	0.997	0.997	0.999	0.987	0.935	0.836	Same
WC1-WC5	0.998	0.997	0.998	0.991	0.933	0.849	Same
WC1-WC6	0.992	0.996	0.998	0.987	0.929	0.836	Same
WC1-WC7	0.997	ND	ND	0.978	ND	0.805	Same
WC1-WC8	0.998	0.996	0.998	0.995	0.934	0.861	Same
WC1-WC9	0.997	0.995	0.996	0.996	0.928	0.868	Same
WC1-WC10	0.997	0.997	0.996	0.998	0.930	0.874	Same

Table A.2.2 cont.

Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	$SI_{cpn60\ UT}$	Three-gene model	One-gene model	Species prediction (both models)
WC1-WC11	ND	0.996	0.998	0.996	ND	0.868	Same
WC1-WC12	1.000	0.995	0.994	0.993	0.927	0.855	Same
WC2-WC3	0.997	0.997	0.997	0.987	0.931	0.836	Same
WC2-WC4	0.997	0.997	0.996	0.995	0.930	0.861	Same
WC2-WC5	0.998	0.997	0.998	0.995	0.933	0.861	Same
WC2-WC6	0.998	0.999	0.998	0.991	0.931	0.849	Same
WC2-WC7	0.997	ND	ND	0.986	ND	0.830	Same
WC2-WC8	0.998	0.999	0.997	0.984	0.933	0.824	Same
WC2-WC9	0.997	1.000	0.999	0.989	0.937	0.843	Same
WC2-WC10	0.997	0.997	0.998	0.991	0.933	0.849	Same
WC2-WC11	ND	0.999	0.997	0.986	ND	0.830	Same
WC2-WC12	1.000	0.997	0.996	0.986	0.931	0.830	Same
WC3-WC4	1.000	1.000	0.998	0.991	0.936	0.849	Same
WC3-WC5	0.998	1.000	0.999	0.989	0.938	0.843	Same
WC3-WC6	0.992	0.999	0.999	0.993	0.934	0.855	Same
WC3-WC7	1.000	ND	ND	0.996	ND	0.868	Same
WC3-WC8	0.998	0.999	0.998	0.975	0.936	0.793	Same
WC3-WC9	1.000	0.997	0.996	0.980	0.931	0.811	Same
WC3-WC10	1.000	1.000	0.996	0.982	0.933	0.818	Same
WC3-WC11	ND	0.999	0.998	0.976	ND	0.799	Same
WC3-WC12	0.997	0.997	0.994	0.976	0.927	0.799	Same
WC4-WC5	0.998	1.000	0.997	0.993	0.933	0.855	Same
WC4-WC6	0.998	0.999	0.997	0.993	0.929	0.855	Same
WC4-WC7	1.000	ND	ND	0.991	ND	0.849	Same
WC4-WC8	0.998	0.999	0.999	0.982	0.937	0.818	Same

Table A.2.2 cont.

Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	$SI_{cpn60\ UT}$	Three-gene model	One-gene model	Species prediction (both models)
WC4-WC9	1.000	0.997	.996	0.987	0.933	0.836	Same
WC4-WC10	1.000	1.000	0.997	0.989	0.934	0.843	Same
WC4-WC11	ND	0.999	0.998	0.984	ND	0.824	Same
WC4-WC12	0.997	0.997	0.995	0.984	0.928	0.824	Same
WC5-WC6	1.000	0.999	1.000	0.996	0.939	0.868	Same
WC5-WC7	0.994	ND	ND	0.987	ND	0.836	Same
WC5-WC8	0.998	0.999	0.998	0.986	0.934	0.830	Same
WC5-WC9	1.000	0.997	0.997	0.991	0.933	0.849	Same
WC5-WC10	0.998	1.000	0.997	0.993	0.933	0.855	Same
WC5-WC11	ND	0.999	0.999	0.987	ND	0.836	Same
WC5-WC12	0.998	0.997	0.995	0.987	0.929	0.836	Same
WC6-WC7	0.992	ND	ND	0.991	ND	0.849	Same
WC6-WC8	0.994	1.000	0.998	0.982	0.932	0.818	Same
WC6-WC9	0.992	0.999	0.997	0.987	0.929	0.836	Same
WC6-WC10	0.992	0.999	0.997	0.989	0.929	0.843	Same
WC6-WC11	ND	1.000	0.999	0.984	ND	0.824	Same
WC6-WC12	0.992	0.999	0.995	0.984	0.926	0.824	Same
WC7-WC8	0.998	ND	ND	0.976	ND	0.799	Same
WC7-WC9	1.000	ND	ND	0.978	ND	0.805	Same
WC7-W10	1.000	ND	ND	0.980	ND	0.811	Same
WC7-W11	ND	ND	ND	0.975	ND	0.793	Same
WC7-W12	0.997	ND	ND	0.978	ND	0.805	Same
WC8-WC9	0.998	0.999	0.998	0.995	0.934	0.861	Same
WC8-WC10	0.998	0.999	0.998	0.993	0.934	0.855	Same
WC8-WC11	ND	1.000	0.998	0.998	ND	0.874	Same

Table A.2.2 cont.

Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	$SI_{cpn60\ UT}$	Three-gene model	One-gene model	Species prediction (both models)
WC8-WC12	0.998	0.999	0.994	0.995	0.928	0.861	Same
WC9-WC10	1.000	0.997	0.998	0.998	0.936	0.874	Same
WC9-WC11	ND	0.999	0.998	0.996	ND	0.868	Same
WC9-WC12	0.997	0.997	0.995	0.993	0.928	0.855	Same
WC10-WC11	ND	0.999	0.998	0.995	ND	0.861	Same
WC10-WC12	0.997	0.997	0.995	0.995	0.928	0.861	Same
WC11-WC12	ND	0.999	0.994	0.993	ND	0.855	Same
T. thermohydrosulfuricus DSM 567 - *T. brockii* subsp. *brockii* HTD4	0.966	0.969	0.879	0.877	0.691	0.456	Different
T. thermohydrosulfuricus DSM 567 - *T. brockii* subsp. *finnii* Ako-1	0.966	0.969	0.879	0.877	0.691	0.456	Different
T. thermohydrosulfuricus DSM 567 - *T. ethanolicus* CCSD1	0.966	0.969	0.878	0.877	0.689	0.456	Different
T. thermohydrosulfuricus DSM 567 - *T. italicus* Ab9	0.900	0.942	0.871	0.868	0.627	0.425	Different
T. thermohydrosulfuricus DSM 567 - *T. mathranii* subsp. *mathranii* A3[f]	0.902	0.941	0.870	0.868	0.635	0.425	Different
T. thermohydrosulfuricus DSM 567 - *T. pseudethanolicus* 39E	0.966	0.969	0.879	0.877	0.691	0.456	Different
T. thermohydrosulfuricus DSM 567 - *Cal. subterraneus* subsp. *tengcongensis* MB4	0.715	0.784	0.754	0.815	0.241	0.244	Different
T. thermohydrosulfuricus DSM 567 - *Th. thermosaccharolyticum* DSM 571	0.651	0.729	0.648	0.777	-0.008	0.113	Different
T. thermohydrosulfuricus DSM 567 - *T. wiegelii* Rt8.B1	ND	0.999	0.984	0.982	ND	0.818	Same
T. thermohydrosulfuricus DSM 567 - *Thermoanaerobacter* sp. X513	0.965	0.969	0.879	0.877	0.690	0.456	Different

Table A.2.2 cont.

Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	$SI_{cpn60\ UT}$	Three-gene model	One-gene model	Species prediction (both models)
T. thermohydrosulfuricus DSM 567 - *Thermoanaerobacter* sp. X514	0.965	0.969	0.879	0.877	0.690	0.456	Different
T. thermohydrosulfuricus DSM 567 - *Thermoanaerobacter* sp. X561	0.965	0.969	0.879	0.877	0.690	0.456	Different
T. brockii subsp. *brockii* HTD4 - *T. brockii* subsp. *finnii* Ako-1	1.000	1.000	0.993	1.000	0.928	0.880	Same
T. brockii subsp. *brockii* HTD4 - *T. ethanolicus* CCSD1	1.000	1.000	0.999	1.000	0.939	0.880	Same
T. brockii subsp. *brockii* HTD4 - *T. italicus* Ab9	0.912	0.949	0.896	0.944	0.682	0.687	Different
T. brockii subsp. *brockii* HTD4 - *T. mathranii* A3	0.911	0.954	0.899	0.944	0.691	0.687	Different
T. brockii subsp. *brockii* HTD4 - *T. pseudethanolicus* 39E	1.000	1.000	0.996	1.000	0.932	0.880	Same
T. brockii subsp. *brockii* HTD4 - *T. wiegelii* Rt8.B1	ND	0.971	0.890	0.891	ND	0.505	Different
T. brockii subsp. *brockii* HTD4 - *Thermoanaerobacter* sp. X513	0.998	1.000	0.999	1.000	0.937	0.880	Same
T. brockii subsp. *brockii* HTD4 - *Thermoanaerobacter* sp. X514	0.998	1.000	0.994	1.000	0.929	0.880	Same
T. brockii subsp. *brockii* HTD4 - *Thermoanaerobacter* sp. X561	0.998	1.000	0.999	1.000	0.937	0.880	Same
T. brockii subsp. *brockii* HTD4 - *Cal. subterraneus* subsp. *tengcongensis* MB4	0.711	0.790	0.749	0.830	0.234	0.294	Different
T. brockii subsp. *brockii* HTD4 - *Th. thermosaccharolyticum* DSM 571	0.646	0.728	0.653	0.774	-0.003	0.101	Different
T. brockii subsp. *finnii* Ako-1 - *T. ethanolicus* CCSD1	1.000	1.000	0.999	1.000	0.938	0.880	Same

Table A.2.2 cont.

Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	$SI_{cpn60\ UT}$	Three-gene model	One-gene model	Species prediction (both models)
T. brockii subsp. *finnii* Ako-1 - *T. italicus* Ab9	0.938	0.952	0.891	0.944	0.689	0.687	Different
T. brockii subsp. *finnii* Ako-1 - *T. mathranii* subsp. *mathranii* A3	0.935	0.954	0.890	0.944	0.686	0.687	Different
T. brockii subsp. *finnii* Ako-1 - *T. pseudethanolicus* 39E	1.000	1.000	1.000	1.000	0.940	0.880	Same
T. brockii subsp. *finnii* Ako-1 - *T. wiegelii* Rt8.B1	ND	0.971	0.884	0.891	ND	0.505	Different
T. brockii subsp. *finnii* Ako-1 - *Thermoanaerobacter* sp. X513	0.997	1.000	0.998	1.000	0.935	0.880	Same
T. brockii subsp. *finnii* Ako-1 - *Thermoanaerobacter* sp. X514	0.997	1.000	0.998	1.000	0.935	0.880	Same
T. brockii subsp. *finnii* Ako-1 - *Thermoanaerobacter* sp. X561	0.997	1.000	0.998	1.000	0.935	0.880	Same
T. brockii subsp. *finnii* Ako-1 - *Cal. subterraneus* subsp. *tengcongensis* MB4	0.713	0.788	0.737	0.830	0.212	0.295	Different
T. brockii subsp. *finnii* Ako-1 - *Th. thermosaccharolyticum* DSM 571	0.654	0.732	0.581	0.774	-0.125	0.103	Different
T. ethanolicus CCSD1 - *T. italicus* Ab9	0.912	0.949	0.895	0.944	0.681	0.687	Different
T. ethanolicus CCSD1 - *T. mathranii* subsp. *mathranii* A3	0.911	0.954	0.898	0.944	0.689	0.687	Different
T. ethanolicus CCSD1 - *T. pseudethanolicus* 39E	1.000	1.000	0.995	1.000	0.931	0.880	Same
T. ethanolicus CCSD1 - *T. wiegelii* Rt8.B1	ND	0.971	0.883	0.891	ND	0.505	Different
T. ethanolicus CCSD1 - *Thermoanaerobacter* sp. X513	0.998	1.000	0.998	1.000	0.935	0.880	Same
T. ethanolicus CCSD1 - *Thermoanaerobacter* sp. X514	0.998	1.000	0.994	1.000	0.928	0.880	Same

Table A.2.2 cont.

Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	$SI_{cpn60\ UT}$	Three-gene model	One-gene model	Species prediction (both models)
T. ethanolicus CCSD1 - *Thermoanaerobacter* sp. X561	0.998	1.000	0.998	1.000	0.935	0.880	Same
T. ethanolicus CCSD1 - *Cal. subterraneus* subsp. *tengcongensis* MB4	0.711	0.790	0.749	0.830	0.233	0.294	Different
T. ethanolicus CCSD1 - *Th. thermosaccharolyticum* DSM 571	0.646	0.728	0.654	0.774	-0.002	0.101	Different
T. italicus Ab9 - *T. mathranii* subsp. *mathranii* A3	0.979	0.995	0.995	1.000	0.917	0.880	Same
T. italicus Ab9 - *T. ethanolicus* CCSD1	0.912	0.949	0.899	0.944	0.689	0.687	Different
T. italicus Ab9 - *T. wiegelii* Rt8.B1	ND	0.943	0.879	0.882	ND	0.474	Different
T. italicus Ab9 - *Thermoanaerobacter* sp. X513	0.912	0.949	0.899	0.944	0.689	0.687	Different
T. italicus Ab9 - *Thermoanaerobacter* sp. X514	0.912	0.949	0.899	0.944	0.689	0.687	Different
T. italicus Ab9 - *Thermoanaerobacter* sp. X561	0.912	0.949	0.899	0.944	0.689	0.687	Different
T. italicus Ab9 – *Cal. subterraneus* subsp. *tengcongensis* MB4	0.691	0.780	0.753	0.819	0.226	0.257	Different
T. italicus Ab9 - *Th. thermosaccharolyticum* DSM 571	0.644	0.730	0.658	0.761	0.006	0.057	Different
T. mathranii subsp. *mathranii* A3 - *T. pseudethanolicus* 39E	0.911	0.954	0.895	0.944	0.683	0.687	Different
T. mathranii subsp. *mathranii* A3 - *T. wiegelii* Rt8.B1	ND	0.944	0.878	0.884	ND	0.481	Different
T. mathranii subsp. *mathranii* A3 - *Thermoanaerobacter* sp. X513	0.912	0.954	0.899	0.944	0.691	0.687	Different
T. mathranii subsp. *mathranii* A3 - *Thermoanaerobacter* sp. X514	0.912	0.954	0.895	0.944	0.683	0.687	Different

Table A.2.2 cont.

Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	$SI_{cpn60\ UT}$	Three-gene model	One-gene model	Species prediction (both models)
T. mathranii subsp. *mathranii* A3 - *Thermoanaerobacter* sp. X561	0.912	0.954	0.899	0.944	0.691	0.687	Different
T. mathranii subsp. *mathranii* A3 - *Cal. subterraneus* subsp. *tengcongensis* MB4	0.693	0.781	0.748	0.819	0.218	0.257	Different
T. mathranii subsp. *mathranii* A3 - *Th. thermosaccharolyticum* DSM 571	0.644	0.737	0.652	0.761	-0.000	0.057	Different
T. pseudethanolicus 39E - *T. wiegelii* Rt8.B1	ND	0.971	0.884	0.891	ND	0.505	Different
T. pseudethanolicus 39E - *Thermoanaerobacter* sp. X513	0.998	1.000	0.999	1.000	0.937	0.880	Same
T. pseudethanolicus 39E - *Thermoanaerobacter* sp. X514	0.998	1.000	0.999	1.000	0.937	0.880	Same
T. pseudethanolicus 39E - *Thermoanaerobacter* sp. X561	0.998	1.000	0.999	1.000	0.937	0.880	Same
T. pseudethanolicus 39E - *Cal. subterraneus* subsp. *tengcongensis* MB4	0.711	0.790	0.753	0.830	0.240	0.294	Different
T. pseudethanolicus 39E - *Th. thermosaccharolyticum* DSM 571	0.646	0.728	0.656	0.774	0.002	0.101	Different
T. wiegelii Rt8.B1 - *Thermoanaerobacter* sp. X513	ND	0.971	0.884	0.891	ND	0.505	Different
T. wiegelii Rt8.B1 - *Thermoanaerobacter* sp. X514	ND	0.971	0.884	0.891	ND	0.505	Different
T. wiegelii Rt8.B1 - *Thermoanaerobacter* sp. X561	ND	0.971	0.884	0.891	ND	0.505	Different
T. wiegelii Rt8.B1 - *Cal. subterraneus* subsp. *tengcongensis* MB4	ND	0.785	0.750	0.832	ND	0.302	Different
T. wiegelii Rt8.B1 - *Th. thermosaccharolyticum* DSM 571	ND	0.731	0.584	0.786	ND	0.144	Different

Table A.2.2 cont.

Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	$SI_{cpn60\ UT}$	Three-gene model	One-gene model	Species prediction (both models)
Thermoanaerobacter sp. X513 - *Thermoanaerobacter* sp. X514	1.000	1.000	0.996	1.000	0.932	0.880	Same
Thermoanaerobacter sp. X513 - *Thermoanaerobacter* sp. X561	1.000	1.000	1.000	1.000	0.940	0.880	Same
Thermoanaerobacter sp. X513 - *Cal. subterraneus* subsp. *tengcongensis* MB4	0.711	0.790	0.748	0.830	0.232	0.294	Different
Thermoanaerobacter sp. X513 - *Th. thermosaccharolyticum* DSM 571	0.646	0.728	0.652	0.774	-0.006	0.101	Different
Thermoanaerobacter sp. X561 - *Cal. subterraneus* subsp. *tengcongensis* MB4	0.711	0.790	0.748	0.830	0.232	0.294	Different
Thermoanaerobacter sp. X561 - *Th. thermosaccharolyticum* DSM 571	0.646	0.728	0.657	0.774	0.003	0.101	Different
Cal. subterraneus subsp. *tengcongensis* MB4 - *Th. thermosaccharolyticum* DSM 571	0.635	0.716	0.651	0.763	-0.018	0.064	Different

[a]Abbreviations: *T*. = *Thermoanaerobacter*; *Th. = Thermoanaerobacterium; Cal*. = *Caldanaerobacter.*

[b]Three-gene model involving *recN*, *rpoA* and *thdF*.

[c]One gene *cpn60* UT model.

[d]ND = Not determined due to a lack of available sequence for comparison.

[e]Unable to determine a high quality *rpoA* or *thdF* sequence for isolate WC7.

[f]Unable to detemine a high quality sequence of *recN* for WC11.

[e]Modified from original publication (Chapter 2, footnote). Updated data for *T. mathranii* subsp. *mathranii* A3 and *T. wiegelii* Rt8.B1 included in here due to the availability of new sequence data.

Table A.2.3. Prediction of genome relatedness[a] based on sequence identity scores for the *Thermoanaerobacter* isolates using Zeigler's one-gene *recN* model.

Pair	SI *recN*	One-gene model (*recN*)	Species prediction
Isolates as a group (low – WC1-WC6)[b]	0.992	0.775	Same species
Isolates as a group (high – WC1-WC2)	1.000	0.950	Same species
T. thermohydrosulfuricus DSM 567-isolates (low – WC6)	0.989	0.925	Same species
T. thermohydrosulfuricus DSM 567-isolates (high -- WC6)	0.997	0.943	Same species
T. brockii subsp. *brockii* HTD4-isolates (low -- WC3)	0.966	0.874	Same species
T. brockii subsp. *brockii* HTD4-isolates (high -- WC1)	0.969	0.880	Same species
T. pseudethanolicus 39E-isolates (low -- WC3)	0.966	0.874	Same species
T. pseudethanolicus 39E-isolates (high -- WC1)	0.969	0.880	Same species

[a]A cutoff of 70% genome sequence identity was used to determine species identity, consistent with Zeigler (2003).

[b]"low" and "high" refer to the extremes of the ranges of sequence identities observed in the groups of sequences compared, based on the sequence identities of the *recN*. The isolate(s) that exhibited these ranges are indicated.

Table A.2.4. Application of the three-gene model of Zeigler (2003) to the strain comparisons of Richter and Roselló-Móra (2009).

Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	Three-gene model	Species prediction
Phylum: *Firmicutes*					
Bacillus cereus ATCC 10987 - *Bacillus cereus* ATCC 14579^T	0.937	0.996	0.924	0.769	Same species
Streptococcus agalactiae 2603 V/R - *Streptococcus agalactiae* A909	0.999	0.998	1.000	0.939	Same species
Streptococcus agalactiae 2603 V/R - *Streptococcus agalactiae* NEM316	0.999	0.999	0.994	0.929	Same species
Streptococcus agalactiae A909 - *Streptococcus agalactiae* NEM 316	1.000	0.999	0.994	0.929	Same species
Phylum: *Bacteroidetes/Chlorobi*					
Bacteroides fragilis NCTC 9343 - *Bacteroides fragilis* YCH46	0.994	0.997	0.994	0.925	Same species
Bacteroides fragilis NCTC 9343 - *Bacteroides thetaiotomicron* VPI-5482	0.739	0.915	0.754	0.322	Different species
Bacteroides thetaiotomicron VPI-5482 - *Bacteroides fragilis* YCH46	0.739	0.916	0.753	0.320	Different species
Phylum: *Alphproteobacteria*					
Brucella canis ATCC 23365 - *Brucella melitensis* ATCC 23457 biovar 2	0.998	0.994	0.997	0.930	Same species
Brucella canis ATCC 23365 - *Brucella ovis* ATCC 25840	0.998	0.995	0.994	0.926	Same species
Brucella canis ATCC 23365 - *Brucella suis* 1330	1.000	0.999	0.998	0.937	Same species
Brucella melitensis ATCC 23457 biovar 2 - *Brucella ovis* ATCC 25840	0.997	0.993	0.997	0.930	Same species
Brucella melitensis ATCC 23457 biovar 2 - *Brucella suis* 1330	0.998	0.995	0.998	0.934	Same species
Brucella ovis ATCC 25840 - *Brucella suis* 1330	0.998	0.996	0.995	0.929	Same species
Phylum: *Betaproteobacteria*					
Nitrosomonas europea ATCC 19718 - *Nitrosomonas eutropha* C91	0.775	0.806	0.804	0.373	Different species
Phylum: *Gammaproteobacteria*					
Escherichia coli CFT073 - *Escherichia coli* E2348/69	0.996	1.000	0.996	0.931	Same species
Escherichia coli CFT073 - *Escherichia coli* K12 MG1655	0.970	1.000	0.950	0.836	Same species
Escherichia coli CFT073 - *Escherichia coli* O157:H7 EDL933	0.972	1.000	0.947	0.832	Same species

Table A.2.4 cont.

Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	Three-gene model	Species prediction
Escherichia coli CFT073 - *Escherichia coli* 042	0.970	1.000	0.958	0.849	Same species
Escherichia coli CFT073 - *Shigella flexneri* 2a 2457T	0.969	0.998	0.953	0.839	Same species
Escherichia coli CFT073 - *Shigella sonnei* 53G	ND[a]	ND[a]	ND[a]	ND[a]	ND[a]
Pseudomonas aeruginosa PAO1 - *Pseudomonas fluorescens* Pf-5	0.785	0.877	0.786	0.383	Different species
Pseudomonas aeruginosa PAO1 - *Pseudomonas fluorescens* PfO1	0.761	0.879	0.762	0.329	Different species
Pseudomonas aeruginosa PAO1 - *Pseudomonas fluorescens* SBW25	0.754	0.877	0.753	0.308	Different species
Pseudomonas aeruginosa PAO1 - *Pseudomonas syringae* pv. Syringae B728a	0.760	0.865	0.768	0.333	Different species
Pseudomonas aeruginosa PAO1 - *Pseudomonas syringae* pv. Tomato	0.750	0.872	0.753	0.304	Different species
Pseuromonas fluorescens PfO1 - *Pseudomonas fluorescens* Pf-5	0.831	0.982	0.822	0.527	Different species
Pseuromonas fluorescens PfO1 - *Pseudomonas fluorescens* SBW25	0.812	0.977	0.807	0.487	Different species
Shewanella amazonensis SB2B - *Shewanella oneidensis* MR-1	0.702	0.873	0.755	0.282	Different species
Shewanella amazonensis SB2B - *Shewanella putrefaciens* CN-32	0.676	0.887	0.728	0.228	Different species
Shewanella amazonensis SB2B - *Shewanella woodyi* ATCC 51908	0.672	0.848	0.731	0.212	Different species
Shewanella baltica OS155 - *Shewanella baltica* OS185	0.961	0.998	0.982	0.888	Same species
Shewanella baltica OS155 - *Shewanella baltica* OS195	0.970	0.999	0.922	0.786	Same species
Shewanella baltica OS155 - *Shewanella baltica* OS223	0.973	0.999	0.925	0.792	Same species
Shewanella baltica OS155 - *Shewanella putrefaciens* CN-32	0.911	0.968	0.814	0.546	Different species
Shewanella baltica OS155 - *Shewanella oneidensis* MR-1T	0.794	0.956	0.808	0.468	Different species
Phylum: *Thermotogae*					
Thermotoga maritima MSB8 - *Thermotoga neapolitana* DSM 4359	ND[b]	0.781	0.751	ND	ND
Thermotoga maritima MSB8 - *Thermotoga petrophila* RKU-1	ND[b]	0.943	0.948	ND	ND

Table A.2.4 cont.

Pair	SI_{recN}	SI_{rpoA}	SI_{thdF}	Three-gene model	Species prediction
Thermotoga neapolitana DSM 4359 - *Thermotoga petrophila* RKU-1	ND[b]	0.783	0.755	ND	ND
Phylum: *Deinococcus-Thermus*					
Deinococcus deserti VCD115 - *Deinococcus radiodurans* R1	0.695	0.825	0.720	0.193	Different species

[a]ND, not determined since genome sequence for *S. sonnei* 53G not available in GenBank at time of analysis.

[b]ND, not determined as no *recN* orthologue identified in *Thermotoga* spp.

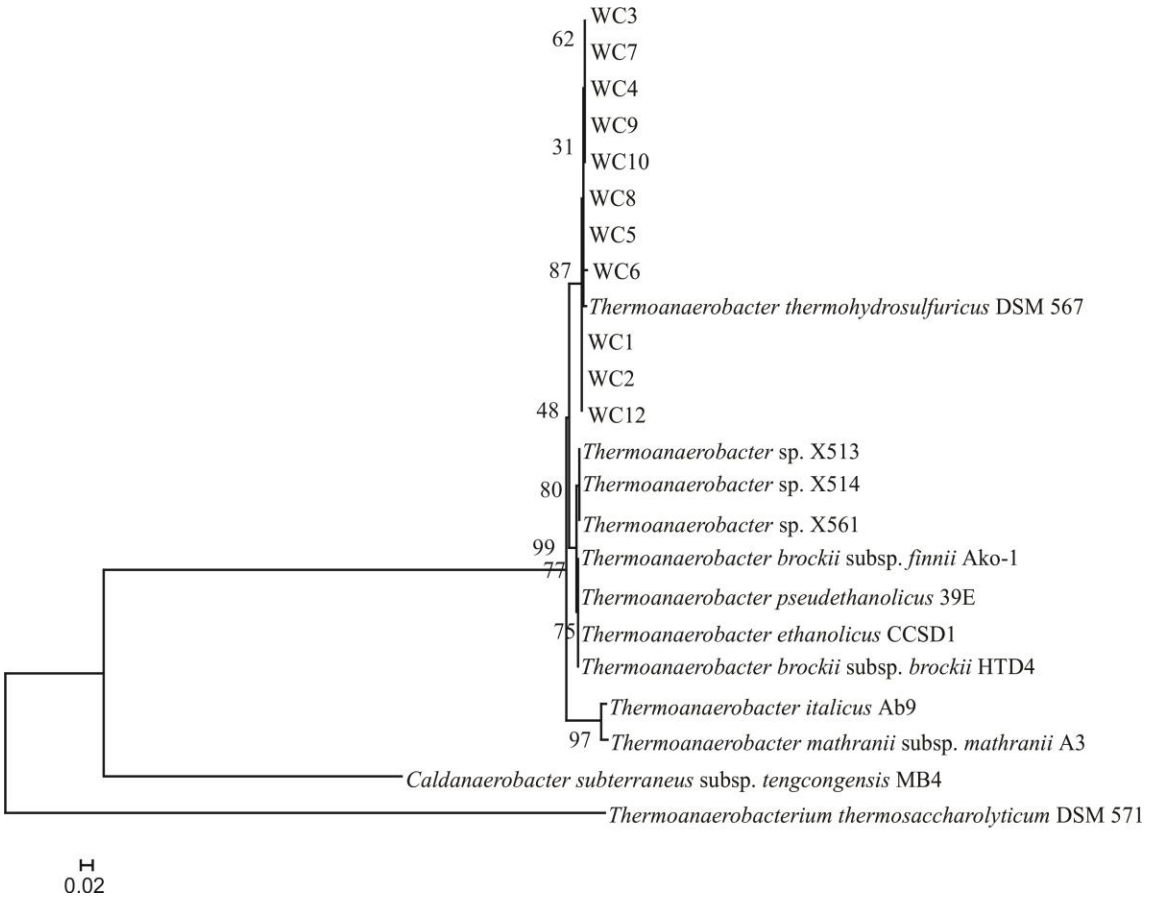


Fig. A.2.1. Neighbour-joining tree of *recN* sequences from *Thermoanaerobacter* spp. Tree was constructed as described in Chapter 3.3.3. Scale bar represents the number of changes per nucleotide.

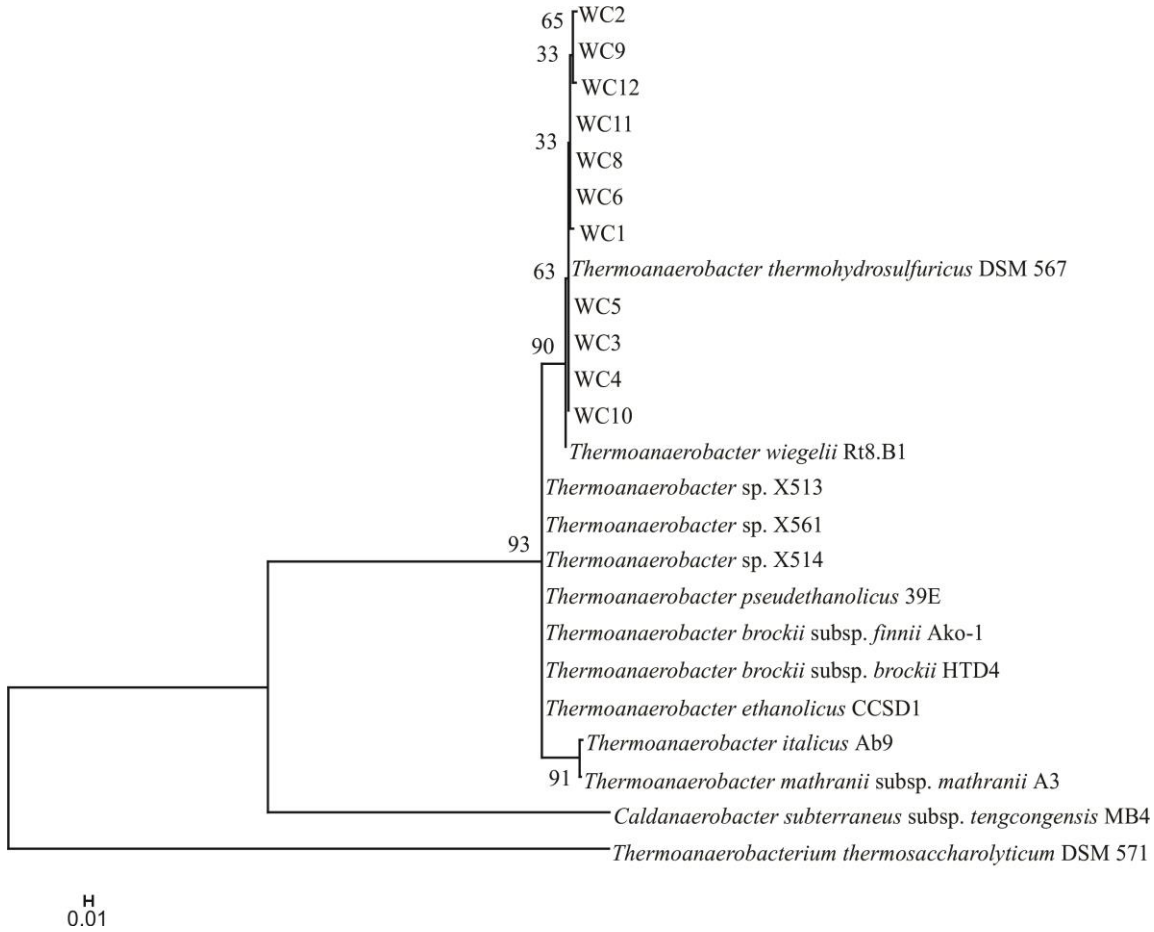


Fig. A.2.2. Neighbour-joining tree of *rpoA* sequences from *Thermoanaerobacter* spp. Tree was constructed as described in Chapter 3.3.3. Scale bar represents the number of changes per nucleotide.

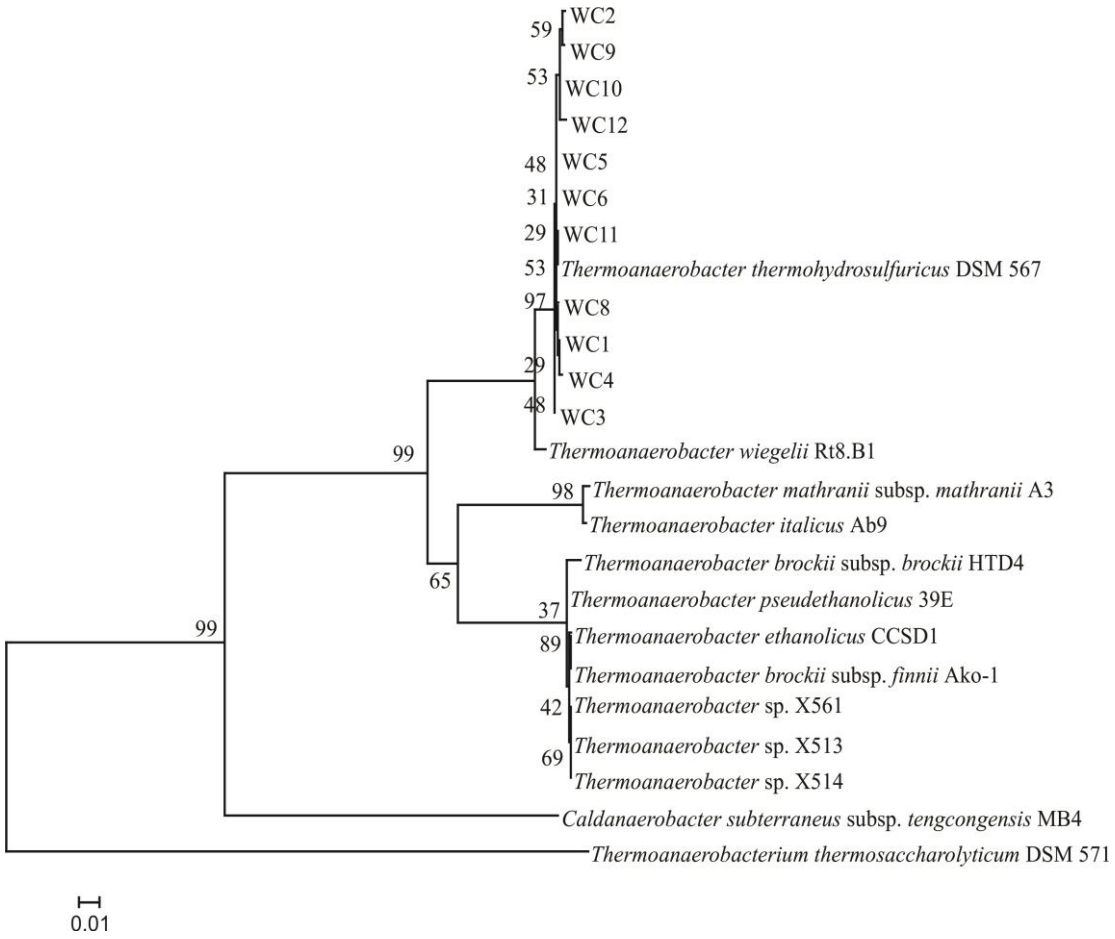


Fig. A.2.3. Neighbour-joining tree of *thdF* sequences from *Thermoanaerobacter* spp. Tree was constructed as described in Chapter 3.3.3. Scale bar represents the number of changes per nucleotide.

A.3 Supplemental material for Chapter 4: Genomic evaluation of *Thermoanaerobacter* spp. for the construction of designer co-cultures to improve lignocellulosic biofuel production.

A.3.1 Supplemental results

Table A.3.1. Selected genome metadata for sequenced *Thermoanaerobacter* spp.

Clade	Genome	Status[a]	Genome Size (Mb)	% G+C	# of CDS	Locus Tag Prefix
1	*T. brockii* subsp. *finnii* Ako-1	F	2.34	34.50	2,336	Thebr_
	T. ethanolicus CCSD1	D	2.20	34.24	2,367	TeCCSD1DRAFT_
	T. pseudethanolicus 39E	F	2.36	34.51	2,291	Teth39_
	Thermoanaerobacter sp. X513	F	2.45	34.52	2,462	Thet_
	Thermoanaerobacter sp. X514	F	2.45	34.52	2,397	Teth514_
	Thermoanaerobacter sp. X561	D	2.38	34.32	2,423	Teth561DRAFT_
2	*T. italicus* Ab9	F	2.45	34.15	2,407	Thit_
	T. mathranii subsp. *mathranii* A3	F	2..30	34.32	2,286	Tmath_
3	*T. siderophilus* SR4	P	2.54	34.23	2,557	ThesiDRAFT1_
	T. thermohydrosulfuricus WC1	PD	2.57	34.35	2,552	TthWC1_
	T. wiegelii Rt8.B1	F	2.78	34.34	2,800	Thewi_

[a]Sequencing status at time of analyses is designated as either finished (F), draft (D) or permanent draft (P).

Table A.3.2. All CAZyme designated gene sequences within sequenced *Thermoanaerobacter* strains as are available within the CAZy database or identified through *de novo* analysis.

The size of the table prohibits its inclusion in this thesis in printable format. An electronic version of Table A.3.2 can be downloaded in Excel format (.xls) at:

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0059362#s4

Note: the electronic and published version of Table A.3.2 is titled "Table S2."

Table A.3.3. Identified genes associated with the utilization of the major carbohydrates produced through lignocellulose hydrolysis in sequenced *Thermoanaerobacter* strains.

The size of the table prohibits its inclusion in this thesis in printable format. An electronic version of Table A.3.3 can be downloaded in Excel format (.xls) at:

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0059362#s4

Note: the electronic and published version of Table A.3.3 is titled "Table S3."

Table A.3.4. Genes associated with pyruvate catabolism in sequenced *Thermoanaerobacter* strains.

The size of the table prohibits its inclusion in this thesis in printable format. An electronic version of Table A.3.4 can be downloaded in Excel format (.xls) at:

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0059362#s4

Note: the electronic and published version of Table A.3.4 is titled "Table S4."

Table A.3.5. Genes associated with transmembrane ion gradient generating and consuming reactions involved with cellular energetics in sequenced *Thermoanaerobacter* genomes.

The size of the table prohibits its inclusion in this thesis in printable format. An electronic version of Table A.3.5 can be downloaded in Excel format (.xls) at:

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0059362#s4

Note: the electronic and published version of Table A.3.5 is titled "Table S5."

Table A.3.6. Key amino acid residues responsible for imparting predicted substrate specificity and K^+ dependence in annotated *Thermoanaerobacter* V-type pyrophosphatases.

		Amino acid positions					
		Cation export specificity			K^+ dependence		
Clade	Genome (locus tag)	180	242	246	478	481	Predicted specificity[a]
1	*T. brockii* subsp. *finnii* Ako-1 (Thebr_1608)	S	E	G	A	G	Na^+ export; K^+ dependent
	T. brockii subsp. *finnii* Ako-1(Thebr_1629)	S	E	G	A	G	Na^+ export; K^+ dependent
	T. ethanolicus CCSD1 (TeCCSD1DRAFT_1224)	S	E	G	A	G	Na^+ export; K^+ dependent
	T. pseudethanolicus 39E (Teth39_1572)	S	E	G	A	G	Na^+ export; K^+ dependent
	T. pseudethanolicus 39E (Teth39_1590)	S	E	G	A	G	Na^+ export; K^+ dependent
	Thermoanaerobacter sp. X513 (Thet_0681)	S	E	G	A	G	Na^+ export; K^+ dependent
	Thermoanaerobacter sp. X514 (Teth514_2253)	S	E	G	A	G	Na^+ export; K^+ dependent
	Thermoanaerobacter sp. X561 (Teth561DRAFT_0879)	S	E	G	A	G	Na^+ export; K^+ dependent
2	*T. italicus* Ab9 (Thit_0710)	S	E	G	A	G	Na^+ export; K^+ dependent
	T. mathranii subsp. *mathranii* A3 (Tmath_0761)	S	E	G	A	G	Na^+ export; K^+ dependent
3	*T. siderophilus* SR4 (ThesiDRAFT1_0299)	S	E	A	K	T	H^+ export; K^+ independent
	T. thermohydrosulfuricus WC1 (TthWC1_1224)	S	E	A	K	T	H^+ export; K^+ independent
	T. wiegelii Rt8.B1 (Thewi_00010550)	S	E	A	K	T	H^+ export; K^+ independent

[a]Specificity inferred based upon the determination of key amino acid residues as described (Luoto *et al*., 2011).

Table A.3.7. Identification of key residues characteristic of ATP or PPi dependent 6-phosphofructokinase genes in *Thermoanaerobacter*.

		Amino acid positions[a]	
Clade	Strain (gene)	104	124
PPi dependent			
1	*T. brockii* subsp. *finnii* Ako-1 (Thebr_0507)	D	K
	T. ethanolicus CCSD1 (TeCCSD1DRAFT_0939)	D	K
	T. pseudethanolicus 39E (Teth39_0494)	D	K
	Thermoanaerobacter sp. X513 (Thet_1945)	D	K
	Thermoanaerobacter sp. X514 (Teth514_0971)	D	K
	Thermoanaerobacter sp. X561 (Teth561DRAFT_1398)	D	K
3	*T. siderophilus* SR4 (ThesiDRAFT1_0659)	D	K
	T. thermohydrosulfuricus WC1 (TthWC1_2408)	D	K
	T. wiegelii Rt8.B1 (Thewi_00000380)	D	K
ATP dependent			
1	*T. brockii* subsp. *finnii* Ako-1 (Thebr_0701)	G	G
	T. ethanolicus CCSD1 (TeCCSD1DRAFT_1787)	G	G
	T. pseudethanolicus 39E (Teth39_0683)	G	G
	Thermoanaerobacter sp. X513 (Thet_1716)	G	G
	Thermoanaerobacter sp. X514 (Teth514_1194)	G	G
	Thermoanaerobacter sp. X561 (Teth561DRAFT_0551)	G	G
2	*T. italicus* Ab9 (Thit_1630)	G	G
	T. mathranii subsp. *mathranii* A3 (Tmath_1620)	G	G
3	*T. siderophilus* SR4 (ThesiDRAFT1_0426)	G	G
	T. thermohydrosulfuricus WC1 (TthWC1_1707)	G	G
	T. wiegelii Rt8.B1 (Thewi_00026650)	G	G
	T. siderophilus SR4 (ThesiDRAFT1_0274)	G	A
	T. thermohydrosulfuricus WC1 (TthWC1_0878)	G	A
	T. wiegelii Rt8.B1 (Thewi_00010820)	G	A
Unknown[b]			
2	*T. italicus* Ab9 (Thit_0261)	G	L

[a]Positions given in reference to *E. coli* phopshofructokinase numbering.

[b]Specific combination not identified by Bapteste *et al*. (2003); dependence can therefore not be inferred.

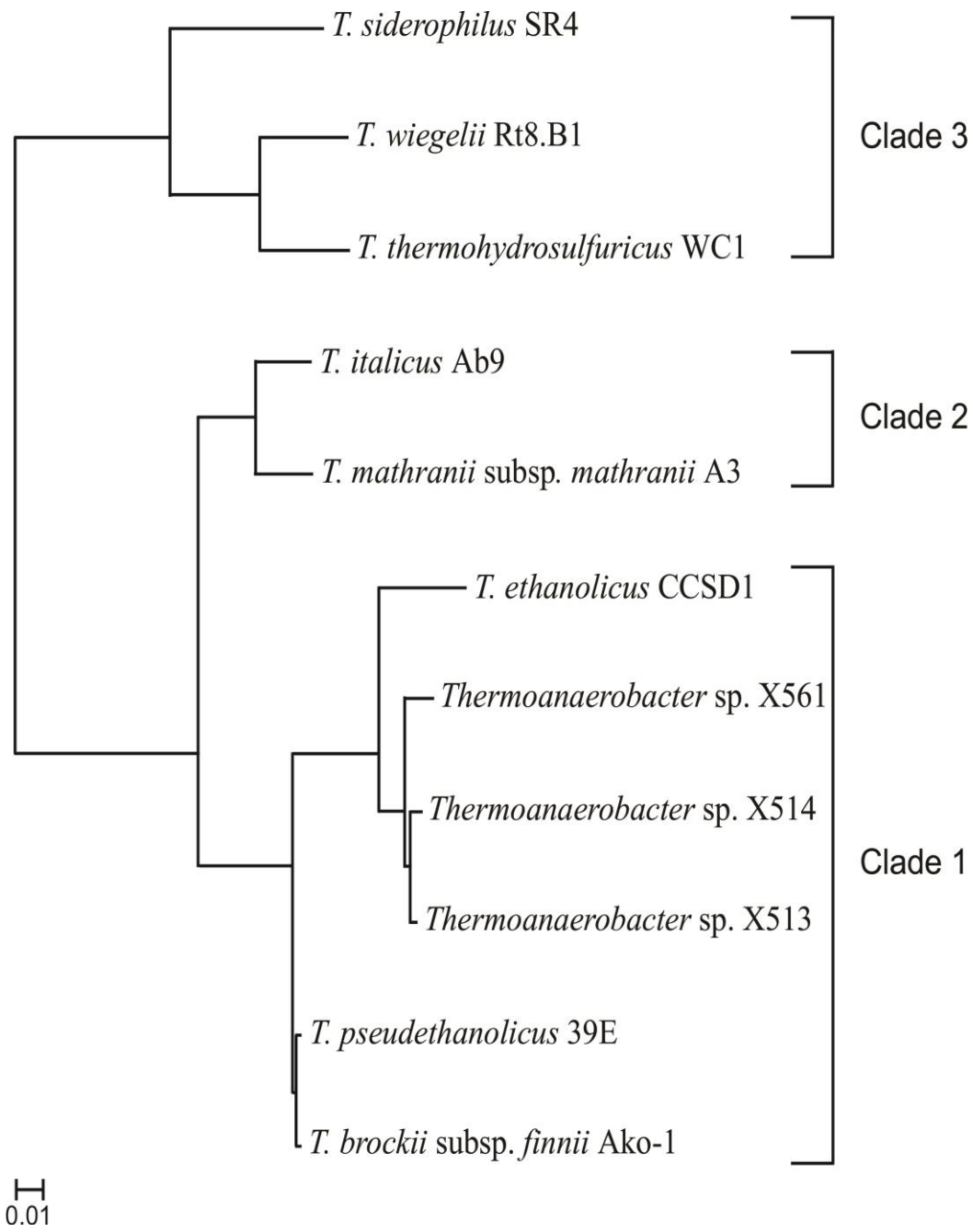


Fig. A.3.1. Phylogram of annotated KO functional profiles for sequenced *Thermoanaerobacter* strains. Cluster analysis performed as described in Fig. 4.2.

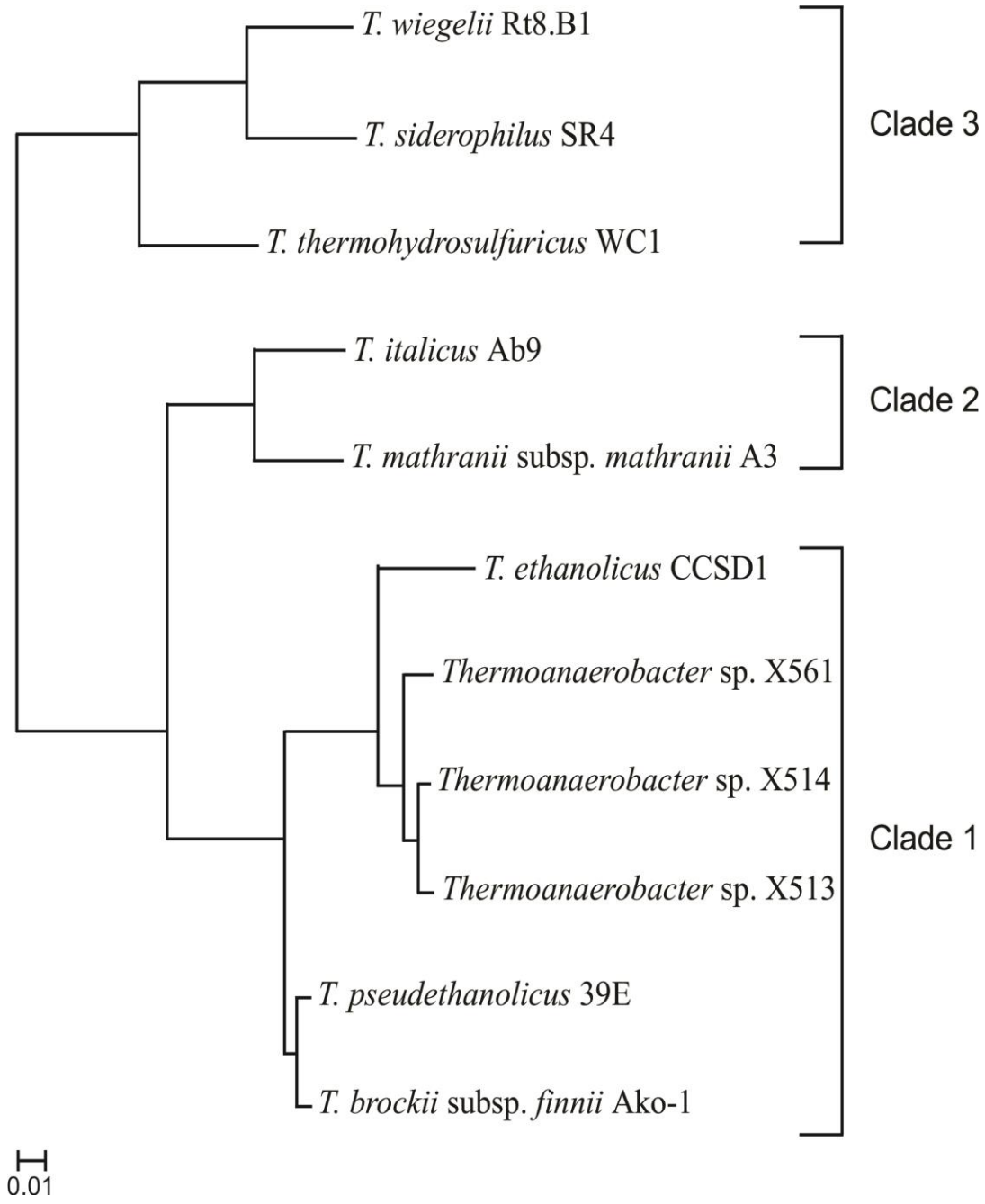


Fig. A.3.2. Phylogram of annotated TIGRFAM functional profiles for sequenced *Thermoanaerobacter* strains. Cluster analysis performed as described in Fig. 4.2.

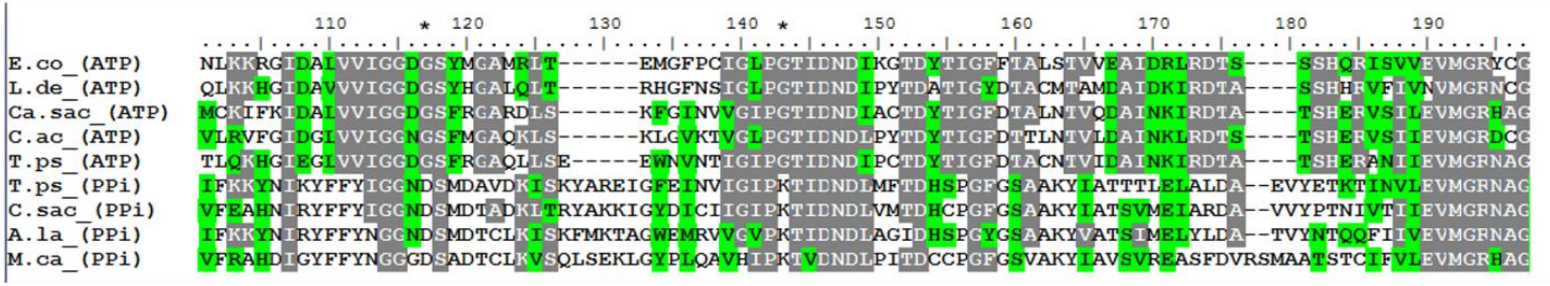


Fig. A.3.3. Partial sequence alignment of selected PFK genes in different bacteria. Sequences are similar to those chosen by Bielen *et al*. (2010). The PFK sequences annotated in *T. pseudethanolicus* 39E were chosen as representative examples of PPi and ATP-dependent PFK genes within the genus *Thermoanaerobacter*. Colour scheme identifies identical residues (grey), similar residues (green) or dissimilar residues (white). Abbreviations are as follows: E.co (*Escherichia coli*), L.de (*Lactobacillus delbrueckii*), C.sa (*Ca. saccharolyticus*), C.ac (*Clostridium acetobutylicum*), *T.ps* (*T. pseudethanolicus* 39E), A.la (*Archoleplasma laidlawii*), M.ca (*Methylococcus capsulatus*). Marked positions (*) indicate conserved residues important in determining PPi or ATP dependence. Amino acid signatures for all *Thermoanaerobacter* strains are available in Table A.3.7.

A.4 Supplemental material for Chapter 5: Metabolic and label-free quantitative proteomics analysis of *Thermoanaerobacter thermohydrosulfuricus* WC1 on single and mixed substrates

A.4.1 Supplemental results

Table A.4.1. Potential *cre* sequences identified within the *T. thermohydrosulfuricus* WC1 genome.

Genomic scaffold[a]	Start site	Stop site	Sequence	Corresponding genes	Position[b]
NZ_KB731277.1	35388	35403	ATGGAAACGAACTCAA	TthWC1_0041	IS
NZ_KB731278.1	5297	5312	TTGAAACCGATGGCAC	TthWC1_0306	IO
NZ_KB731278.1	39064	39079	TTGTAAGCGCTTACAG	TthWC1_0339-TthWC1_0340	IN
NZ_KB731279.1	50905	50920	ATGGAATCGTAATCAA	TthWC1_0568-TthWC1_0569	IN
NZ_KB731279.1	167813	167828	ATGAAACCGATTGCAT	TthWC1_0680-TthWC1_0681	IN
NZ_KB731279.1	167815	167830	ATGCAATCGGTTTCAT	TthWC1_0680-TthWC1_0681	IN
NZ_KB731281.1	15244	15259	TTGAAAACGAATGCAA	TthWC1_0892	IS
NZ_KB731282.1	79259	79274	ATGAAAACGGAAACAC	TthWC1_1061	IS
NZ_KB731283.1	3220	3235	TTGCAAACGGTGTCTC	TthWC1_1101	IS
NZ_KB731284.1	39411	39426	ATGCAATCGAATGCAT	TthWC1_1236-TthWC1_1237	IN
NZ_KB731284.1	39497	39512	ATGCAACCGAATGCAT	TthWC1_1236-TthWC1_1237	IN
NZ_KB731285.1	96678	96693	ATGTAATCGGTAACTA	TthWC1_1375-TthWC1_1376	IN
NZ_KB731285.1	96866	96881	ATGTAATCGGTAACTA	TthWC1_1375-TthWC1_1376	IN
NZ_KB731285.1	97075	97090	ATGTAATCGGTAACTA	TthWC1_1376-End of scaffold	IN
NZ_KB731285.1	56410	56425	TTGCAAACGTTGTCTT	TthWC1_1337	IS
NZ_KB731287.1	61902	61917	TTGGAACCGGATACTT	TthWC1_1531	IO
NZ_KB731288.1	61666	61681	TTGCAAGCGCATACAA	TthWC1_1609	IO
NZ_KB731289.1	63313	63328	ATGGAAACGTAAACTA	TthWC1_1684	IS
NZ_KB731290.1	21049	21064	TTGGAATCGTAGTCAC	TthWC1_1710	IO
NZ_KB731294.1	28092	28108	GAGAAAAAGTTTTCTTG[c]	TthWC1_1946-TthWC1_1947	IN
NZ_KB731298.1	1686	1701	ATGAAAACGATAACTC	TthWC1_2104	IO

Table A.4.1 cont.

Genomic scaffold[a]	Start site	Stop site	Sequence	Corresponding genes	Position[b]
NZ_KB731303.1	24223	24238	TTGAAAACGTAATCAT	TthWC1_2309	IS
NZ_KB731304.1	3655	3670	ATGAAAGCGTATTCTC	TthWC1_2320	IS
NZ_KB731309.1	6547	6562	ATGAAACCGCTGCCAT	TthWC1_2468	IO

[a]GenBank accession number.

[b]Relative to gene sequence. Abbreviations: IS = internal to coding sequence, same strand as annotated gene; IO = internal to coding sequence, opposite strand to annotated gene; IN = intergenic region between annotated genes.

[c]Sequence homologous to *C. difficile* consensus sequence (Antunes *et al*., 2012).

Table A.4.2. Calculated mass balances for *T. thermohydrosulfuricus* WC1 on 5 g/L xylose, 5 g/L cellobiose or 5 g/L xylose plus 5 g/L cellobiose.

	% Carbon recovery			O/R[a]		
Target optical density	Xylose	Cellobiose	Xylose + cellobiose	Xylose	Cellobiose	Xylose + cellobiose
Post-inoculation[b]	NC[c]	NC	NC	NC	NC	NC
0.10	78.76	110.31	81.08	0.78	0.84	0.77
0.20	84.97	83.73	80.38	0.82	0.83	0.81
0.40	82.76	95.55	88.34	0.87	0.82	0.82
0.70	92.93	87.49	89.44	0.84	0.82	0.88
0.80	91.67	85.11	91.81	0.85	0.83	0.82
0.85	88.46	89.60	85.45	0.86	0.82	0.86

[a]O/R = oxidized end-products formed/ reduced end-products formed.

[b]Post-inoculation refers to cultures sacrificed immediately after inoculation from a xylose-grown parent culture.

[c]NC = Not calculable. Values used for calculation represent net production or net consumption and are corrected for carryover from inoculation.

Table A.4.3. Summary of the identified peptides and proteins resulting from proteomic analysis of *T. thermohydrosulfuricus* WC1 on 5 g/L xylose, 5 g/L cellobiose or 5 g/L xylose plus 5 g/L cellobiose.

		Protein expectation value					
Substrate	Culture	log(e) < -10	log(e) < -3	log(e) < -1	Non-redundant peptides	Identified peptides	Total MS/MS spectra
Xylose	A	808	967	1047	5792	18744	33505
	B	815	978	1057	5901	19704	35224
Cellobiose	A	851	996	1074	6130	19747	33349
	B	831	1010	1088	6280	21246	35692
Xylose plus cellobiose	A	832	1007	1071	6136	19635	33672
	B	829	1001	1068	6080	19625	33867

Table A.4.4. Expression data and calculated expression ratios for *T. thermohydrosulfuricus* WC1 on single or mixed substrates.

The size of the table prohibits its inclusion in this thesis in printable format. An electronic version of Table A.4.4 is available with the electronic version of this thesis from the University of Manitoba's MSpace database:

http://mspace.lib.umanitoba.ca/

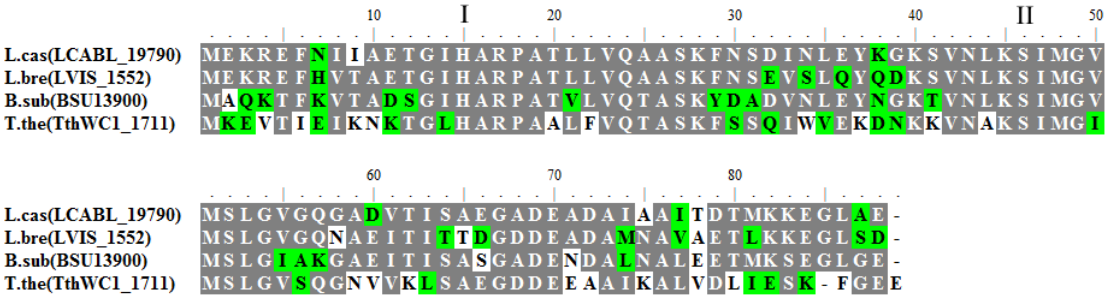


Fig. A.4.1. Alignment of the HPr protein sequence from *T. thermohydrosulfuricus* WC1 against select reference sequences. Reference sequences were chosen from organisms with Ser_{46} residues known to be phosphorylated during CCR (Ye *et al*., 1994; Galinier *et al*., 1999; Viana *et al*., 2000). Positions 15 and 46 (in reference to *B. subtilis*) are marked as (I) or (II) respectively. Locus tags are identified in brackets after the organism name. Shading colour scheme identifies identical residues (grey), similar residues (green) or dissimilar residues (white). Abbreviations: L.cas (*Lactobacillus casei* susbp. *casei* BL23; L.bre (*Lactobacillus brevis* ATCC367); B.sub (*B. subtilis* subsp. *subtilis* 168); T.the (*T. thermohydrosulfuricus* WC1).

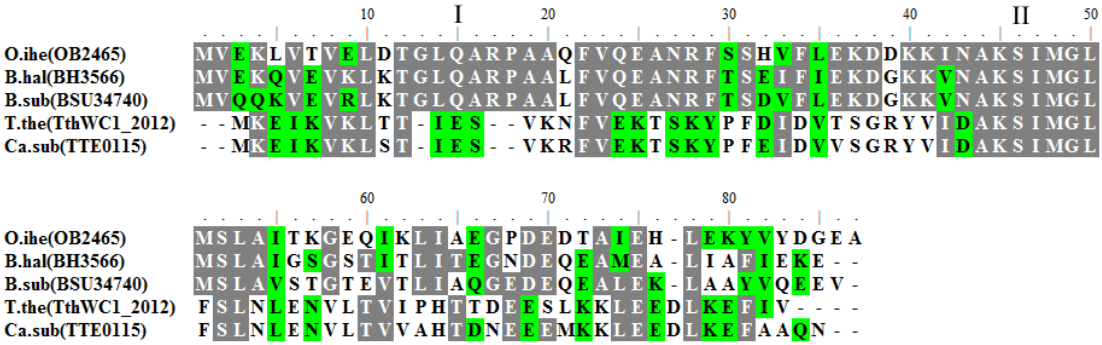


Fig. A.4.2. Alignment of the CrH protein sequence from *T. thermohydrosulfuricus* WC1 against select reference sequences. Reference sequences were chosen from CrH and putative CrH sequences identified by (Warner & Lolkema, 2003). Positions 15 and 46 (in reference to *B. subtilis*) are marked as (I) or (II) respectively. The locus tags are identified in brackets after the organism name. Shading colour scheme identifies identical residues (grey), similar residues (green) or dissimilar residues (white). Abbreviations: O.ihe (*Oceanobacillus iheyensis* HTE831); B.hal (*Bacillus halodurans* C-125); B.sub (*B. subtilis* subsp. *subtilis* 168); T.the (*T. thermohydrosulfuricus* WC1); Ca.sub (*Caldanaerobacter subterraneus* subsp. *tengcongensis* MB4).

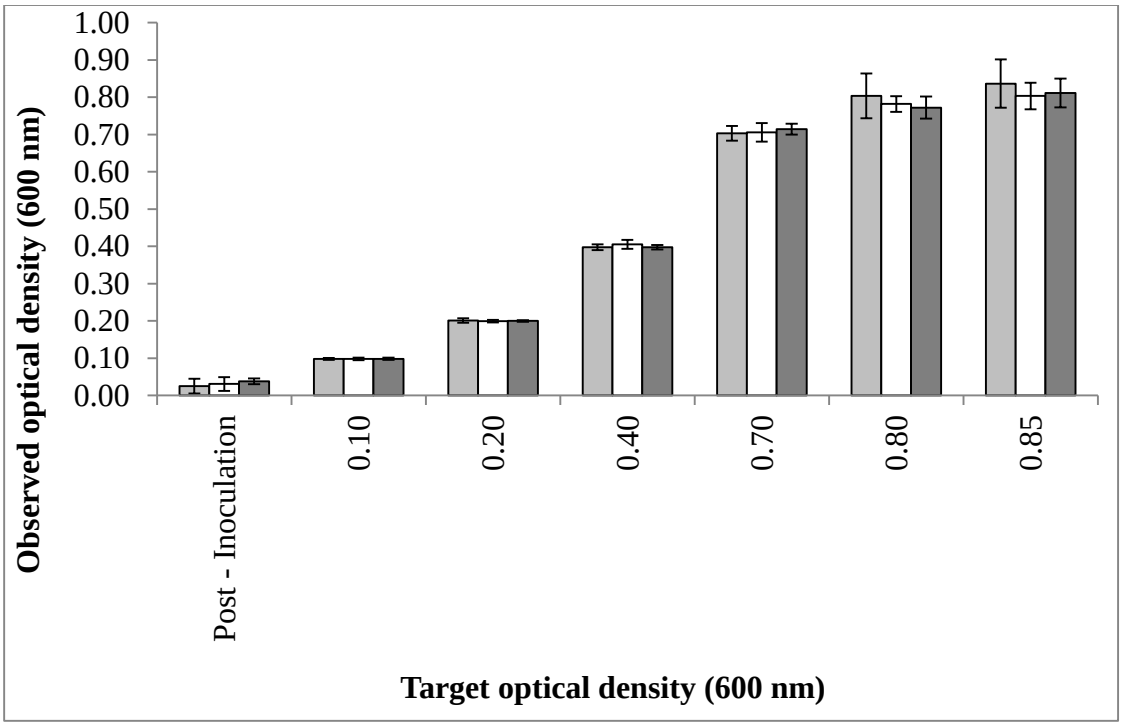


Fig. A.4.3. Comparison of the measured OD_{600} values against the target OD_{600} values. Cultures were grown as described (Chapter 5.3.3) on 5 g/L xylose (light grey), 5 g/L cellobiose (white) or 5 g/L xylose plus 5 g/L cellobiose (dark grey). Post-inoculation refers to cultures sacrificed immediately after inoculation from xylose grown parent cultures.

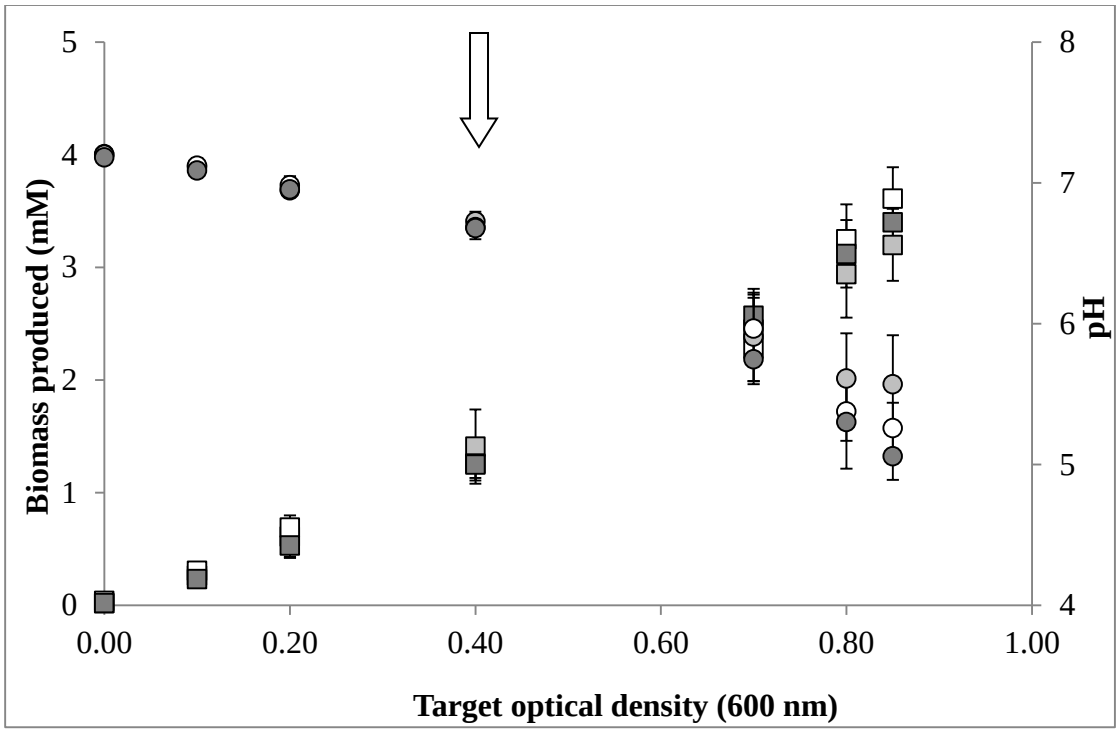
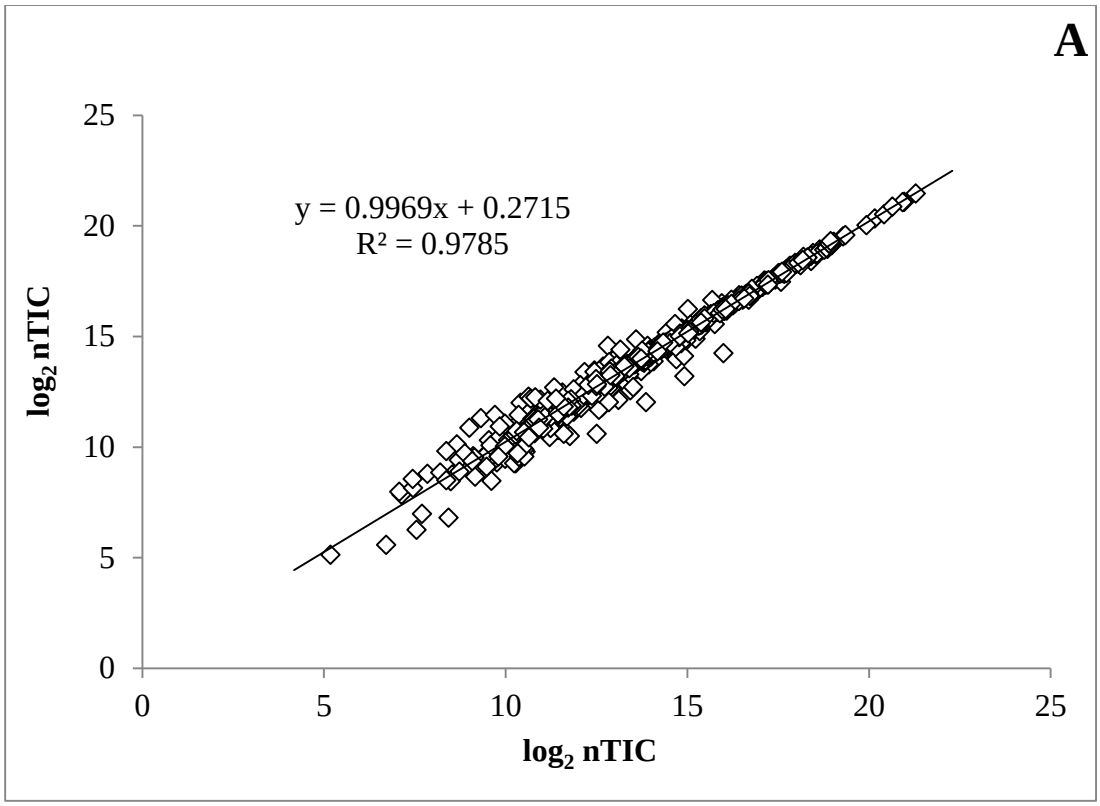
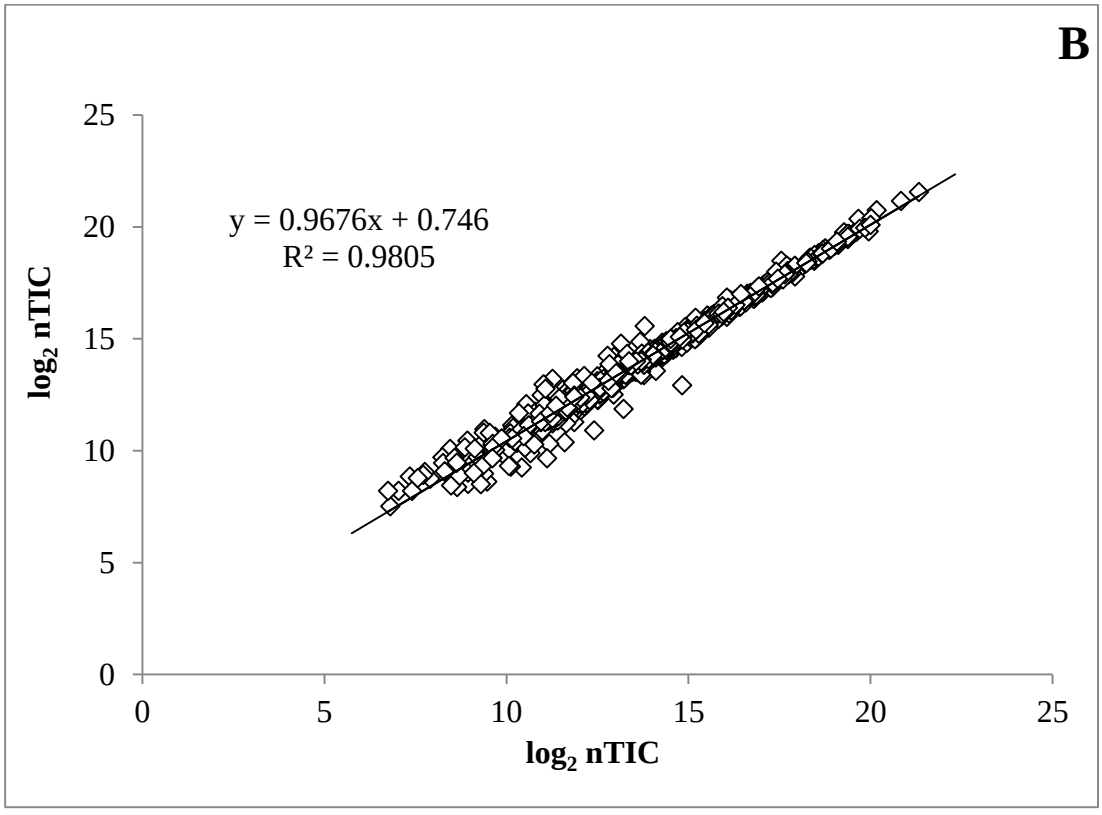


Fig. A.4.4. Biomass synthesis and changes in pH throughout growth of *T. thermohydrosulfuricus* WC1 on single substrate or mixed substrate conditions tested. Growth on 5 g/L xylose (light grey), 5 g/L cellobiose (white) or 5 g/L xylose plus 5 g/L cellobiose (dark grey). Biomass production is represented by squares, while pH measurements are represented by circles. Target optical densities are as described (5.3.3.2). The post-inoculation value is designated 0.00. Error bars represent the standard deviation between replicates. Arrow indicates the physiological point in time used for proteomic analyses.


A
y = 0.9969x + 0.2715
R² = 0.9785
log2 nTIC
log2 nTIC
0
5
10
15
20
25


B
y = 0.9676x + 0.746
R² = 0.9805
log2 nTIC
log2 nTIC
0
5
10
15
20
25

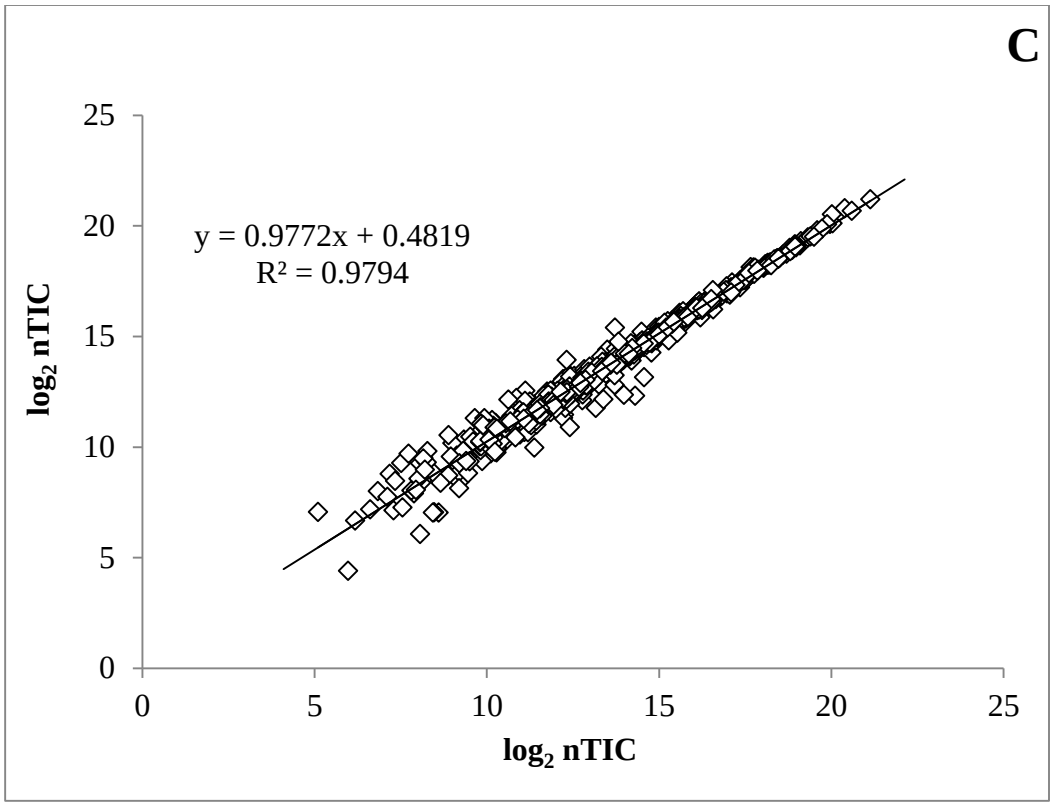
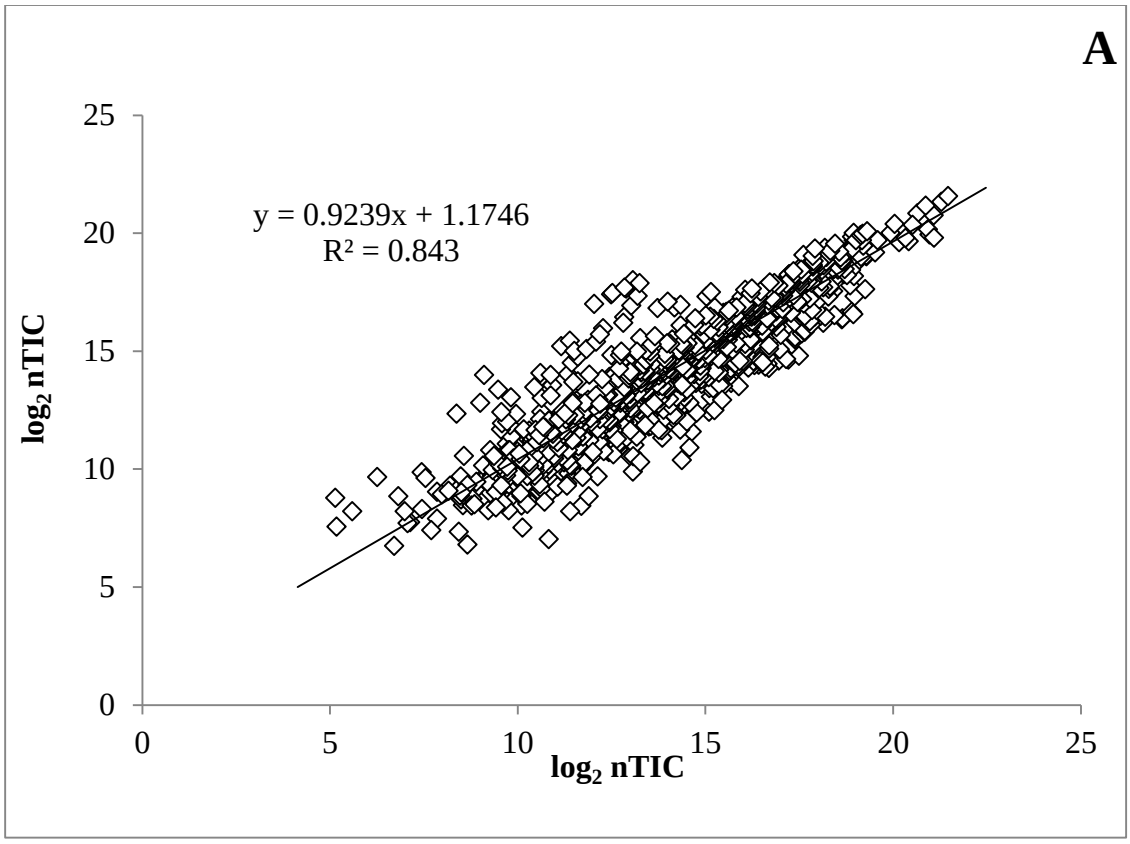
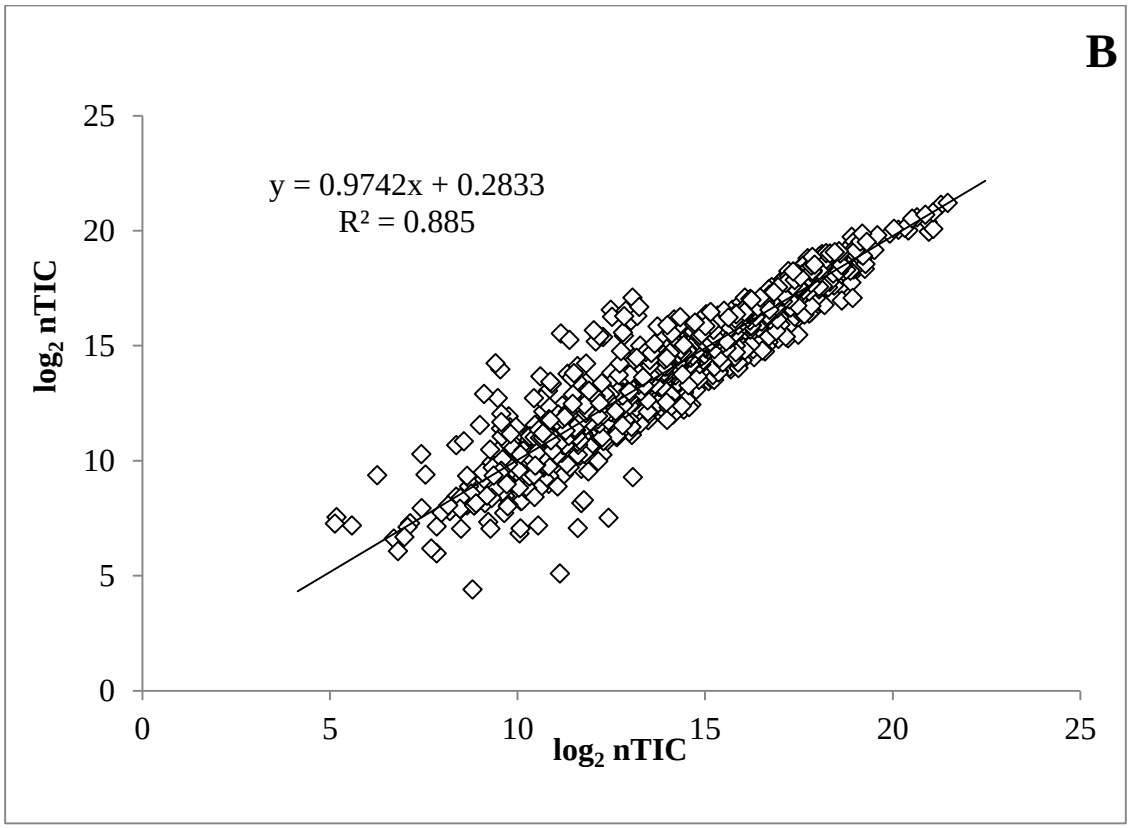


Fig. A.4.5. Linear regression analysis of the $\log_2$ nTIC values for all observed proteins between biological replicates. Values given for replicates grown on 5 g/L xylose (A), 5 g/L cellobiose (B) or 5 g/L xylose plus 5 g/L cellobiose (C). nTIC values observed for Culture A are plotted on the x-axis, while the corresponding nTIC values observed for Culture B are plotted on the y-axis.


A
y = 0.9239x + 1.1746
R² = 0.843
log2 nTIC
log2 nTIC
0
5
10
15
20
25


B
y = 0.9742x + 0.2833
R² = 0.885
log2 nTIC
log2 nTIC
0
5
10
15
20
25

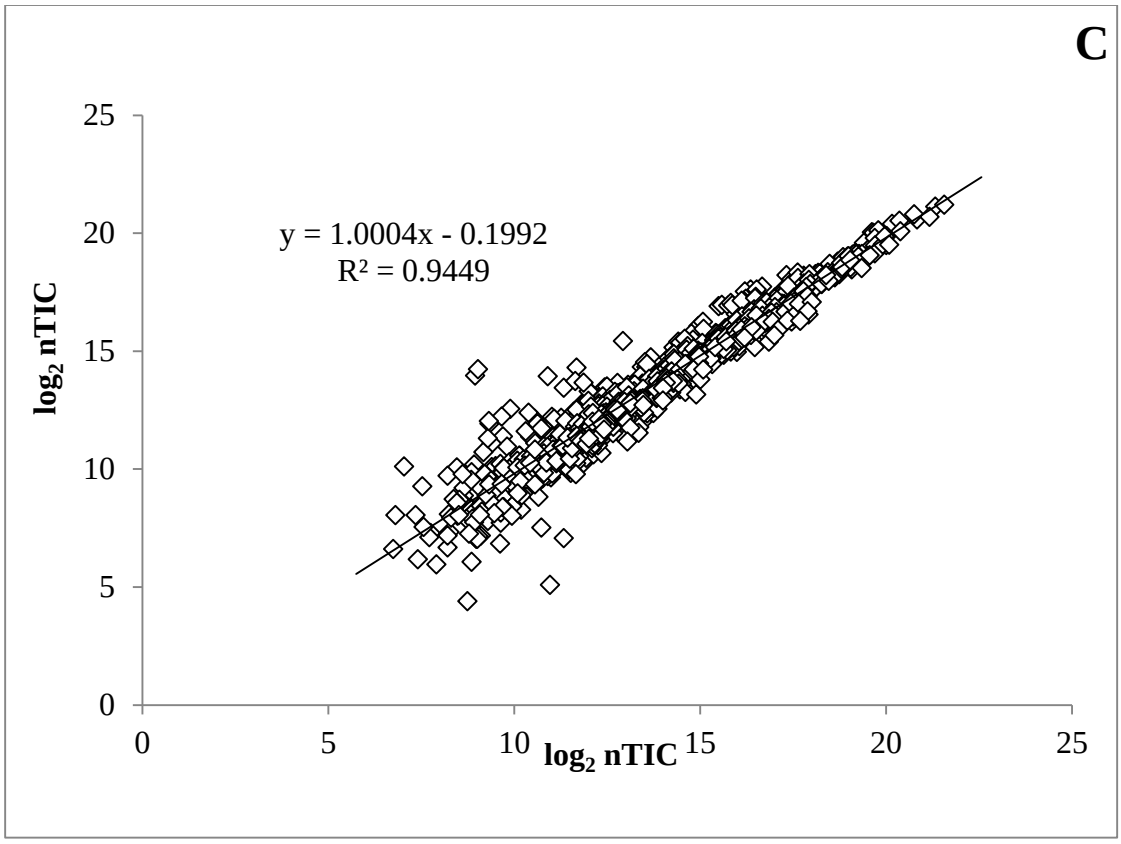


Fig. A.4.6. Linear regression analysis of the $\log_2$ nTIC values for all observed proteins between growth conditions. (A) Xylose grown cells (x-axis) vs. cellobiose grown cells (y-axis). (B) Xylose grown cells (x-axis) vs. sugar mix grown cells (y-axis). (C) Cellobiose grown cells (x-axis) vs. sugar mix grown cells (y-axis). Values plotted along axes are from both Culture A and Culture B.

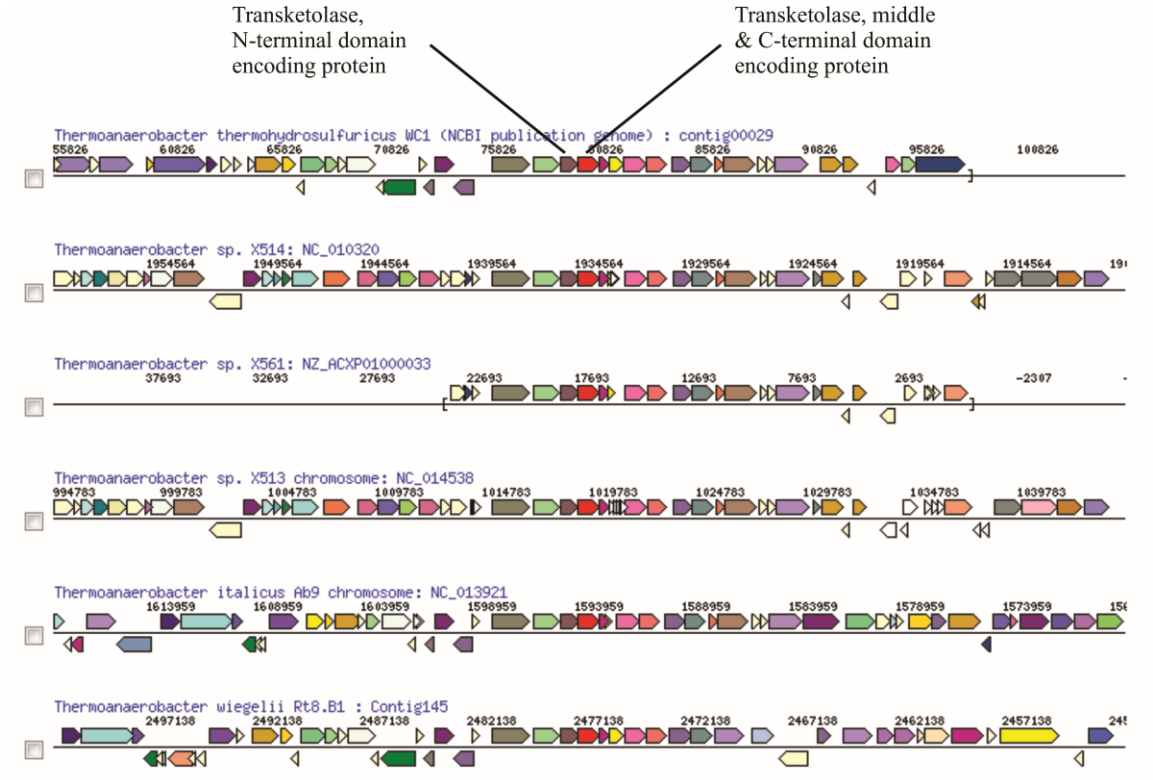


Fig. A.4.7. Illustrative example of transketolase sequences encoded as two separate CDS in sequenced *Thermoanaerobacter* spp. Orthologous sequences aligned with those identified in *T. thermohydrosulfuricus* WC1. Alignment of orthologous sequences conducted using the IMG-ER analysis tool (Markowitz *et al*., 2009).

A.5 Growth and metabolic analyses of *Thermoanaerobacter thermohydrosulfuricus* WC1 on Beechwood xylan.

A.5.1 Introduction

Xylan forms the polymeric backbone of hemicelluloses derived from hardwoods, grasses and cereal crops (Saha, 2003) and can comprise a significant component of the carbohydrate content found in lignocellulosic biomass (Carol & Sommerville, 2009). In CBP platforms, its hydrolysis and subsequent utilization are considered essential for attaining commercially viable second-generation biofuel production (Galbe *et al*., 2007; Gírio *et al*., 2010). Selecting microorganisms capable of these processes is therefore an important component of designing an effective CBP platform.

Within the genus *Thermoanaerobacter*, growth on polymeric xylan has only been reported for a few strains and a link between these strains and the presence of a putative extracellular endoxylanases (GH10 enzymes) encoded within their genomes has been proposed (Chapter 4.4.3.1). Given that the *T. thermohydrosulfuricus* WC1 genome encodes such a putative endoxylanase, its ability to grow on xylan was investigated.

A.5.2 Materials and methods

A.5.2.1 Bacteria and culture conditions

Two independent glycerol stocks of *T. thermohydrosulfuricus* WC1 (DSM 26960), as well as one culture received from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ), were used in this study and treated as one strain. Cultures were grown at 60^{o}C in ATCC 1191 medium as described (Chapter 5.3.1)

and contained either 5 g/L Beechwood xylan (X4252, Sigma-Aldrich, St. Louis, MO) or no substrate. Cultures were passaged every 24 hours on 5 g/L xylan for three times prior to inoculating experimental tubes.

A.5.2.2 Growth and metabolic analyses

For growth and metabolic analyses, three biological replicates were sacrificed at specific time points in each experiment and entire experiments were repeated three independent times. End-product and biomass determinations were identical to those described previously (Chapter 5.3.3.2) with the exception of protein measurements. For protein estimations, 1 ml samples of culture were removed from sacrificed tubes at specific time points and were centrifuged at 13,000 x *g* for 10 min. The resulting pellets were washed in 0.9% (wt/vol) NaCl and then resuspended in 2 M NaOH. After incubating the resuspended cells in a boiling water bath for 10 min, samples were centrifuged for an additional 10 min and the supernatant retained for protein analyses.

To account for potentially contaminating proteins and/or interfering compounds originating from the xylan substrate itself, a xylan extract solution was prepared. In brief, uninoculated tubes of ATCC 1191 medium containing 5 g/L xylan were incubated at 60^{o}C for the duration of the experiment. After 72 hours, the residual xylan was processed for protein analysis identically to inoculated cultures. The solution resulting after boiling the xylan in 2 M NaOH for 10 min was used as the xylan extract solution.

Protein measurements were performed using a Qubit 2.0 fluorometer (Life Technologies, Burlington, ON, Canada). For measurements, protein samples were diluted 1: 20 in xylan extract solution (total volume 200 µl) and the fluorescent signal

measured. Measurements were compared against a standard curve created using known amounts of bovine serum albumin dissolved in xylan extract solution. All samples, as well as all standards, represent average values calculated from three technical replicates.

A.5.3 Results and discussion

Growth curves of *T. thermohydrosulfuricus* WC1 were conducted to evaluate its ability to utilize polymeric xylan as a fermentable substrate. Biomass production increased for 32 hours post-inoculation (Fig. A.5.1). This was in contrast to the no-substrate control condition, which showed only weak growth in the basal medium. Growth by the no-substrate control is likely attributable to utilization of yeast extract components in the medium (Chapter 2). Alternatively, the availability of residual soluble sugars, generated through xylan hydrolysis in the inoculum cultures, is unknown. If present, these sugars may have been carried over during inoculation of the experimental and control bottles permitting trace amounts of growth in the no-substrate control.

While biomass production stopped after 32 hours the cumulative end-products measured continued to increase until 48 hours post-inoculation (Fig. A.5.2). Similar end-product profiles were observed to those reported for growth of *T. thermohydrosulfuricus* WC1 on 5 g/L xylose (Fig. 5.2; Chapter 5.4.2) with lactate being the most abundant end-formed. Mass balances of all investigated products suggest that all major products of metabolism were accounted for (Table A.5.1).

The data presented confirms that *T. thermohydrosulfuricus* WC1 is capable of growth and end-product formation on Beechwood xylan. This is consistent with the correlation proposed between *Thermoanaerobacter* spp. which possess an extracellular

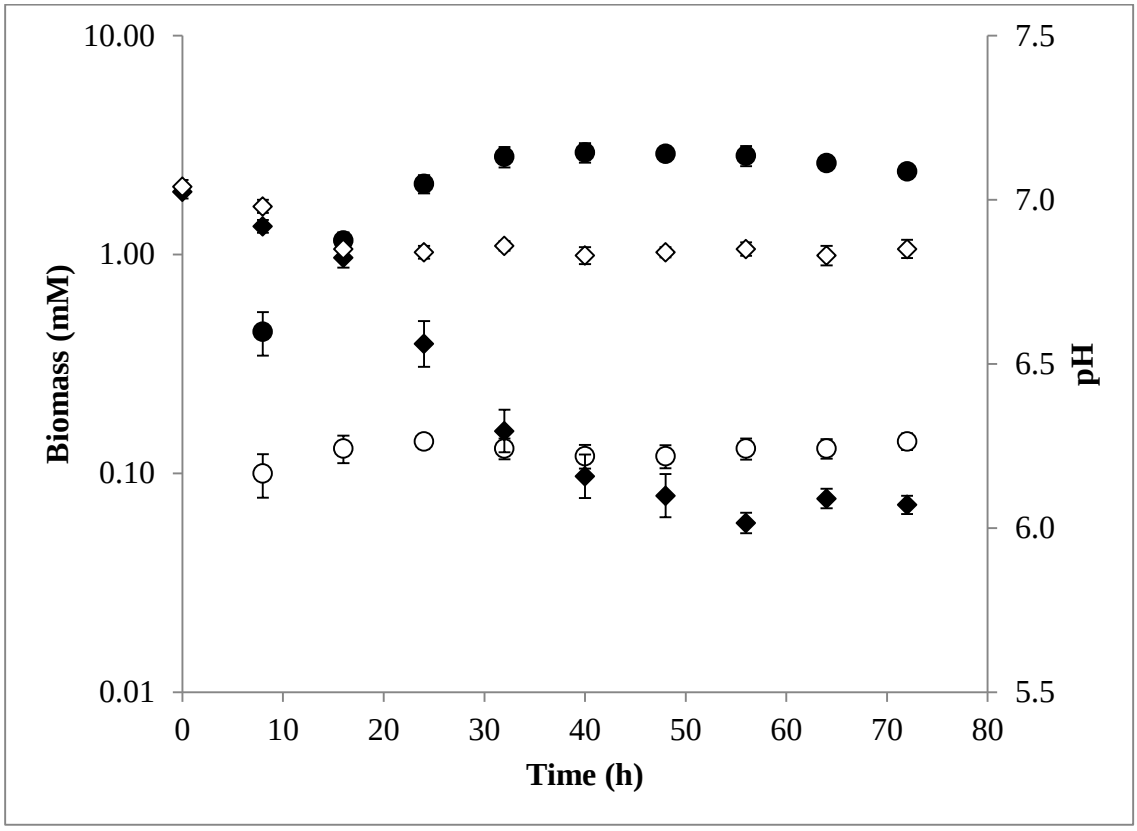


Fig. A.5.1. Biomass production and change in medium pH throughout growth of *T. thermohydrosulfuricus* WC1 on 5 g/L Beechwood xylan. Biomass production is represented by circles, while pH is represented by diamonds. Closed symbols represent cultures containing 5 g/L xylan, while open symbols represent the no-substrate control.

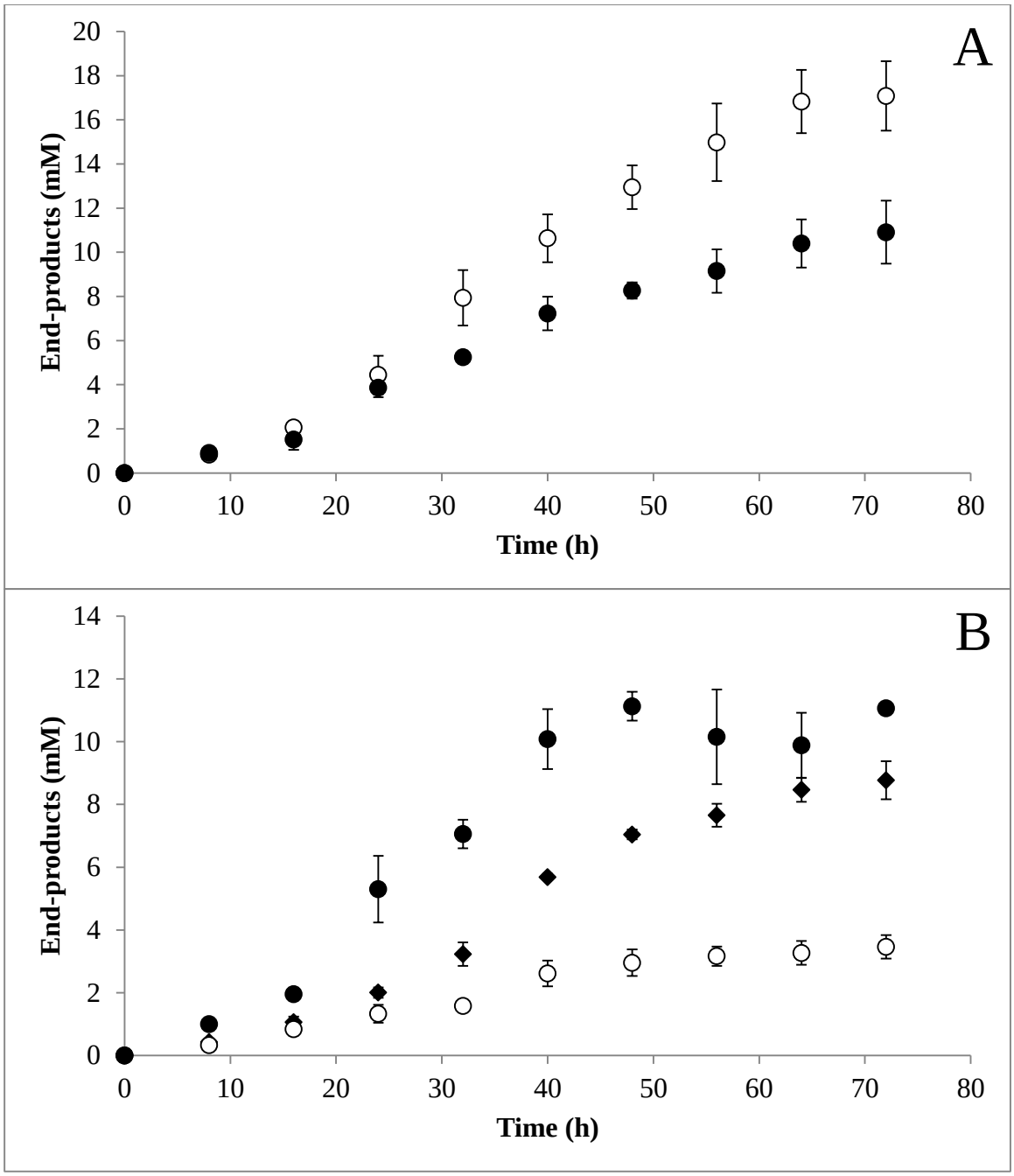


Fig. A.5.2. Average values of fermentative end-products by *T. thermohydrosulfuricus* WC1 on 5 g/L Beechwood xylan. (A) Formation of gaseous end-products. Closed circles represent H_2 production, while open circles represent CO_2 formation. (B) Production of non-gaseous end-products. Symbols are as follows: closed circles – lactate; open circles – acetate; closed diamonds – ethanol.

Table A.5.1. Calculated mass balances of end-product formation for *T. thermohydrosulfuricus* WC1 on 5 g/L Beechwood xylan.

Time (h)	O/R[a]	C1/C2[b]
0	NC[c]	NC
8	0.69	0.69
16	0.77	0.67
24	0.81	0.82
32	1.00	1.04
40	0.93	0.95
48	0.97	1.01
56	1.04	1.10
64	1.08	1.17
72	1.07	1.17

[a]Oxidized products formed/reduced products formed.

[b]One carbon end-products/two carbon end-products.

[c]Values used for calculations represent net production and cannot be calculated for T=0 measurements.

GH10 endoxylanase and growth on polymeric xylan (Chapter 4.4.3.1). However, it is yet unconfirmed if the putative GH10 enzyme identified is responsible for catalyzing xylan hydrolysis.

It is also important to note that it has not yet been determined if *T. thermohydrosulfuricus* WC1 is capable of growth on xylan-based hemicelluloses from different plant sources. Beechwood xylan is reported to be comprised principally of xylose (97%), and hexuronic acids (2.7%) with only trace amounts of arabinose and glucose as side-chains to the xylan-polymeric backbone (Hespell & Cotta, 1995). As xylan-based, as well as glucomannan-based, hemicellulose composition is known to vary dependent on plant species (Scheller & Ulvskov, 2010) different chemical and structural xylan-polymers may not be suitable for supporting growth in *T. thermohydrosulfuricus* WC1 and would be worthwhile to investigate in the future.

A.6 The effect of elevated levels of vitamin B_{12} in the fermentation medium on ethanol production by *Thermoanaerobacter thermohydrosulfuricus* WC1

A.6.1 Introduction

The study by Mori (1995) provides the first report of a *Thermoanaerobacter* strain capable of producing vitamin B_{12}-like compounds. In that investigation, the vitamin biosynthesis capabilities of *Thermoanaerobacter thermohydrosulfuricus* YM3 were found to complement nutritional auxotrophies in *Clostridium thermocellum* YM4. More than two decades later, He *et al*., (2011) reported B_{12} biosynthesis capabilities in *Thermoanaerobacter* sp. X514 and identified a link between B_{12} and ethanol production in some *Thermoanaerobacter* strains. In that study, it was found that the addition of B_{12} to the fermentation medium of *Thermoanaerobacter pseudethanolicus* 39E – *C. thermocellum* LQR1 co-cultures increased ethanol production to yields comparable to those observed in *Thermoanaerobacter* sp. X514 – *C. thermocellum* LQR1 co-culture without B_{12} addition. These findings were later validated by Hemme *et al*., (2011) when investigating the same *Thermoanaerobacter* strains in both mono- and co-culture. As such, the vitamin B_{12} biosynthesis capabilities of some *Thermoanaerobacter* strains may serve not only to complement nutritional auxotrophies in potential co-cultures, but may also be important to increasing ethanol production in *Thermoanaerobacter* spp.

Despite the apparent value to CBP systems for B_{12} synthesis in some strains, the physiological basis for B_{12}-linked ethanol production has not yet been resolved; though two mechanisms have been proposed based on *in silico* analyses. Hemme *et al*., (2011; see Fig. 2) propose a mechanism in some *Thermoanaerobacter* spp. whereby an

increased availability of B_{12} increases activity of a B_{12}-dependent methylmalonyl-CoA mutase. This reaction leads to increasing levels of methylmalonyl-CoA, which subsequently drives the activity of a membrane-bound, energy-conserving methylmalonyl-CoA decarboxylase. The energy conserved by the methylmalonyl-CoA decarboxylase (in the form of sodium-motive force) is proposed to have indirect effects on the overall physiology of *Thermoanaerobacter* sp. X514 or *T. pseudethanolicus* 39E, leading to increased ethanol yields.

A second mechanism has also been proposed for *Thermoanaerobacter* sp. X514 by Lin *et al*., (2012). Through transcriptomic analyses of *Thermoanaerobacter* sp. X514 grown on cellobiose, fructose, xylose, glucose, or xylose plus glucose, a putative link was established between B_{12}-biosynthesis enzymes and genes whose products may catalyze the consumption of ethanolamine or 1,2-propanediol (see Fig. S5D; Lin *et al*., 2012). Both ethanolamine and 1,2-propanediol were proposed to serve as alternative energy sources for ethanol production in that strain. As the proposed ethanolamine and 1,2-propanediol consuming enzymes are B_{12}-dependent, their subsequent activity could increase with elevated co-factor availability. The increased consumption of these potential energy sources would therefore lead to elevated ethanol production.

Based on the proposed role of B_{12} in other *Thermoanaerobacter* spp., the potential biosynthesis capabilities, as well as the putative B_{12}-binding enzymes, in *T. thermohydrosulfuricus* WC1 were evaluated through genomic analyses. Further, the effect of elevated levels of B_{12} in the fermentation medium was also investigated.

A.6.2 Materials and methods

A.6.2.1 Genomic analyses

Comparative genomic analyses of genes whose products are involved with vitamin B_{12} biosynthesis was conducted within the IMG-ER platform (Markowitz *et al*., 2009) as previously described (Chapter 4.3.3). To identify putative B_{12}-binding enzymes, genome specific searches were conducted within the IMG-ER database using specific protein family (pfam) (Finn *et al*., 2010) or Interpro (Hunter *et al*., 2009) qualifiers. Specific annotation qualifiers chosen include all those putatively involved with B_{12}-binding as reported within the IMG-ER database and are: pfam02310; pfam02607; pfam02965; pfam06522; pfam08471; IPR016176; and IPR013344.

A.6.2.2 Growth, biomass determination and ethanol production

Fifty millilitre cultures of *T. thermohydrosulfuricus* WC1 were grown in ATCC 1191 medium as previously described (Chapter 2.3.8). Filter sterilized anoxic solutions containing either cellobiose or cellobiose plus vitamin B_{12} were added such that the final concentrations in the medium contained either 5 g/L cellobiose + 5 μg/L vitamin B_{12} (found in the basal medium) (control condition) or 5 g/L cellobiose + 50 μg/L vitamin B_{12} (experimental condition). A final concentration of 50 μg/L vitamin B_{12} was considered sufficient as it was in excess of the exogenous vitamin B_{12} added in similar experiments (30 μg/L; He *et al*., 2010). Growth experiments were inoculated from cultures transferred twice (every 24 hours) on 5 g/L cellobiose.

At specific time points, 1 ml of culture was aseptically removed from actively growing bottles and centrifuged at 13,000 x *g* for 10 min. The supernatants were assayed

for ethanol produced as described (Chapter 2.3.8). Pellets resulting from centrifugation were washed in 0.9% NaCl and protein extracted via boiling in 2 M NaOH for 10 min. Protein determination and biomass estimates was conducted as previously described (Appendix A.5.2.2) with the exception that protein samples and standards were prepared/diluted using 2 M NaOH rather than a xylan extract solution. Reported values are averages of two independent experiments with each experiment consisting of five biological replicates. Further, assays of each biological replicate were done in triplicate.

A.6.3 Results and discussion

To assess the potential for vitamin B_{12} biosynthesis in *T. thermohydrosulfuricus* WC1, comparative genomic analysis of the strain was conducted against *Thermoanaerobacter* sp. X514 and *T. pseudethanolicus* 39E. As shown in Table A.6.1, *T. thermohydrosulfuricus* WC1 lacks many of the genes found in *Thermoanaerobacter* sp. X514; the only strain of the three with known B_{12} biosynthesis capabilities. Specifically, *T. thermohydrosulfuricus* WC1 lacks all genes for corrin ring biosynthesis and all but one gene, glutamate-tRNA ligase (TthWC1_2534), potentially involved in tetrapyrrole biosynthesis. The gene content profile identified is most comparable to *Caldanaerobacter subterraneus* subsp. *tengcongensis* MB4 (see Fig. 3 in Hemme *et al*., 2011).

While *T. thermohydrosulfuricus* WC1 lacks vitamin B_{12} biosynthetic capabilities, the potential for improved ethanologenesis may still exist with increased B_{12} availability in the fermentation medium as was observed for *T. pseudethanolicus* 39E (Hemme *et al*., 2011). As an initial assessment, all putative B_{12}-binding enzymes were identified in the

Table A.6.1. Genes involved[a] with vitamin B_{12} biosynthetic pathways in select *Thermoanaerobacter* spp.

		Strains		
Annotated function	Gene	*Thermoanaerobacter* sp. X514	*T. pseudethanolicus* 39E	*T. thermohydrosulfuricus* WC1
Glutamate-tRNA ligase	*gltX*	Teth514_2102	Teth39_1419	TthWC1_2534
Glutamyl-tRNA reductase	*hemA*	Teth514_0317		
Glutamate-1-semialdehyde-2,1-aminomutase	*hemL*	Teth514_0321	Teth39_1888	
Porphobilinogen synthase	*hemB*	Teth514_9320	Teth39_1889	
Porphobilinogen deaminase	*hemC*	Teth514_0318	Teth39_1891	
Uroporphyrin-III C-methyltransferase	*hemD*	Teth514_0319	Teth39_1890	
Cobalamin biosynthesis protein	*cbiX*	Teth514_0307	Teth39_1893	
Precorrin-2 C20-methyltransferase	*cbiL*	Teth514_0312	Teth39_1892	
Precorrin-3B C(17)-methyltransferase	*cbiH*	Teth514_0315		
Precorrin-4 C(11)-methyltransferase	*cbiF*	Teth514_0313		
Co-precorrin 5A	*cbiG*	Teth514_0314		
Co-precorrin 6A synthase	*cbiD*	Teth514_0309		
Precorrin-6A reductase	*cbiJ*	Teth514_0316		
Precorrin-8X methylmutase	*cbiC*	Teth514_0308		
Precorrin-6Y C15-methyltransferase subunit	*cbiT*	Teth514_0311		
Precorrin-6Y C15-methyltransferase subunit	*cbiE*	Teth514_0310		

Table A.6.1 cont.

Annotated function	Gene	*Thermoanaerobacter* sp. X514	*T. pseudethanolicus* 39E	*T. thermohydrosulfuricus* WC1
Cobyrinic acid a,c-diamide synthase	*cbiA*	Teth514_0299		
Cobyrinic acid a,c-diamide adenosyltransferase	*cobO*	Teth514_2187	Teth39_1505	TthWC1_0329
Cobyric acid synthase	*cbiP*	Teth514_0300	Teth39_1900	TthWC1_0972
Adenosylcobinamide-phosphate synthase	*cbiB*	Teth514_0301	Teth39_1899	TthWC1_0973
Adenosylcobinamide kinase	*cobU*	Teth514_0305	Teth39_1895	TthWC1_0977
Adenosylcobinamide-GDP ribazoletransferase	*cobS*	Teth514_0306	Teth39_1894	TthWC1_0978

[a]Gene designations and annotations derived from those reported by Hemme *et al*., (2011).

genomes of the three strains evaluated here (Table A.6.2). With the exception of hypothetical proteins, or protein ambiguously annotated as "cobalamin B_{12}-binding domain proteins," *T. thermohydrosulfuricus* WC1 differed from the other *Thermoanaerobacter* strains evaluated only by the absence of an annotated ribonucleotide reductase and a glycerol dehydratase. The glycerol dehydratase, which is found only in *Thermoanaerobacter* sp. X514, is one of the B_{12}-dependent enzymes proposed by Lin *et al*., (2011) to consume 1,2-propanediol.

Of additional note, an ortholog of the methylmalonyl-CoA mutase, a key player in the proposed mechanism of Hemme *et al*., (2011) was identified in *T. thermohydrosulfuricus* WC1. However, previous genomic analyses has identified that the putative membrane-bound methylmalonyl-CoA decarboxylase, catalyzing the second step in the proposed mechanism, is limited only to Clade 1 strains, and is not present in the *T. thermohydrosulfuricus* WC1 genome (Chapter 4.4.7.1). As such, based on the model presented by Hemme *et al*., (2011), an increased availability of B_{12} may drive synthesis of methylmalonyl-CoA, but would not drive sodium ion translocation (via a membrane-bound methylmalonyl-CoA decarboxylase) in *T. thermohydrosulfuricus* WC1. Based on these differences in genome content, it was considered unlikely that B_{12} have a stimulatory effect on ethanol production by *T. thermohydrosulfuricus* WC1 via the mechanisms proposed.

To test this, and identify if vitamin B_{12}-could have an unexpected effect on ethanol production in *T. thermohydrosulfuricus* WC1, cultures were grown on 5 g/L cellobiose with 50 μg/L exogenous B_{12} added. Increased levels of exogenous B_{12} had no

Table A.6.2. Identification of putative B_{12}-binding enzymes in the genomes of select *Thermoanaerobacter* spp.

		Strains		
Protein family (pfam and/or Interpro)	Genome annotation	*Thermoanaerobacter* sp. X514	*T. pseudethanolicus* 39E	*T. thermohydrosulfuricus* WC1
pfam02310	Radical SAM domain protein	Teth514_1404	Teth39_0963	TthWC1_2319
	Hypothetical protein	Teth514_2107		
	Cobalamin B_{12}-binding domain protein	Teth514_1854	Teth39_0903	TthWC1_0283
				TthWC1_0386
pfam02965	Methionine synthase	Teth514_1438	Teth39_0998	TthWC1_2541
pfam02310 or pfam02607	Homocysteine S-methyltransferase	Teth514_1211	Teth39_0697	TthWC1_1104
pfam02310 or IPR016176	Cobalamin B_{12}-binding domain protein	Teth514_1481	Teth39_1044	
			Teth39_1424	
IPR016176	Hypothetical protein	Teth514_1480	Teth39_1043	
	Methylmalonyl-CoA mutase, large subunit	Teth514_1855	Teth39_0902	TthWC1_1510
	Glycerol dehydratase	Teth514_1953		
IPR013344	Ribonucleotide reductase	Teth514_0821	Teth39_0329	

apparent stimulatory effect on ethanol production under the conditions tested (Fig. A.6.1). Further, no effect on biomass synthesis was observed.

The genomic analyses presented here identifies that, unlike some *Thermoanaerobacter* spp., *T. thermohydrosulfuricus* WC1 has no apparent vitamin B_{12} biosynthesis capabilities. Therefore, it is expected to have no obvious capacity to complement B_{12}-auxotrophic strains in potential co-cultures. Further, as increased ethanol production was also not observed, alternative models to those proposed (Hemme *et al*., 2011; Lin *et al*., 2011), which potentially explain the stimulatory effect of B_{12} in some strains, are not possible. It is important to realize though, that these findings neither confirm, nor deny, the proposed mechanism for B_{12}-linked ethanol production in *Thermoanaerobacter* sp. X514 or *T. pseudethanolicus* 39E, but rather only confirm that B_{12} has no overt effect in *T. thermohydrosulfuricus* WC1.

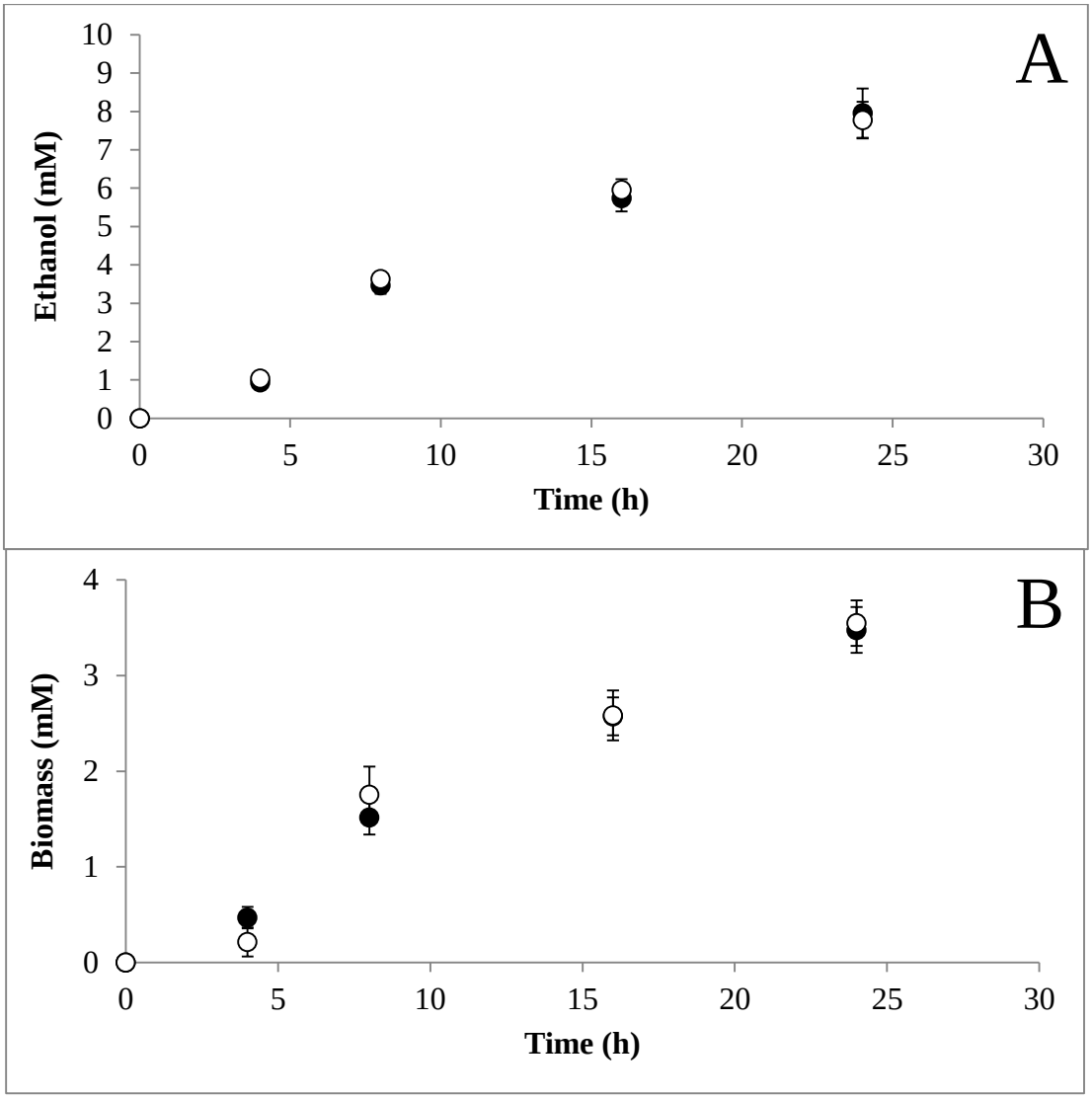


Fig. A.6.1. The influence of elevated vitamin B_{12} levels in the fermentation medium on ethanol production or biomass synthesis by *T. thermohydrosulfuricus* WC1. (A) Ethanol production. (B) Biomass synthesis.

A.7 A fluorescent *in situ* hybridization (FISH) protocol for distinguishing between *Clostridium thermocellum* and *Thermoanaerobacter* spp.

A.7.1 Background

Cellulose utilization by *Clostridium thermocellum* requires that the organism adhere to the substrate, resulting in naturally forming biofilms (Dumitrache *et al*., 2013). However, in co-cultures involving *C. thermocellum* and *Thermoanaerobacter* spp., the spatial organization of these strains is currently unknown. Given that many *Thermoanaerobacter* spp. lack the hydrolytic enzymes needed for lignocellulose hydrolysis (Chapter 4.4.3), understanding the physical relationships between *C. thermocellum* and *Thermoanaerobacter* spp. in co-cultures may provide insights into the mechanisms of biomass hydrolysis and utilization in co-culture settings.

Using fluorescent *Caldicellulosiruptor* spp. specific antibodies, a recent co-culture study by Kridelbaugh *et al*., (2013) involving *Caldicellulosiruptor bescii* and *C. thermocellum* has shown that both strains adhere to/associate with cellulosic biomass during its hydrolysis. Additionally, previous studies investigating natural lignocellulosic communities through the use of fluorescent *in situ* hybridization methodologies have also shown that non-cellulosic microorganism associate with lignocellulosic biomass during hydrolysis (Burrell *et al*., 2004; O'Sullivan *et al*., 2005). However, attempts to understand the spatial organization of *Thermoanaerobacter* spp. in relation to substrate-bound *C. thermocellum* have not yet been reported.

A.7.2 Materials and methods

A.7.2.1 Probe design

For development of a *Thermoanaerobacter* spp. molecular probe, nucleotide sequences of the 16S rRNA-encoding genes from *T. thermohydrosulfuricus* WC1, *T. brockii* subsp. *brockii* HTD4, *T. thermohydrosulfuricus* DSM 567 and *T. pseudethanolicus* 39E (all four 16S rRNA sequences; Chapter 2.4.2) were aligned using BioEdit v.7.0.9.0 (Hall, 1999). Regions conserved amongst all sequences were identified through manual observation and the reverse complement of those regions, corresponding to sequences complementary to the 16S rRNA molecule, was used to design probes using Primer3 software (Untergrasser *et al*., 2012). Candidate probes were restricted to those having a predicted melting temperature ≥57^{o}C as suggested by Hugenholtz *et al*. (2002). To further limit the number of potential probes tested, the complement sequence of candidate probes was used to query the genome of *C. thermocellum* through BLAST analysis in the Integrated Microbial Genomes database (Markowitz *et al*., 2012). The probe with the highest predicted melting temperature, and whose complement had the fewest aligned nucleotides against the *C. thermocellum* genome, was selected. An identical procedure was used for development of a *C. thermocellum* specific probe with the following exceptions: i) only the *C. thermocellum* ATCC 27405 16S rRNA encoding sequence was used to identify potential probes; and ii) candidate probes were searched against the genomes of *T. pseudethanolicus* 39E and *T brockii* subsp. *finnii* Ako-1 only as genome sequence data for *T. brockii* subsp. *brockii* HTD4, *T. thermohydrosulfuricus* WC1 and *T. thermohydrosulfuricus* DSM 567 was not available at the time of analysis.

A.7.2.2 Probe synthesis

To differentiate between *C. thermocellum* and *Thermoanaerobacter* spp. in co-cultures, probes were synthesized (Integrated DNA Technologies, Coralville, IA) with either a Cy3- or Cy5-fluorophore linked at its 5'-terminus. As such, the *Thermoanaerobacter* spp. probe (probe: Tbac) used in this study had the sequence: 5'-Cy3-CTCACCTTTCGGCTCGCTACT-3', while the *C. thermocellum* specific probe (probe: Cth188) had the sequence: 5'-Cy5- ATAACAGGACGATGCCGCC-3'.

A.7.2.3 Cell growth, cell fixation, probe hybridization and microscopy

Cultures of *T. thermohydrosulfuricus* WC1, *T. thermohydrosulfuricus* DSM 567, *T. brockii* subsp. *brockii* HTD4, *T. pseudethanolicus* 39E and *C. thermocellum* ATCC 27405 were grown in ATCC 1191 medium as previously described (Chapter 2.3.8). Cellobiose (2 g/L) was used as the carbon source for all *Thermoanaerobacter* spp. cultures, while 2 g/L Avicel was used for growth of *C. thermocellum*. Overnight cultures of all strains were used for analyses.

A modified version of the protocol by Hugenholtz *et al*., (2002) was used for cell fixation and probe hybridization. In brief, cells grown overnight were fixed by adding 300 µl of a 4% (wt/vol) paraformadledhye – phosphate buffered saline (PBS) (8.0 g/L NaCl; 0.2 g/L KCl; 1.44 g/L Na_2HPO_4; 0.2 NaH_2PO_4; pH = 7.0) solution to 100 µl of cell culture and incubating at either room temperature for 2 hours or at 4^oC for 24 hours. After incubation, cells were washed (5,000 x *g*) twice with PBS solution (without paraformaldehyde) and the pellet was resuspended in 300 µl of 100% ethanol.

Fixed samples (5 µl) were applied to 8-well Teflon coated microscope slides (Thermo Scientific, Ottawa, ON) and allowed to dry for 10 min at 46°C. Upon drying, samples were dehydrated by immersing the slides in 50%, 80% and 100% ethanol in consecutive fashion for 3 min at each concentration. Residual ethanol was removed through evaporation prior to probe hybridization. The basal hybridization buffer used was composed of: 360 µl of 5 M NaCl; 40 µl of 1 M Tris-HCl (pH = 7.2); 2 µl of 10% (wt/vol) SDS; *x* µl of formamide and water to 2 ml. Formamide concentrations ranging from 10-50% (vol/vol) (i.e. 200-1,000 µl of the hybridization buffer) were tested. A solution of 9 µl hybridization buffer containing 50 ng of either the Tbac probe or the Cth188 probe was applied to the corresponding samples and incubated at 46°C for 18 h. After hybridization, excess probe was removed by immersing the microscope slides in wash buffer containing: 1 ml 1 M Tris-HCl (pH=7.2); 0.5 ml of 0.5 M EDTA (pH=8.0); 0.05 ml of 10% SDS; *z* µl of 5 M NaCl; and water to 50 ml. The concentration of 5 M NaCl used in the wash buffer was adjusted according to the concentration of formamide used in the hybridization buffer as described (Hugenholtz *et al*., 2002). Slides were immersed in wash buffer for 20 min at 48°C and then air dried prior to visualization.

Microscopy was conducted using a Zeiss confocal laser scanning microscope using a 100× Plan Apochromat objective. Bound fluorophores were excited using either a 488 nm argon laser (Cy3) or a 633 nm He-Ne laser (Cy5).

A.7.3 Results and discussion

Understanding the spatial organization of strains in a designer co-culture may help shed insight into interspecies interactions and potential synergies, particularly

relating to biomass hydrolysis. To address this, probes targeting either the 16S rRNA molecule of *C. thermocellum* or select *Thermoanaerobacter* spp. were designed and tested. Staining of both planktonic as well as substrate bound *C. thermocellum* cells was observed using the Cth188 probe (Fig. A.7.1). Further, no overt differences were observed when different concentrations of formamide (10-50%) were used in the hybridization buffer. Fluorescence of the cellulose particles in regions lacking bound cells was also observed (Fig. A.7.1C-A.7.1D), which may represent the incomplete removal of the nucleotide probe during the wash steps. Alternatively, it may be attributable to the autofluorescence of cellulose as has been reported elsewhere (O'Sullivan *et al*., 2005).

Hybridization of the Tbac probe to all *Thermoanaerobacter* spp. tested was also observed (Fig. A.5.2) using 10-50% formamide hybridization buffer solutions. Additionally, repeated attempts to hybridize the Cth188 probe to any of the *Thermoanaerobacter* spp., or the Tbac probe to *C. thermocellum* were not successful at any concentration of formamide in the hybridization buffer tested.

The methodologies and results presented here represent initial attempts at designing molecular probes for potential application of distinguishing between *C. thermocellum* and select *Thermoanaerobacter* spp. in co-cultures through fluorescent *in situ* hybridization. While initially successful, additional efforts are needed to further validate the efficacy of these probes.

First, quantification of hybridization efficiencies would help to validate, or identify limitations, of the methodologies/probes used. The effectiveness of probe hybridization has previously been validated by using general nucleic acid stains

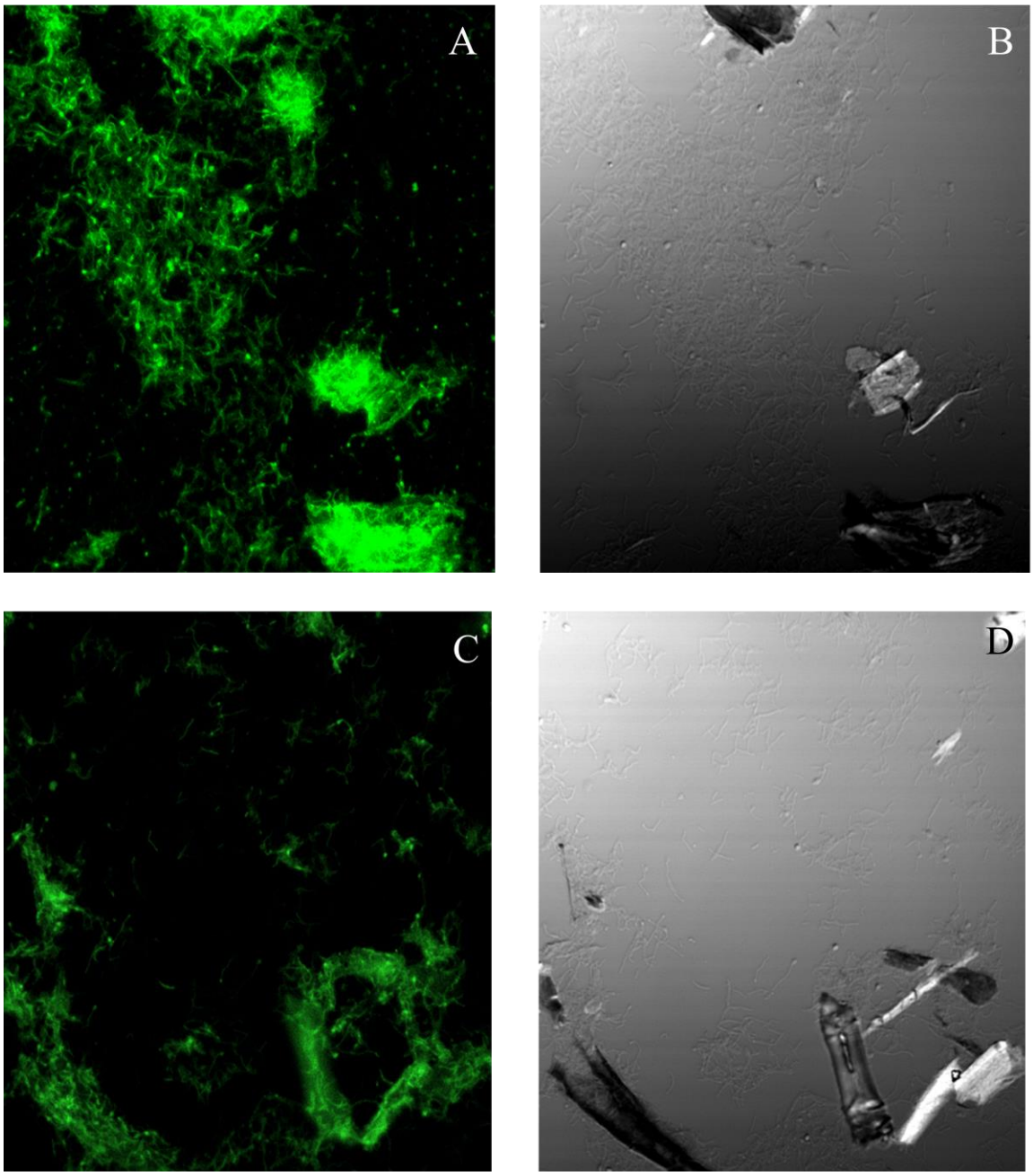


Fig. A.7.1. Binding of the Cth188 probe to *C. thermocellum* cultures. Comparison of the fluorescent (A) and light microscopy (B) images of cells growing on Avicel. Fluorescent (C) and light microscopy (D) images of *C. thermocellum* cells showing autofluorescence of the cellulose. Probe hybridization was done using 30% (A) or 20% (C) formamide hybridization buffers.

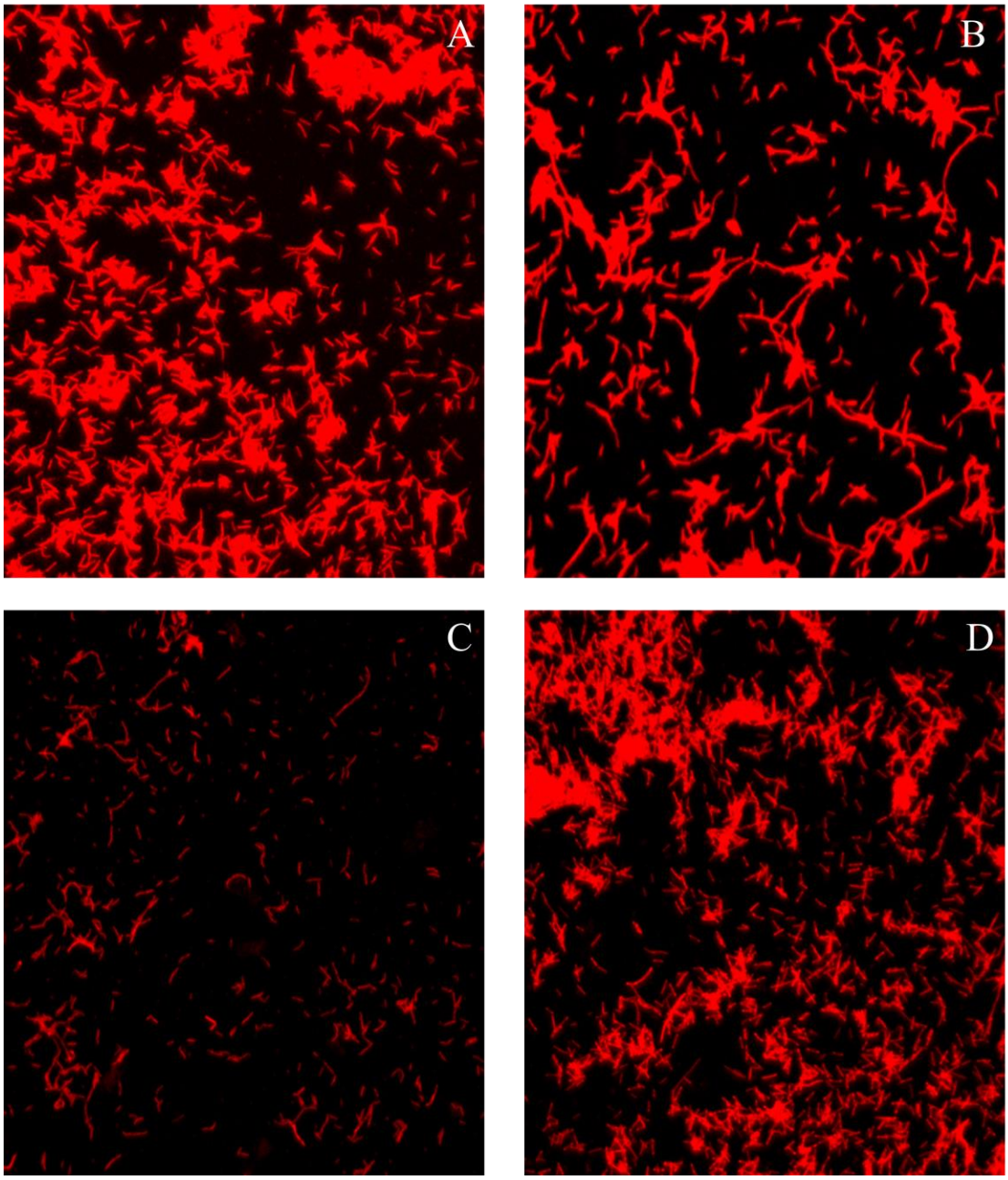


Fig. A.7.2. Hybridization of the Tbac probe to *Thermoanaerobacter* spp. tested. Hybridization to *T. thermohydrosulfuricus* WC1 (A), *T. thermohydrosulfuricus* DSM 567 (B), *T. brockii* subsp. *brockii* HTD4 (C) and *T. pseudethanolicus* 39E (D). For all images, probe hybridization was performed using a 50% formamide hybridization buffer.

(e.g. 4',6-diamidino-2-phenylindole (DAPI); propidium iodide) or universal FISH probes in conjunction with flow cytometry or quantitative microscopy techniques (Zoetendal *et al*., 2002; Rigottier-Gois *et al*., 2003). Similar approaches may help develop this work further.

Secondly, while cross-reactivity of the probes against non-target strains was not observed, the effect that simultaneous and/or sequential hybridization of both probes to a single sample has on hybridization efficiency has not yet been determined. Finally, the protocols used incorporate centrifugation of the cells/cellulosic biomass, which may have unintended shearing effects on the adherence of the microorganisms to the lignocellulose. In moving forward, efforts to identify suitable immobilization procedures, such as cryogenic fixing of adhered cells (Okabe *et al*., 1999) should also be explored.

Despite these challenges, these initial efforts show promise for incorporating FISH methodologies into designer co-culture studies using *C. thermocellum* and/or *Thermoanaerobacter* spp. Further, through database (probeBase; Loy *et al*., 2007) and literature searches, to the best of our knowledge, FISH probes specifically targeting these strains have not yet been reported. Therefore, the designed probes may also have potential applications in ecological studies of lignocellulosic communities.

A.8 Taxonomic designations of *Thermoanaerobacter* spp.

A.8.1 Background

The genus *Thermoanaerobacter* has routinely been transformed since its first description by Wiegel & Ljungdahl (1981). Further, numerous strain reassignments and altered taxonomic designations have made identifying the breadth of literature relating to many strains challenging. As such, Table A.8.1 is provided as a guide to assist readers establish connections between strains discussed in this thesis and the literature pertaining to those, as well as other, *Thermoanaerobacter* strains

Table A.8.1. Taxonomic designations of current and former *Thermoanaerobacter* spp.[a]

Current name	Reclassification reference	Former name(s)	Reference(s)
Thermoanaerobacter acetoethylicus HTB2/W	Rainey & Stackebrandt, 1993	*Thermobacteroides acetoethylicus* HTB2/W	Ben-Bassat & Zeikus, 1981
Thermoanaerobacter brockii subsp. *brockii* HTD4	Cayol *et al*., 1995	*Thermoanaerobacter brockii* HTD4	Lee et al., 1993
		Thermoanaerobium brockii HTD4	Zeikus *et al*., 1979
Thermoanaerobacter brockii susbp. *finnii* Ako-1	Cayol *et al*., 1995	*Thermoanaerobacter finnii* Ako-1	Schmid *et al*., 1986
Thermoanaerobacter brockii subsp. *lactiethylicus* SEBR 5268	NRC[b]	NRC	Cayol *et al*., 1995

Table A.8.1 cont.

Current name	Reclassification reference	Former name(s)	Reference(s)
Thermoanaerobacter ethanolicus JW200	NRC	NRC	Wiegel & Ljungdahl, 1981
Thermoanaerobacter italicus Ab9	NRC	NRC	Kozianowski *et al*., 1997
Thermoanaerobacter keratinophilus 2KXI	NRC	NRC	Riessen & Antranikian, 2001
Thermoanaerobacter kivui LKT-1	Collins *et al*., 1994	*Acetogenium kivui* LKT-1	Leigh *et al*., 1991
Thermoanaerobacter mathranii subsp. *alimentarius* AIP 505.99	NRC	NRC	Carlier *et al*., 2006
Thermoanaerobacter mathranii subsp. *mathranii* A3	NRC	NRC	Larsen *et al*., 1997
Thermoanaerobacter pseudethanolicus 39E	Onyenwoke *et al*., 2007	*Thermoanaerobacter ethanolicus* 39E	Lee *et al*., 1993
		Clostridium thermohydrosulfuricum 39E	Zeikus *et al*., 1980
Thermoanaerobacter siderophilus SR4	NRC	NRC	Slobodkin *et al*.,
Thermoanaerobacter sulfuigignens JW/SL-NZ826	NRC	NRC	Lee *et al*., 2007
Thermoanaerobacter sulfurophilus L-64	NRC	NRC	Bonch-Osmolovskaya *et al*., 1997
Thermoanaerobacter thermocopriae IAM13577	Collins *et al*., 1994	*Clostridium thermocopriae* IAM13577	Jin *et al*., 1988
Thermoanaerobacter thermohydrosulfuricus E100-69 *(*DSM 567)	Lee *et al*., 1993	*Clostridium thermohydrosulfuricum* E100-69	Hollaus & Klaushofer, 1973

Table A.8.1 cont.

Current name	Reclassification reference	Former name(s)	Reference(s)
Thermoanaerobacter wiegelii Rt8.B1	NRC	NRC	Cook *et al*., 1996
Caldanaerobacter subterraneus subsp. *subterraneus* SEBR 7858	Fardeau *et al*., 2004	*Thermoanaerobacter subterraneus* SEBR 7858	Fardeau *et al*., 2000
Caldanaerobacter subterraneus subsp. *tengcongensis* MB4	Fardeau *et al*., 2004	*Thermoanaerobacter tengcongensis* MB4	Xue *et al*., 2001
Caldanaerobacter subterraneus subsp. *yonsiensis* KB-1	Fardeau *et al*., 2004	*Thermoanaerobacter yonsiensis* KB-1	Kim *et al*., 2001

[a]Only strains with currently recognized and valid taxonomic designations (Wiegel, 2009) are provided in this table. All strains listed, with the exception of *Thermoanaerobacter keratinophilus*, are recognized type strains.

[b]NRC - strain has not been reclassified since its original description.

Cited literature

Abdou L, Boileau C, de Philip P, Pagès S, Fiérobe H-P, Tardif C. 2008. Transcriptional regulation of the *Clostridium cellulolyticum cip-cel* operon: a complex mechanism involving a catabolite-responsive element. J Bacteriol **190**:1499–1506.

Acinas SG, Haverkamp THA, Huisman J, Stal LJ. 2009. Phenotypic and genetic diversification of *Pseudanabaena* spp. (cyanobacteria). ISME J **3**:31–46.

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF. 2004a. Fine-scale phylogenetic architecture of a complex bacterial community. Nature **430**:551–554.

Acinas SG, Marcelino LA, Klepac-ceraj V, Polz MF. 2004b. Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. J Bacteriol **186**:2629–2635.

Adékambi T, Drancourt M, Raoult D. 2009. The *rpoB* gene as a tool for clinical microbiologists. Trends Microbiol **17**:37–45.

Adelsberger H, Hertel C, Glawischnig E, Zverlov V V, Schwarz WH. 2004. Enzyme system of *Clostridium stercorarium* for hydrolysis of arabinoxylan: reconstitution of the *in vivo* system from recombinant enzymes. Microbiol **150**:2257–2266.

Agbor VB, Cicek N, Sparling R, Berlin A, Levin DB. 2011. Biomass pretreatment: fundamentals toward application. Biotechnol Adv **29**:675–685.

Aich S, Delbaere LTJ. 2007. Phylogenetic study of the evolution of PEP-carboxykinase. Evol Bioinform Online **3**:333-340.

Allgaier M, Reddy A, Park JI, Ivanova N, D'haeseleer P, Lowry S, Sapra R, Hazen TC, Simmons BA, VanderGheynst JS, Hugenholtz P. 2010. Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community. PloS One **5**:e8812.

Amore A, Pepe O, Ventorino V, Birolo L, Giangrande C, Faraco V. 2013. Industrial waste based compost as a source of novel cellulolytic strains and enzymes. FEMS Microbiol Lett **339**:93–101.

Antizar-Ladislao B, Turrion-Gomez JL. 2008. Second-generation biofuels and local bioenergy systems. Biofuels Bioproducts Biorefining **2:**455–469.

Antunes A, Camiade E, Monot M, Courtois E, Barbut F, Sernova NV, Rodionov D A, Martin-Verstraete I, Dupuy B. 2012. Global transcriptional control by glucose and carbon regulator CcpA in *Clostridium difficile*. Nucleic Acids Res **40**:10701–10718.

Arahal DR, Sanchez E, Macian MC, Garay E. 2008. Value of *recN* sequences for species identification and as a phylogenetic marker within the family "*Leuconostocaceae*." Int Microbiol **11**:33-39.

Argyros DA, Tripathi SA, Barrett TF, Rogers SR, Feinberg LF, Olson DG, Foden JM, Miller BB, Lynd LR, Hogsett DA, Caiazza NC. 2011. High ethanol titers from cellulose by using metabolically engineered thermophilic, anaerobic microbes. Appl Environ Microbiol **77**:8288–8294.

Bahl H, Gottwald M, Kuhn A, Rale V, Andersch W, Gottschalk G. 1986. Nutritional factors affecting the ratio of solvents produced by *Clostridium acetobutylicum*. Appl Environ Microbiol **52**:169–172.

Balat M, Balat H. 2009. Recent trends in global production and utilization of bio-ethanol fuel. Appl Energy **86**:2273–2282.

Banat IM, Marchant R, Rahman TJ. 2004. *Geobacillus debilis* sp. nov., a novel obligately thermophilic bacterium isolated from a cool soil environment, and reassignment of *Bacillus pallidus* to *Geobacillus pallidus* comb. nov. Int J Syst Evol Microbiol **54**:2197–2201.

Bao Q, Tian Y, Li W, Xu Z, Xuan Z, Hu S, Dong W, Yang J, Chen Y, Xue Y, Xu Y, Lai X, Huang L, Dong X, Ma Y, Ling L, Tan H, Chen R, Wang J, Yu J, Yang H. 2002. A complete sequence of the *T. tengcongensis* genome. Genome Res **12**:689–700.

Bapteste E, Moreira D, Philippe H. 2003. Rampant horizontal gene transfer and phospho-donor change in the evolution of the phosphofructokinase. Gene **318**:185-191.

Beckner M, Ivey ML, Phister TG. 2011. Microbial contamination of fuel ethanol fermentations. Lett Appl Microbiol **53**:387–394.

Ben-Bassat A, Lamed R, Zeikus JG. 1981. Ethanol production by thermophilic bacteria: metabolic control of end product formation in *Thermoanaerobium brockii*. J Bacteriol **146**:192–199.

Ben-Bassat A, Zeikus J. 1981. *Thermobacteroides acetoethylicus* gen. nov. and spec. nov., a new chemoorganotrophic, anaerobic, thermophilic bacterium. Arch Microbiol **128**:365–370.

Bessen DE, Manoharan A, Luo F, Wertz JE, Robinson DA. 2005. Evolution of

transcription regulatory genes is linked to niche specialization in the bacterial pathogen *Streptococcus pyogenes*. J Bacteriol **187**:4163–4172.

Bhandiwad A, Guseva A, Lynd LR. 2013. Metabolic engineering of *Thermoanaerobacterium thermosaccharolyticum* for increased n-butanol production. Adv Microbiol **3**:46–51.

Biegel E, Schmidt S, González JM, Müller V. 2011. Biochemistry, evolution and physiological function of the Rnf complex, a novel ion-motive electron transport complex in prokaryotes. Cell Mol Life Sci **68**:613-634.

Biegel E, Schmidt S, Müller V. 2009. Genetic, immunological and biochemical evidence for a Rnf complex in the acetogen *Acetobacterium woodii*. Environ Microbiol **11**:1438-1443.

Bielen AAM, Willquist K, Engman J, van der Oost J, van Niel EWJ, Kengen SWM. 2010. Pyrophosphate as a central energy carrier in the hydrogen-producing extremely thermophilic *Caldicellulosiruptor saccharolyticus*. FEMS Microbiol Lett **307**:48–54.

Bischoff KM, Skinner-Nemec KA, Leathers TD. 2007. Antimicrobial susceptibility of *Lactobacillus* species isolated from commercial ethanol plants. J Ind Microbiol Biotechnol **34**:739–744.

Blokesch M, Paschos A, Theodoratou E, Bauer A, Hube M, Huth S, Bock A. 2002. Metal insertion into NiFe-hydrogenases. Biochem Soc Trans **30**:674–680.

Blouzard J-C, Coutinho PM, Fierobe H-P, Henrissat B, Lignon S, Tardif C, Pagès S,

de Philip P. 2010. Modulation of cellulosome composition in *Clostridium cellulolyticum*: Adaptation to the polysaccharide environment revealed by proteomic and carbohydrate- active enzyme analyses. Proteomics **10**:541–554.

Blumer-Schuette SE, Lewis DL, Kelly RM. 2010. Phylogenetic, microbiological, and glycoside hydrolase diversities within the extremely thermophilic, plant biomass-degrading genus *Caldicellulosiruptor*. Appl Environ Microbiol **76**:8084–8092.

Boisvert S, Laviolette F, Corbeil J. 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol **17**:1519–1533.

Bonch-Osmolovskaya E, Miroshnichenko M, Chernykh N, Kostrikina N, Pikuta E, Rainey F. 1997. Reduction of elemental sulfur by moderately thermophilic organotrophic bacteria and the description of *Thermoanaerobacter sulfurophilus* sp. nov . Microbiol **66**:581–587.

Booth IR. 1985. Regulation of cytoplasmic pH in bacteria. Microbiol Mol Biol Rev **49**:359-378.

Brechtel E, Bahl H. 1999. In *Thermoanaerobacterium thermosulfurigenes* EM1 S-layer homology domains do not attach to peptidoglycan. J Bacteriol **181**:5017–5023.

Bredholt S, Sonne-Hansen J, Nielsen P, Mathrani IM, Ahring BK. 1999. *Caldicellulosiruptor kristjanssonii* sp. nov., a cellulolytic, extremely thermophilic, anaerobic bacterium. Int J Syst Bacteriol **49**:991–996.

Brenner K, You L, Arnold FH. 2008. Engineering microbial consortia: a new frontier in synthetic biology. Trends Biotechnol **26**:483-489.

Breves R, Bronnenmeier K, Wild N, Lottspeich F, Staudenbauer WL, Hofemeister

J. 1997. Genes encoding two different β-glucosidases of *Thermoanaerobacter brockii* are clustered in a common operon. Appl Environ Microbiol **63**:3902–3910.

Brockhurst MA, Colegrave N, Hodgson DJ, Buckling A. 2007. Niche occupation limits adaptive radiation in experimental microcosms. PloS One **2**:e193.

Brown MV, Fuhrman JA. 2004. Marine bacterial microdiversity as revealed by internal transcribed spacer analysis. Aquat Microb Ecol **41**:15-23.

Brückner R, Titgemeyer F. 2002. Carbon catabolite repression in bacteria: choice of the carbon source and autoregulatory limitation of sugar utilization. FEMS Microbiol Lett **209**:141–148.

Bryant FO. 1991. Characterization of the fructose 1,6-bisphosphate-activated L(+)-lactate dehydrogenase from *Thermoanaerobacter ethanolicus*. J Enzyme Inhib **5**:235–248.

Bryant FO, Wiegel J, Ljungdahl LG. 1988. Purification and properties of primary and secondary alcohol dehydrogenases from *Thermoanaerobacter ethanolicus*. Appl Environ Microbiol **54**:460-465.

Buckling A, Wills MA, Colegrave N. 2003. Adaptation limits diversification of experimental bacterial populations. Science **302**:2107–2109.

Burdette DS, Jung S, Shen G-J, Hollingsworth RI, Zeikus JG. 2002. Physiological function of alcohol dehydrogenases and long-chain (C_{30}) fatty acids in alcohol tolerance of *Thermoanaerobacter ethanolicus*. Appl Environ Microbiol **68**:1914–1918.

Burdette DS, Vieille C, Zeikus JG. 1996. Cloning and expression of the gene encoding

the *Thermoanaerobacter ethanolicus* 39E secondary-alcohol dehydrogenase and biochemical characterization of the enzyme. Biochem J **316**:115-122.

Burdette DS, Zeikus JG. 1994. Purification of acetaldehyde dehydrogenase and alcohol dehydrogenases from *Thermoanaerobacter ethanolicus* 39E and characterization of the secondary-alcohol dehydrogenase (2^o Adh) as a bifunctional alcohol dehydrogenase--acetyl-CoA reductive thioesterase. Biochem J **302**:163-170.

Burrell PC, Sullivan CO, Song H, Clarke WP, Burrell PC, Sullivan CO, Song H, Clarke WP, Blackall LL. 2004. Identification, detection, and spatial resolution of *Clostridium* populations responsible for cellulose degradation in a methanogenic landfill leachate bioreactor. Appl Environ Microbiol **70**:2414–2419.

Caccamo D, Gugliandolo C, Stackebrandt E, Maugeri TL. 2000. *Bacillus vulcani* sp. nov., a novel thermophilic species isolated from a shallow marine hydrothermal vent. Int J Syst Evol Microbiol **50**:2009–2012.

Calusinska M, Happe T, Joris B, Wilmotte A. 2010. The surprising diversity of clostridial hydrogenases: a comparative genomic perspective. Microbiol **156**:1575-1588.

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucl Acids Res **37**: D233-D238.

Cardona CA, Sánchez OJ. 2007. Fuel ethanol production: process design trends and integration opportunities. Bioresour Technol **98**:2415–2457.

Carere CR, Kalia V, Sparling R, Cicek N, Levin DB. 2008. Pyruvate catabolism and

hydrogen synthesis pathway genes of *Clostridium thermocellum* ATCC 27405. Indian J Microbiol **48**:252–266.

Carere CR, Rydzak T, Verbeke TJ, Cicek N, Levin DB, Sparling R. 2012. Linking genome content to biofuel production yields: a meta-analysis of major catabolic pathways among select H_2 and ethanol-producing bacteria. BMC Microbiol **12**:295.

Carere, RC. 2013. Ph.D. thesis. University of Manitoba, Winnipeg, MB, Canada. Genomics of cellulolytic Clostridia and development of rational metabolic engineering strategies.

Carlier J-P, Bonne I, Bedora-Faure M. 2006. Isolation from canned foods of a novel *Thermoanaerobacter* species phylogenetically related to *Thermoanaerobacter mathranii* (Larsen 1997): emendation of the species description and proposal of *Thermoanaerobacter mathranii* subsp. *Alimentarius* subsp. Nov. Anaerobe **12**:153–159.

Carroll A, Somerville C. 2009. Cellulosic biofuels. Annu Rev Plant Biol **60**:165-182.

Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD. 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res **36**:D623–D631.

Cayol JL, Ollivier B, Patel BKC, Ravot G, Magot M, Agernon E, Grimont PAD, Garcia JL. 1995. Description of *Thermoanaerobacter brockii* subsp. *lactiethylicus* susbp. nov., isolated from a deep subsurface French oil well, a

proposal to reclassify *Thermoanaerobacter finnii* as *Thermoanaerobacter brockii* subsp. *finnii* comb. nov., and an emended description of *Thermoanaerobacter brockii*. Int J Syst Bacteriol **45**:783-789.

Cha M, Chung D, Elkins JG, Guss AM, Westpheling J. 2013. Metabolic engineering of *Caldicellulosiruptor bescii* yields increased hydrogen production from lignocellulosic biomass. Biotechnol Biofuels **6**:85.

Chain PSG, Grafham D V, Fulton RS, FitzGerald MG, Hostetler D, Ali MJ, Birren B, Bruce DC, Bulhay C, Cole JR, Ding Y, Dugan S, Field D, Garrity GM, Gibbs R, Graves T, Han CS, Harrison SH, Highlander S, Hugenholtz P, Khouri HM, Kodira CD, Kolker E, Kyrpides NC, Lang D, Lapidus A, Malfatti SA, Markowitz V, Metha T, Nelson KE, Parkhill J, Pitluck S, Qin X, Read TD, Schmutz J, Sozhamannan S, Sterk P, Strausberg RL, Sutton G, Thomson NR, Tiedje JM, Weinstock G, Wollam A, Consortium GSCHMPJ, Detter JC. 2009. Genome project standards in a new era of sequencing. Science **326**:4–5.

Chen Z, Wen B, Qang Q, Tong W, Guo J, Bai X, Zhao J, Sun Y, Tang Q, Lin Z, Lin L, Liu S. 2013. Quantitative proteomics reveals the temperature-dependent proteins encoded by a series of cluster genes in *Thermoanaerobacter tengcongensis*. Mol Cell Proteomics **12**:2266–2277.

Cheng J, Guffanti AA, Krulwich TA. 1997. A two-gene ABC-type transport system that extrudes Na^+ in *Bacillus subtilis* is induced by ethanol or protonophore. Mol Microbiol **23**:1107-1120.

Chundawat SPS, Beckham GT, Himmel ME, Dale BE. 2011. Deconstruction of

lignocellulosic biomass to fuels and chemicals. Annu Rev Chem Biomol Eng **2**:121–45.

Chung D, Cha M, Farkas J, Westpheling J. 2013a. Construction of a stable replicating shuttle vector for *Caldicellulosiruptor* species: use for extending genetic methodologies to other members of this genus. PloS One **8**:e62881.

Chung D, Farkas J, Huddleston JR, Olivar E, Westpheling J. 2012. Methylation by a unique α-class N4-cytosine methyltransferase is required for DNA transformation of *Caldicellulosiruptor bescii* DSM6725. PloS One **7**:e43844.

Chung D, Farkas J, Westpheling J. 2013b. Overcoming restriction as a barrier to DNA transformation in *Caldicellulosiruptor* species results in efficient marker replacement. Biotechnol Biofuels **6**:82.

Clark DP. 1989. The fermentation pathways of *Escherichia coli*. FEMS Microbiol Rev **63**:223–234.

Clavel T, Lippman R, Gavini F, Doré J, Blaut M. 2007. *Clostridium saccharogumia* sp. nov. and *Lactonifactor longoviformis* gen. nov., sp. nov., two novel human faecal bacteria involved in the conversion of the dietary phytoestrogen secoisolariciresinol diglucoside. Syst Appl Microbiol **30**:16–26.

Clomburg JM, Gonzalez R. 2010. Biofuel production in *Escherichia coli*: the role of metabolic engineering and synthetic biology. Appl Microbiol Biotechnol **86**:419–434.

Cohan FM. 2002. What are bacterial species? Ann Rev Microbiol **56**:457-487.

Collins MD, Lawson PA, Willems A, Cordoba JJ, Fernandez-Garayzabal J, Garcia

P, Cai J, Hippe H, Farrow JA. 1994. The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. Int J Syst Bacteriol **44**:812–826.

Comas I, González-Candelas F, Zúñiga M. 2008. Unraveling the evolutionary history of the phosphoryl-transfer chain of the phosphoenolpyruvate: phosphotransferase system through phylogenetic analyses and genome context. BMC Evol Biol **8**:147.

Cook GM. 2000. The intracellular pH of the thermophilic bacterium *Thermoanaerobacter wiegelii* during growth and production of fermentation acids. Extremophiles **4**:279–84.

Cook GM, Janssen PH, Morgan HW. 1993. Uncoupler-resistant glucose uptake by the thermophilic glycolytic anaerobe *Thermoanaerobacter thermosulfuricus* (*Clostridium thermohydrosulfuricum*). Appl Environ Microbiol **59**:2984-2990.

Cook GM, Janssen PH, Russell JB, Morgan HW. 1994b. Dual mechanisms of xylose uptake in the thermophilic bacterium *Thermoanaerobacter thermohydrosulfuricus*. FEMS Microbiol Lett **116**:257–262.

Cook GM, Morgan HW. 1994a. Hyperbolic growth of *Thermoanaerobacter thermohydrosulfuricus* (*Clostridium thermohydrosulfuricum*) increases ethanol production in pH-controlled batch culture. Appl Microbiol Biotechnol **41**:84-89.

Cook GM, Rainey FA, Patel BKC, Morgan HW. 1996. Characterization of a new obligately anaerobic thermophile, *Themoanaerobacter wiegelii* sp. nov. Int J Syst Bacteriol **46**:123–127.

Cooper JE, Feil EJ. 2004. Multilocus sequence typing - what is resolved? Trends

Microbiol **12**:373–377.

Coorevits A, Dinsdale AE, Halket G, Lebbe L, De Vos P, Van Landschoot A, Logan NA. 2012. Taxonomic revision of the genus *Geobacillus*: emendation of *Geobacillus, G. stearothermophilus, G. jurassicus, G. toebii, G. thermodenitrificans* and *G. thermoglucosidans* (nom. corrig., formerly "*thermoglucosidasius*"); transfer of *Bacillus thermantarcticus* to the genus as *G. thermantarcticus* comb. nov.; proposal of *Caldibacillus debilis* gen. nov., comb. nov.; transfer of *G. tepidamans* to *Anoxybacillus* as *A. tepidamans* comb. nov.; and proposal of *Anoxybacillus caldiproteolyticus* sp. nov. Int J Syst Evol Microbiol **62**:1470–85.

Countinho PM, Stam M, Blanc E, Henrissat B. 2003. Why are there so many carbohydrate-active enzyme-related genes in plants? Trends Plant Sci **8**:563–565.

Cripps RE, Eley K, Leak DJ, Rudd B, Taylor M, Todd M, Boakes S, Martin S, Atkinson T. 2009. Metabolic engineering of *Geobacillus thermoglucosidasius* for high yield ethanol production. Metab Eng **11**:398–408.

Crocetti GR, Hugenholtz P, Bond PL, Schuler A, Keller J, Jenkins D, Blackall LL. 2000. Identification of polyphosphate-accumulating organisms and design of 16S rRNA-directed probes for their detection and quantitation. Appl Environ Microbiol **66**:1175–1182.

Cuadros-Orellana S, Martin-Cuadrado A-B, Legault B, D'Auria G, Zhaxybayeva O, Papke RT, Rodriguez-Valera F. 2007. Genomic plasticity in prokaryotes: the case of the square haloarchaeon. ISME J **1**:235–245.

Čuboňováa L, Katano M, Kanai T, Atomi H, Reeve JN, Santangelo TJ. 2012. An

archaeal histone is required for transformation of *Thermococcus kodakarensis*. J Bacteriol **194**:6864–6874.

Currie DH, Herring CD, Guss AM, Olson DG, Hogsett DA, Lynd LR. 2013. Functional heterologous expression of an engineered full length CipA from *Clostridium thermocellum* in *Thermoanaerobacterium saccharolyticum*. Biotechnol Biofuels **6**:32.

Dam P, Kataeva I, Yang S-J, Zhou F, Yin Y, Chou W, Poole FL, Westpheling J, Hettich R, Giannone R, Lewis DL, Kelly R, Gilbert HJ, Henrissat B, Xu Y, Adams MWW. 2011. Insights into plant biomass conversion from the genome of the anaerobic thermophilic bacterium *Caldicellulosiruptor bescii* DSM 6725. Nucleic Acids Res **39**:3240–3254.

Daniels L, Belay N, Rajagopal BS. 1986. Assimilatory reduction of sulfate and sulfite by methanogenic bacteria. Appl Environ Microbiol **51**:703–9.

Darrett RH, Grisham CM. 1995. Biochemistry. Saunders College Publishing. New York, NY.

Das SP, Ravindran R, Ahmed S, Das D, Goyal D, Fontes CMGA, Goyal A. 2012. Bioethanol production involving recombinant *C. thermocellum* hydrolytic hemicellulase and fermentative microbes. Appl Biochem Biotechnol **167**:1475–88.

Datar R, Huang J, Maness P-C, Mohagheghi A, Czernik S, Chornet E. 2007. Hydrogen production from the fermentation of corn stover biomass pretreated with a steam-explosion process. Int J Hydrogen Energy **32**:932–939.

Daubin V, Moran NA, Ochman H. 2003. Phylogenetics and the cohesion of bacterial

genomes. Science **301**:829–832.

de Hoon MJL, Imoto S, Nolan J, Miyano S. 2004. Open source clustering software. Bioinformatics **20**:1453-1454.

de la Cruz F, Davies J. 2000. Horizontal gene transfer and the origin of species: lessons from bacteria. Trends Microbiol **8**:128–133.

Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. 1999. Alignment of whole genomes. Nucleic Acids Res **27**:2369–2376.

de Vrije T, Bakker RR, Budde MA, Lai MH, Mars AE, Claassen PA. 2009. Efficient hydrogen production from the lignocellulosic energy crop *Miscanthus* by the extreme thermophilic bacteria *Caldicellulosiruptor saccharolyticus* and *Thermotoga neapolitana*. Biotechnol Biofuels **2**:12.

de Vrije T, Budde MAW, Lips SJ, Bakker RR, Mars AE, Claassen PAM. 2010. Hydrogen production from carrot pulp by the extreme thermophiles *Caldicellulosiruptor saccharolyticus* and *Thermotoga neapolitana*. Int J Hydrogen Energy **35**:13206–13213.

de Vrije T, Mars AE, Budde MAW, Lai MH, Dijkema C, de Waard P, Claassen PAM. 2007. Glycolytic pathway and hydrogen yield studies of the extreme thermophile *Caldicellulosiruptor saccharolyticus*. Appl Microbiol Biotechnol **74**:1358–1367.

Dekkers LC, Phoelich CC, van der Fits L, Lugtenberg BJ. 1998. A site-specific recombinase is required for competitive root colonization by *Pseudomonas fluorescens* WCS365. Proc Natl Acad Sci U S A **95**:7051–7056.

Dellomonaco C, Fava F, Gonzalez R. 2010. The path to next generation biofuels:

successes and challenges in the era of synthetic biology. Microb Cell Fact **9**:3.

Delucchi MA. 2010. Impacts of biofuels on climate change, water use, and land use. Ann N Y Acad Sci **1195**:28–45.

Demain AL, Newcomb M, Wu JHD. 2005. Cellulase, Clostridia, and ethanol. Microbiol Mol Biol Rev **69**:124–154.

Demirbaş A. 2005. Bioethanol from cellulosic materials: a renewable motor fuel from biomass. Energy Sources **27**:327-337.

Desai SG, Guerinot ML, Lynd LR. 2004. Cloning of L-lactate dehydrogenase and elimination of lactic acid production via gene knockout in *Thermoanaerobacterium saccharolyticum* JW/SL-YS485. Appl Microbiol Biotechnol **65**:600–605.

Desvaux M. 2006. Unravelling carbon metabolism in anaerobic cellulolytic bacteria. Biotechnol Prog **22**:1229-1238.

Deutscher J. 2008. The mechanisms of carbon catabolite repression in bacteria. Curr Opin Microbiol **11**:87–93.

Doležal P, Vaňáčová Š, Tachezy J & Hrdý I. 2004. Malic enzymes of *Trichomonas vaginalis*: two enzyme families, two distinct origins. Gene **329**:81-92.

Dombek PE, Johnson LK, Zimmerley ST, Sadowsky MJ. 2000. Use of repetitive DNA sequences and the PCR To differentiate *Escherichia coli* isolates from human and animal sources. Appl Environ Microbiol **66**:2572–2577.

Dumitrache A, Wolfaardt G, Allen G, Liss SN, Lynd LR. 2013. Form and function of *Clostridium thermocellum* biofilms. Appl Environ Microbiol **79**:231–239.

Dürre P. 2007. Biobutanol: an attractive biofuel. Biotechnol J **2**:1525–34.

Erbeznik M, Dawson KA, Strobel HJ. 1998a. Cloning and characterization of transcription of the *xylAB* operon *in Thermoanaerobacter ethanolicus*. J Bacteriol **180**:1103–9.

Erbeznik M, Hudson SE, Herman AB, Strobel HJ. 2004. Molecular analysis of the *xylFGH* operon, coding for xylose ABC transport, in *Thermoanaerobacter ethanolicus*. Curr Microbiol **48**:295-299.

Erbeznik M, Ray M, Dawson KA, Strobel HJ. 1998b. Xylose transport by the anaerobic thermophile *Thermoanaerobacter ethanolicus* and the characterization of a D-xylose binding protein. Curr Microbiol **37**:295-300.

Fan Y-T, Zhang Y-H, Zhang S-F, Hou H-W, Ren B-Z. 2006. Efficient conversion of wheat straw wastes into biohydrogen gas by cow dung compost. Bioresour Technol **97**:500–505.

Fardeau ML, Magot M, Patel BK, Thomas P, Garcia JL, Ollivier B. 2000. *Thermoanaerobacter subterraneus* sp. nov., a novel thermophile isolated from oilfield water. Int J Syst Evol Microbiol **50**:2141–2149.

Fardeau ML, Patel BK, Magot M, Ollivier B. 1997. Utilization of serine, leucine, isoleucine, and valine by *Thermoanaerobacter brockii* in the presence of thiosulfate or *Methanobacterium* sp. as electron acceptors. Anaerobe **3**:405–410.

Fardeau M-L, Salinas MB, L'Haridon S, Jeanthon C, Verhé F, Cayol J-L, Patel BKC, Garcia JL, Ollivier B. 2004. Isolation from oil reservoirs of novel thermophilic anaerobes phylogenetically related to *Thermoanaerobacter subterraneus*: reassignment of *T. subterraneus*, *Thermoanaerobacter yonseiensis*, *Thermoanaerobacter tengcongensis* and *Carboxydibrachium pacificum* to

Caldanaerobacter subterraneus gen. nov., sp. nov., comb. nov. as four novel subspecies. Int J Syst Evol Microbiol **54**:467–474.

Faudon C, Fardeau ML, Heim J, Patel B, Magot M, Ollivier B. 1995. Peptide and amino acid oxidation in the presence of thiosulfate by members of the genus *Thermoanaerobacter*. Curr Microbiol **31**:152–157.

Feng X, Mouttaki H, Lin L, Huang R, Wu B, Hemme CL, He Z, Zhang B, Hicks LM, Xu J, Zhou J, Tang YJ. 2009. Characterization of the central metabolic pathways in *Thermoanaerobacter* sp. strain X514 via isotopomer-assisted metabolite analysis. Appl Environ Microbiol **75**:5001–5008.

Fiala G, Stetter KO. 1986. *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100^{o}C. Arch Microbiol **145**:56-61.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A. 2010. The Pfam protein families database. Nucleic Acids Res **38:**D211–222.

Fischer CR, Klein-Marcuschamer D, Stephanopoulos G. 2008. Selection and optimization of microbial hosts for biofuels production. Metab Eng **10**:295–304.

Freier D, Mothershed CP, Wiegel J. 1988. Characterization *of Clostridium thermocellum* JW20. Appl Environ Microbiol **54**:204–211.

Frejer-Schröder D, Wiegel J, Gottschalk G. 1989. Butanol formation by *Clostridium thermosaccharolyticum* at neutral pH. Biotechnol Lett **11**:831–836.

Frock AD, Gray SR, Kelly RM. 2012. Hyperthermophilic *Thermotoga* species differ

with respect to specific carbohydrate transporters and glycoside hydrolases. Appl Environ Microbiol **78**:1978–1986.

Furdui C, Ragsdale SW. 2002. The roles of coenzyme A in the pyruvate:ferredoxin oxidoreductase reaction mechanism: rate enhancement of electron transfer from a radical intermediate to an iron-sulfur cluster. Biochem **41:**9921–9937.

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

Galinier A, Deutscher J, Martin-Verstraete I. 1999. Phosphorylation of either Crh or HPr mediates binding of CcpA to the *Bacillus subtilis xyn cre* and catabolite repression of the *xyn* operon. J Mol Biol 286:307–314.

Galinier A, Haiech J, Kilhoffer MC, Jaquinod M, Stülke J, Deutscher J, Martin-Verstraete I. 1997. *The Bacillus subtilis crh* gene encodes a HPr-like protein involved in carbon catabolite repression. Proc Natl Acad Sci U S A **94**:8439–8444.

Garcia-Martinez D V, Shinmyo A, Madia A, Demain AL. 1980. Studies on cellulase production by *Clostridium thermocellum*. Eur J Appl Microbiol Biotechnol **9**:189–197.

Gavrilov SN, Bonch-Osmolovskaya EA, Slobodkin AI. 2003. Physiology of organotrophic and lithotrophic growth of the thermophilic iron-reducing bacteria *Thermoterrabacterium ferrireducens* and *Thermoanaerobacter siderophilus*. Microbiol **72**:160-167.

Georgieva TI, Mikkelsen MJ, Ahring BK. 2007b. High ethanol tolerance of the

thermophilic anaerobic ethanol producer *Thermoanaerobacter* BG1L1. Cent Eur J Biol **2**:364–377.

Georgieva TI, Skiadas IV, Ahring BK. 2007a. Effect of temperature on ethanol tolerance of a thermophilic anaerobic ethanol producer *Thermoanaerobacter* A10: modeling and simulation. Biotechnol Bioeng **98**:1161-1170.

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Peer Y Van De, Vandamme P, Thompson FL, Swings J. 2005. Re-evaluating prokaryotic species. Nat Rev Microbiol **3**:733–739.

Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, Bonner R, Aebersold R. 2012. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics **11**:O111.016717.

Gírio FM, Fonseca C, Carvalheiro F, Duarte LC, Marques S, Bogel-Łukasik R. 2010. Hemicelluloses for fuel ethanol: a review. Biores Technol **101**:4775–4800.

Glazunova OO, Raoult D, Roux V. 2009. Partial sequence comparison of the *rpoB*, *sodA*, *groEL* and *gyrB* genes within the genus *Streptococcus*. Int J Syst Evol Microbiol **59**:2317–2322.

Gold ND, Martin VJJ. 2007. Global view of the *Clostridium thermocellum* cellulosome revealed by quantitative proteomic analysis. J Bacteriol **189**:6787–6795.

Goldemberg J. 2007. Ethanol for a sustainable energy future. Science. **315**:808–810.

Görke B, Stülke J. 2008. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. Nat Rev Microbiol **6**:613–624.

Goupil N, Corthier G, Ehrlich SD, Renault P. 1996. Imbalance of leucine flux in

Lactococcus lactis and its use for the isolation of diacetyl-overproducing strains. Appl Environ Microbiol **62**:2636–2640.

Graham-Rowe D. 2011. Agriculture: beyond food versus fuel. Nature. **474**:S6-S8.

Gregg D, Saddler JN. 1996. A techno-economic assessment of the pretreatment and fractionation steps of a biomass-to-ethanol process. Appl Biochem Biotechnol **57**:711-727.

Grotkjaer T, Christakopoulos P, Nielsen J, Olsson L. 2005. Comparative metabolic network analysis of two xylose fermenting recombinant *Saccharomyces cerevisiae* strains. Metab Eng **7**:437–444.

Guedon E, Payot S, Desvaux M, Petitdemange H. 1999. Carbon and electron flow in *Clostridium cellulolyticum* grown in chemostat culture on synthetic medium. J Bacteriol **181**:3262–3269.

Guss AM, Olson DG, Caiazza NC, Lynd LR. 2012. Dcm methylation is detrimental to plasmid transformation in *Clostridium thermocellum*. Biotechnol Biofuels **5**:30.

Haft DH. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res **31**:371–373.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser **41**:95-98.

Hamelinck CN, Hooijdonk G Van, Faaij AP. 2005. Ethanol from lignocellulosic biomass: techno-economic performance in short-, middle- and long-term. Biomass Bioenergy **28**:384–410.

Hamilton-Brehm SD, Mosher JJ, Vishnivetskaya T, Podar M, Carroll S, Allman S,

Phelps TJ, Keller M, Elkins JG. 2010. *Caldicellulosiruptor obsidiansis* sp. nov., an anaerobic, extremely thermophilic, cellulolytic bacterium isolated from Obsidian Pool, Yellowstone National Park. Appl Environ Microbiol **76**:1014–1020.

Haruta S, Cui Z, Huang Z, Li M, Ishii M, Igarashi Y. 2002. Construction of a stable microbial community with high cellulose-degradation ability. Appl Microbiol Biotechnol **59**:529–534.

Harvey SH, Krien MJ, O'Connell MJ. 2002. Structural maintenance of chromosomes (SMC) proteins, a family of conserved ATPases. Genome Biol **3**:REVIEWS3003.

Haverkamp THA, Schouten D, Doeleman M, Wollenzien U, Huisman J, Stal LJ. 2009. Colorful microdiversity of *Synechococcus* strains (picocyanobacteria) isolated from the Baltic Sea. ISME J **3**:397–408.

Havlík P, Schneider UA, Schmid E, Böttcher H, Fritz S, Skalský R, Aoki K, De Cara S, Kindermann G, Kraxner F, Leduc S, McCallum I, Mosnier A, Sauer T, Obersteiner M. 2011. Global land-use implications of first and second generation biofuel targets. Energy Policy **39**:5690–5702.

Hayashi H, Takehara M, Hattori T, Kimura T, Karita S, Sakka K, Ohmiya K. 1999. Nucleotide sequences of two contiguous and highly homologous xylanase genes *xynA* and *xynB* and characterization of XynA from *Clostridium thermocellum*. Appl Microbiol Biotechnol **51**:348–357.

He M-X, Wu B, Shui Z-X, Hu Q-C, Wang W-G, Tan F-R, Tang X-Y, Zhu Q-L, Pan K, Li Q, Su X-H. 2012. Transcriptome profiling of *Zymomonas mobilis* under ethanol stress. Biotechnol Biofuels **5**:75.

He Q, Hemme CL, Jiang H, He Z, Zhou J. 2011. Mechanisms of enhanced cellulosic bioethanol fermentation by co-cultivation of *Clostridium* and *Thermoanaerobacter* spp. Bioresour Technol **102**:9586–9592.

He Q, Lokken PM, Chen S, Zhou J. 2009. Characterization of the impact of acetate and lactate on ethanolic fermentation by *Thermoanaerobacter ethanolicus*. Bioresour Technol **100**:5955–5965.

Heider J, Mai X, Adams MWW. 1996. Characterization of 2-ketoisovalerate ferredoxin oxidoreductase, a new and reversible coenzyme A-dependent enzyme involved in peptide fermentation by hyperthermophilic archaea. J Bacteriol **178:**780–787.

Heinonen JK, Drake HL. 1988. Comparative assessment of inorganic pyrophosphate and pyrophosphatase levels of *Escherichia coli*, *Clostridium pasteurianum*, and *Clostridium thermoaceticum*. FEMS Microbiol Lett **52**:205–208.

Helling RB, Vargas CN, Adams J. 1987. Evolution of *Escherichia coli* during growth in a constant environment. Genet **116**:349–358.

Hemme CL, Fields MW, He Q, Deng Y, Lin L, Tu Q, Mouttaki H, Zhou A, Feng X, Zuo Z, Ramsay BD, He Z, Wu L, Van Nostrand J, Xu J, Tang YJ, Wiegel J, Phelps TJ, Zhou J. 2011. Correlation of genomic and physiological traits of *Thermoanaerobacter* species with biofuel yields. Appl Environ Microbiol **77**:7998–8008.

Hemmingsen SM, Woolford C, van der Vies SM, Tilly K, Dennis DT, Georgopoulos CP, Hendrix RW, Ellis RJ. 1988. Homologous plant and bacterial proteins chaperone oligomeric protein assembly. Nature **333**:330–334.

Henkin TM. 1996. The role of CcpA transcriptional regulator in carbon metabolism in

Bacillus subtilis. FEMS Microbiol Lett **135**:9-15

Hertel TW, Tyner WE, Birur DK. 2010. The global impacts of biofuel mandates. Energy J **31**:75–100.

Hespell RB, Cotta MA. 1995. Degradation and utilization by *Butyrivibrio fibrisolvens* H17c of xylans with different chemical and physical properties. Appl Environ Microbiol **61**:3042–3050.

Hess M, Sczyrba A, Egan R, Kim T-W, 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. Sci **331**:463–467.

Hethener P, Brauman A, Garcia J-L. 1992. *Clostridium termitidis* sp. nov., a cellulolytic bacterium from the gut of the wood-feeding termite, *Nasutitermes lujae*. Syst Appl Microbiol **15**:52–58.

Hill JE, Penny SL, Crowell KG, Goh SH, Hemmingsen SM. 2004. cpnDB: a chaperonin sequence database. Genome Res **14**:1669–1675.

Hill JE, Seipp RP, Betts M, Hawkins L, Kessel AG Van, Crosby WL, Hemmingsen SM. 2002. Extensive profiling of a complex microbial community by high-throughput sequencing. Appl Environ Microbiol **68**:3055–3066.

Hill JE, Town JR, Hemmingsen SM. 2006. Improved template representation in *cpn60* polymerase chain reaction (PCR) product libraries generated from complex templates by application of a specific mixture of PCR primers. Environ Microbiol **8**:741–746.

Ho C-Y, Chang J-J, Lin J-J, Chin T-Y, Mathew GM, Huang C-C. 2011.

Establishment of functional rumen bacterial consortia (FRBC) for simultaneous biohydrogen and bioethanol production from lignocellulose. Int J Hydrogen Energy **36**:12168–12176.

Hollaus F, Klaushofer H. 1973. Identification of hyperthermophilic obligate anaerobic bacteria from extraction juices of beet sugar factories. Int Sugar J **75**:237-241.

Hollaus F, Sleytr U. 1972. On the taxonomy and fine structure of some hyperthermophilic saccharolytic Clostridia. Arch Microbiol **86**:129–146.

Holt PJ, Williams RE, Jordan KN, Lowe CR, Bruce NC. 2000. Cloning, sequencing and expression in *Escherichia coli* of the primary alcohol dehydrogenase gene from *Thermoanaerobacter ethanolicus* JW200. FEMS Microbiol Lett **190**:57-62.

Hong S, Bunge J, Leslin C, Jeon S, Epstein SS. 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. ISME J **3**:1365–1373.

Horner-Devine MC, Carney KM, Bohannan BJM. 2004. An ecological perspective on bacterial biodiversity. Proc R Soc Lon **271**:113–122.

Hsu DD, Inman D, Heath GA, Wolfrum EJ, Mann MK, Aden A. 2010. Life cycle environmental impacts of selected U.S. ethanol production and use pathways in 2022. Environ Sci Technol **44**:5289–5297.

Hu H, Wood TK. 2010. An evolved *Escherichia coli* strain for producing hydrogen and ethanol from glycerol. Biochem Biophys Res Comm **391**:1033–1038.

Hu S, Zheng H, Gu Y, Zhao J, Zhang W, Yang Y, Wang S, Zhao G, Yang S, Jiang W. 2011. Comparative genomic and transcriptomic analysis revealed genetic characteristics related to solvent formation and xylose utilization in *Clostridium acetobutylicum* EA 2018. BMC Genomics **12**:93.

Huang C, Patel BK, Mah RA, Baresi L. 1998. *Caldicellulosiruptor owensensis* sp. nov., an anaerobic, extremely thermophilic, xylanolytic bacterium. Int J Syst Bacteriol **48**:91–97.

Huang H, Wang S, Moll J, Thauer RK. 2012. Electron bifurcation involved in the energy metabolism of the acetogenic bacterium *Moorella thermoacetica* growing on glucose or H_2 plus CO_2. J Bacteriol **194:**3689–3699.

Huber R, Langworthy TA, Knig H, Thomm M, Woese CR, Sleytr UB, Stetter KO. 1986. *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90ºC. Arch Microbiol **144**:324–333.

Hugenholtz P, Tyson GW, Blackall LL. 2002. Design and evaluation of 16S rRNA-targeted oligonucleotide probes for fluorescence *in situ* hybridization, p. 29–42. *In* de Muro MA, Rapley R (ed), Gene Probes, Methods in Molecular Biology, Humana Press, Totowa, NJ.

Huhnke RL, Lewis RS, Tanner RS. 2010. Isolation and characterization of novel clostridial species. U.S. patent 7,704,723.

Hung K-S, Liu S-M, Fang T-Y, Tzou W-S, Lin F-P, Sun K-H, Tang S-J. 2011. Characterization of a salt-tolerant xylanase from *Thermoanaerobacterium saccharolyticum* NTOU1. Biotechnol Lett **33**:1441–1447**.**

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJA, Thimma M, Thomas PD, Valentin F,

Wilson D, Wu CH, Yeats C. 2009. InterPro: the integrative protein signature database. Nucleic Acids Res **37**:D211–215.

Hyun HH, Zeikus JG. 1985. Regulation and genetic enhancement of glucoamylase and pullulanase production in *Clostridium thermohydrosulfuricum*. J Bacteriol **164**:1146–1152.

IEA (International Energy Agency). 2012. Key World Energy Statistics. http://www.iea.org/publications/freepublications/publication/name,31287,en.html, accessed August 2, 2013.

Ingram LO, Conway T, Clark DP, Sewell GW, Preston JF. 1987. Genetic engineering of ethanol production in *Escherichia coli*. Appl Environ Microbiol **53**:2420–2425.

International Union of Biochemistry and Molecular Biology (1992). 1992. Enzyme Nomenclature. Academic Press, San Diego, CA.

Islam R, Cicek N, Sparling R, Levin D. 2006. Effect of substrate loading on hydrogen production during anaerobic fermentation by *Clostridium thermocellum* 27405. Appl Microbiol Biotechnol **72**:576–583.

Izquierdo JA, Goodwin L, Davenport KW, Teshima H, Bruce D, Detter C, Tapia R, Han S, Land M, Hauser L, Jeffries CD, Han J, Pitluck S, Nolan M, Chen A, Huntemann M, Mavromatis K, Mikhailova N, Liolios K, Woyke T, Lynd LR. 2012. Complete genome sequence of *Clostridium clariflavum* DSM 19732. Stand Genomic Sci **6**:104–115.

Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A **96**:3801–3806.

Jayaprakash TP, Schellenberg JJ, Hill JE. 2012. Resolution and characterization of

distinct *cpn60*-based subgroups of *Gardnerella vaginalis* in the vaginal microbiota. PloS One **7**:e43009.

Jannasch H, Huber R, Belkin S, Stetter K. 1988. *Thermotoga neapolitana* sp. nov. of the extremely thermophilic, eubacterial genus *Thermotoga*. Arch Microbiol **150**:103–104.

Jaspers E, Overmann J. 2004. Ecological significance of microdiversity : identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. Appl Environ Microbiol **70**:4831–4839.

Jeppsson M, Bengtsson O, Franke K, Lee H, Hahn-Hägerdal B, Gorwa-Grauslund MF. 2006. The expression of a *Pichia stipitis* xylose reductase mutant with higher K_M for NADPH increases ethanol production from xylose in recombinant *Saccharomyces cerevisiae*. Biotechnol Bioeng **93**:665–673.

Jiao Y, Navid A, Stewart BJ, McKinlay JB, Thelen MP, Pett-Ridge J. 2012. Syntrophic metabolism of a co-culture containing *Clostridium cellulolyticum* and *Rhodopseudomonas palustris* for hydrogen production. Int J Hydrogen Energy **37**:11719–11726.

Jin F, Yamasato K, Toda K. 1988. *Clostridium thermocopriae* sp. nov., a cellulolytic thermophile from animal feces, compost, soil, and a hot spring in Japan. Int J Syst Bacteriol **38**:279–281.

Johnson EA, Madia A, Demain AL. 1981. Chemically defined minimal medium for growth of the anaerobic cellulolytic thermophile *Clostridium thermocellum*. Appl Environ Microbiol **41**:1060–1062.

Jones CR, Ray M, Strobel HJ. 2002b. Cloning and transcriptional analysis of the

Thermoanaerobacter ethanolicus strain 39E maltose ABC transport system. Extremophiles **6**:291–299.

Jones CR, Ray M, Strobel HJ. 2002a. Transcriptional analysis of the xylose ABC transport operons in the thermophilic anaerobe *Thermoanaerobacter ethanolicus*. Curr Microbiol **45**:54–62.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. *In* Munro HN (ed), Mammalian Protein Metabolism, vol 3. Academy Press, New York, NY.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. 2008. KEGG for linking genomes to life and the environment. Nucleic Acids Res **36**:D480–484**.**

Kim BC, Grote R, Lee DW, Antranikian G, Pyun YR. 2001. *Thermoanaerobacter yonseiensis* sp. nov., a novel extremely thermophilic, xylose-utilizing bacterium that grows at up to 85ºC. Int J Syst Evol Microbiol **51**:1539–1548.

Kochetov GA, Sevostyanova I A. 2005. Binding of the coenzyme and formation of the transketolase active center. IUBMB Life **57**:491–497.

Kozianowski G, Canganella F, Rainey F a, Hippe H, Antranikian G. 1997. Purification and characterization of thermostable pectate-lyases from a newly isolated thermophilic bacterium, *Thermoanaerobacter italicus* sp. nov. Extremophiles **1**:171–82.

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.

Kresse AU, Dinesh SD, Larbig K, Römling U. 2003. Impact of large chromosomal

inversions on the adaptation and evolution of *Pseudomonas aeruginosa* chronically colonizing cystic fibrosis lungs. Mol Microbiol **47**:145-158.

Kridelbaugh DM, Nelson J, Engle NL, Tschaplinski TJ, Graham DE. 2013. Nitrogen and sulfur requirements for *Clostridium thermocellum* and *Caldicellulosiruptor bescii* on cellulosic substrates in minimal nutrient media. Bioresour Technol **130**:125–135.

Krokhin OV. 2006. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. Anal Chem **78**:7785–7795.

Krug K, Nahnsen S, Macek B. 2011. Mass spectrometry at the interface of proteomics and genomics. Mol Bio **7**:284–291.

Kruus K, Lua AC, Demain AL, Wu JH. 1995. The anchorage function of CipA (CelL), a scaffolding protein of the *Clostridium thermocellum* cellulosome. Proc Natl Acad Sci U S A **92**:9254–9258.

Kuhnert P, Korczak BM, Stephan R, Joosten H, Iversen C. 2009. Phylogeny and prediction of genetic similarity of *Cronobacter* and related taxa by multilocus sequence analysis (MLSA). Int J Food Microbiol **136**:152–158.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol **5**:R12.

Lacis LS, Lawford HG. 1991. *Thermoanaerobacter ethanolicus* growth and product yield from elevated levels of xylose or glucose in continuous cultures. Appl Environ Microbiol **57**:579–585.

Lamed R, Zeikus JG. 1980a. Ethanol production by thermophilic bacteria: relationship between fermentation product yields of and catabolic enzyme activities in *Clostridium thermocellum* and *Thermoanaerobium brockii*. J Bacteriol **144**:569-578.

Lamed R, Zeikus JG. 1980b. Glucose fermentation pathway of *Thermoanaerobium brockii*. J Bacteriol **141**:1251–1257.

Larsen L, Nielsen P, Ahring BK. 1997. *Thermoanaerobacter mathranii* sp. nov., an ethanol-producing , extremely thermophilic anaerobic bacterium from a hot spring in Iceland. Arch Microbiol **168**:114–119.

Lawson A, Soh A, Ralambotianal H, Ollivier- B, Prensier G, Tine E, Garcia J-L. 1991. *Clostridium thermopalmarium* sp. nov., a moderately thermophilic butyrate-producing bacterium isolated from Palm Wine in Senegal. Syst Appl Microbiol **14**:135–139.

Le Coq D, Lindner C, Krüger S, Steinmetz M, Stülke J. 1995. New β-glucoside (*bgl*) genes in *Bacillus subtilis*: the *bglP* gene product has both transport and regulatory functions similar to those of *BglF*, its *Escherichia coli* homolog. J Bacteriol **177**:1527–1535.

Le Ruyet P, Dubourguier HC, Albagnac G. 1984. Homoacetogenic fermentation of cellulose by a coculture of *Clostridium thermocellum* and *Acetogenium kivui*. Appl Environ Microbiol **48**:893–894.

Lee SK, Chou H, Ham TS, Lee TS, Keasling JD. 2008. Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. Curr Opin Biotechnol **19**:556–563.

Lee Y-J, Dashti M, Prange A, Rainey FA, Rohde M, Whitman WB, Wiegel J. 2007. *Thermoanaerobacter sulfurigignens* sp. nov., an anaerobic thermophilic bacterium that reduces 1 M thiosulfate to elemental sulfur and tolerates 90 mM sulfite. Int J Syst Evol Microbiol **57**:1429–1434.

Lee YE, Jain MK, Lee C, Lowe SE, Zeikus JG. 1993. Taxonomic distinction of saccharolytic thermophilic anaerobes: description of *Thermoanaerobacterium xylanolyticum* gen. nov., sp. nov., and *Thermoanaerobacterium saccharolyticum* gen. nov., sp. nov.; reclassification of *Thermoanaerobium brockii, Clostridium thermosulfurogenes,* and *Clostridium thermohydrosulfuricum* E100-69 as *Thermoanaerobacter brockii* comb. nov., *Thermoanaerobacterium thermosulfurigenes* comb. nov., and *Thermoanaerobacter thermohydrosulfuricus* comb. nov., respectively; and transfer of *Clostridium thermohydrosulfuricum* 39E *to Thermoanaerobacter ethanolicus*. Int J Syst Bacteriol **43**:41–51.

Leigh J, Mayer F, Wolfe R. 1981. *Acetogenium kivui*, a new thermophilic hydrogen-oxidizing acetogenic bacterium. Arch Microbiol **129:**275–280.

Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B. 2013. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol Biofuels **6**:41.

Levin DB, Islam R, Cicek N, Sparling R. 2006. Hydrogen production by *Clostridium thermocellum* 27405 from cellulosic biomass substrates. Int J Hydrogen Energy **31**:1496-1503.

Li H, Cann AF, Liao JC. 2010a. Biofuels: biomolecular engineering fundamentals and advances. Annu Rev Chem Biomol Eng **1**:19–36.

Li S, Lai C, Cai Y, Yang X, Yang S, Zhu M, Wang J, Wang X. 2010b. High efficiency hydrogen production from glucose/xylose by the *ldh*-deleted *Thermoanaerobacterium* strain. Bioresour Technol **101**:8718–8724.

Li X, Mupondwa E, Panigrahi S, Tabil L, Sokhansanj S, Stumborg M. 2012. A review of agricultural crop residue supply in Canada for cellulosic ethanol production. Renewable Sustainable Energy Rev **16**:2954–2965.

Li S, Wen J, Jia X. 2011. Engineering *Bacillus subtilis* for isobutanol production by heterologous Ehrlich pathway construction and the biosynthetic 2-ketoisovalerate precursor pathway overexpression. Appl Microbiol Biotechnol **91**:577–589.

Lin L, Ji Y, Tu Q, Huang R, Teng L, Zeng X, Song H, Wang K, Zhou Q, Li Y, Cui Q, He Z, Zhou J, Xu J. 2013. Microevolution from shock to adaptation revealed strategies improving ethanol tolerance and production in *Thermoanaerobacter*. Biotechnol Biofuels **6**:103.

Lin L, Song H, Ji Y, He Z, Pu Y, Zhou J, Xu J. 2010. Ultrasound-mediated DNA transformation in thermophilic gram-positive anaerobes. PloS One **5**:e12582

Lin L, Song H, Tu Q, Qin Y, Zhou A, Liu W, He Z, Zhou J, Xu J. 2011. The *Thermoanaerobacter* glycobiome reveals mechanisms of pentose and hexose co-utilization in bacteria. PLoS Genet **7**:e1002318.

Lin Lu, Xu Jian. 2013. Dissecting and engineering metabolic and regulatory networks of thermophiles for biofuel production. Biotechnol Adv., in press. http://dx.doi.org/10.1016/j.biotechadv.2013.03.003

Liou JS-C, Balkwill DL, Drake GR, Tanner RS. 2005. *Clostridium carboxidivorans*

sp. nov., a solvent-producing clostridium isolated from an agricultural settling lagoon, and reclassification of the acetogen *Clostridium scatologenes* strain SL1 as *Clostridium drakei* sp. nov. Int J Syst Evol Microbiol **55**:2085–2091.

Liu H, Zhang T, Fang HHP. 2003. Thermophilic H_2 production from a cellulose-containing wastewater. Biotechnol Lett **25**:365–369.

Liu L, Zhang L, Tang W, Gu Y, Hua Q, Yang S, Jiang W, Yang C. 2012. Phosphoketolase pathway for xylose catabolism in *Clostridium acetobutylicum* revealed by ^{13}C metabolic flux analysis. J Bacteriol **194**:5413–5422.

Liu Y, Yu P, Song X, Qu Y. 2008. Hydrogen production from cellulose by co-culture of *Clostridium thermocellum* JN4 and *Thermoanaerobacterium thermosaccharolyticum* GD17. Int J Hydrogen Energy **33**:2927–2933.

Lochner A, Giannone RJ, Keller M, Antranikian G, Graham DE, Hettich RL. 2011b. Label-free quantitative proteomics for the extremely thermophilic bacterium *Caldicellulosiruptor obsidiansis* reveal distinct abundance patterns upon growth on cellobiose, crystalline cellulose, and switchgrass. J Proteome Res **10**:5302–5314.

Lochner A, Giannone RJ, Rodriguez Jr M, Shah MB, Mielenz JR, Keller M, Antranikian G, Graham DE, Hettich RL. 2011a. Use of label-free quantitative proteomics to distinguish the secreted cellulolytic systems of *Caldicellulosiruptor bescii* and *Caldicellulosiruptor obsidiansis*. Appl Environ Microbiol **77**:4042–4054.

Löffler FE, Sun Q, Li J, Tiedje JM. 2000. 16S rRNA gene-based detection of

tetrachloroethene-dechlorinating *Desulfuromonas* and *Dehalococcoides* species. Appl Environ Microbiol **66**:1369–1374.

Loftie-Eaton W, Taylor M, Horne K, Tuffin MI, Burton SG, Cowan DA. 2013. Balancing redox cofactor generation and ATP synthesis: key microaerobic responses in thermophilic fermentations. Biotechnol Bioeng **110**:1057–1065.

Lorén JG, Farfán M, Miñana-Galbis D, Fusté MC. 2010. Prediction of whole-genome DNA G + C content within the genus *Aeromonas* based on housekeeping gene sequences. Syst Appl Microbiol **33**:237–242.

Lovitt RW, Longin R, Zeikus JG. 1984. Ethanol production by thermophilic bacteria: physiological comparison of solvent effects on parent and alcohol-tolerant strains of *Clostridium thermohydrosulfuricum*. Appl Environ Microbiol **48**:171–7.

Lovitt RW, Shen GJ, Zeikus JG. 1988. Ethanol production by thermophilic bacteria: biochemical basis for ethanol and hydrogen tolerance in *Clostridium thermohydrosulfuricum*. J Bacteriol **170**:2809–2815

Loy A, Maixner F, Wagner M, Horn M. 2007. probeBase-an online resource for rRNA-targeted oligonucleotide probes: new features 2007. Nucleic Acids Res **35**:D800–804.

Ludwig W. 2007. Nucleic acid techniques in bacterial systematics and identification. Int J Food Microbiol **120**:225–236.

Lund PA. 2009. Multiple chaperonins in bacteria - why so many? FEMS Microbiol Rev **33**:785–800.

Luoto HH, Belogurov GA, Baykov AA, Lahti R, Malinen AM. 2011. Na^+-

translocating membrane pyrophosphatases are widespread in the microbial world and evolutionarily precede H^+-translocating pyrophosphatases. J Biol Chem **286**:21633–21642.

Lüthi E, Love DR, McAnulty J, Wallace C, Caughey PA, Saul D, Bergquist PL. 1990. Cloning, sequence analysis, and expression of genes encoding xylan-degrading enzymes from the thermophile "*Caldocellum saccharolyticum*". Appl Environ Microbiol **56**:1017–1024.

Lynd LR, Currie D, Ciazza N, Herring C, Orem N. 2008. Consolidated bioprocessing of cellulosic biomass to ethanol using thermophilic bacteria, p 55-74. *In* Wall JD, Harwood CS, Demain A, (eds), Bioenergy. ASM Press, Washington.

Lynd LR, Grethlein HE, Wolkin RH. 1989. Fermentation of cellulosic substrates in batch and continuous culture by *Clostridium thermocellum*. Appl Environ Microbiol **55**:3131-3139.

Lynd LR, van Zyl WH, McBride JE, Laser M. 2005. Consolidated bioprocessing of cellulosic biomass: an update. Curr Opin Biotechnol **16**:577–583.

Lynd LR, Weimer PJ, van Zyl WH, Pretorius IS. 2002. Microbial cellulose utilization: fundamentals and biotechnology. Microbiol Mol Biol Rev **66**:506–577.

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

Madden RH, Bryder MJ, Poole NJ. 1982. Isolation and characterization of an anaerobic, cellulolytic bacterium, *Clostridium papyrosolvens* sp. nov. Int J Syst Bacteriol **32**:87–91**.**

Madden RH. 1983. Isolation and characterization of *Clostridium stercorarium* sp. nov., cellulolytic thermophile. Int J Syst Bacteriol **33**:837–840.

Mai V, Wiegel J, Lorenz WW. 2000. Cloning, sequencing, and characterization of the bifunctional xylosidase-arabinosidase from the anaerobic thermophile *Thermoanaerobacter ethanolicus*. Gene **247**:137–143.

Mai V, Wiegel J. 2000. Advances in development of a genetic system for *Thermoanaerobacterium* spp.: expression of genes encoding hydrolytic enzymes, development of a second shuttle vector, and integration of genes into the chromosome. Appl Environ Microbiol **66**:4817–4821.

Mai X, Adams MWW. 1994. Indolepyruvate ferredoxin oxidoreductase from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Biol Chem **269**:16726–16732.

Margeot A, Hahn-Hagerdal B, Edlund M, Slade R, Monot F. 2009. New improvements for lignocellulosic ethanol. Curr Opin Biotechnol **20**:372–380.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes X V, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376–380.

Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res **40**:D115–122.

Markowitz VM, Mavromatis K, Ivanova NN, Chen I-M A, Chu K, Kyrpides NC. 2009. IMG ER: a system for microbial genome annotation expert review and curation. Bioinforma **25**:2271–2278.

Marston EL, Sumner JW, Regnery RL. 1999. Evaluation of intraspecies genetic variation within the 60 kDa heat-shock protein gene (*groEL*) of *Bartonella* species. Int J Syst Bacteriol **49**:1015–1023.

Martin B, Humbert O, Camara M, Guenzi E, Walker J, Mitchell T, Andrew P, Prudhomme M, Alloing G, Hakenbeck R. 1992. A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. Nucleic Acids Res **20**: 3479–3483.

Martín C, Klinke HB, Thomsen AB. 2007. Wet oxidation as a pretreatment method for enhancing the enzymatic convertibility of sugarcane bagasse. Enzyme Microb Technol **40**:426–432.

Mayer MAG, Bronnenmeier K, Schwarz WH, Schertler C, Staudenbauer WL. 1995. Isolation and properties of acetate kinase- and phosphotransacetylase-negative mutants of *Thermoanaerobacter thermohydrosulfuricus*. Microbiol **141**:2891-2896.

McBee RH. 1954. The characteristics of *Clostridium thermocellum*. J Bacteriol **67**:505–506.

Mcbride JE, Rajgarhia V, Shaw AJ, Tripathi SA, Brevnova E, Caiazza N, Van Dijken JP, Froehlich AC, Sillers WR, Flatt JH. December 2012. Production of propanols, alcohols, and polyols in consolidated bioprocessing organisms. US patent 2012/0322078 A1.

McClung LS. 1935. Studies on anaerobic bacteria. IV. Taxonomy of cultures of a thermophilic species causing "swells" of canned food. J Bacteriol **29**:189–202.

McQueen P, Spicer V, Rydzak T, Sparling R, Levin D, Wilkins J A, Krokhin O. 2012. Information-dependent LC-MS/MS acquisition with exclusion lists potentially generated on-the-fly: case study using a whole cell digest of *Clostridium thermocellum*. Proteomics **12:**1160–1169.

Meintanis C, Chalkou KI, Kormas KA, Lymperopoulou DS, Katsifas EA, Hatzinikolaou DG, Karagouni AD. 2008. Application of *rpoB* sequence similarity analysis, REP-PCR and BOX-PCR for the differentiation of species within the genus *Geobacillus*. Lett Appl Microbiol **46**:395–401.

Mendez BS, Pettinar MJ, Ivanier SE, Ramos CA, Sineriz ANDF. 1991. *Clostridium thermopapyrolyticum* sp. nov., a cellulolytic thermophile. Int J Syst Bacteriol **41**:281–283.

Meng B, Qian Z, Wei F, Wang W, Zhou C, Wang Z, Wang Q, Tong W, Wang Q, Ma Y, Xu N, Liu S. 2009. Proteomic analysis on the temperature-dependent complexes in *Thermoanaerobacter tengcongensis*. Proteomics **9**:3189–3200.

Miadenovska Z, Mathrani I, Ahring BK. 1995. Isolation and characterization of *Caldicellulosiruptor lactoaceticus* sp. nov., an extremely thermophilic, cellulolytic, anaerobic bacterium. Arch Microbiol **163**:223–230.

Mikkelsen MJ, Ahring BK. November 2007. *Thermoanaerobacter mathranii* strain BG1. US Patent 20100143998.

Miroshnichenko ML, Kublanov I V, Kostrikina NA, Tourova TP, Kolganova T V, Birkeland N-K, Bonch-Osmolovskaya EA. 2008. *Caldicellulosiruptor kronotskyensis* sp. nov. and *Caldicellulosiruptor hydrothermalis* sp. nov., two extremely thermophilic, cellulolytic, anaerobic bacteria from Kamchatka thermal springs. Int J Syst Evol Microbiol **58**:1492–1496.

Miwa Y, Nakata A, Ogiwara A, Yamamoto M, Fujita Y. 2000. Evaluation and characterization of catabolite-responsive elements (*cre*) of *Bacillus subtilis*. Nucleic Acids Res **28**:1206–1210.

Monedero V, Gosalbes MJ, Perez-Martinez G. 1997. Catabolite repression in *Lactobacillus casei* ATCC 393 is mediated by CcpA . J Bacteriol **179**:6657-6664.

Moore LR, Rocap G, Chisholm SW. 1998. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature **393**:464–467.

Morag E, Bayer E A, Lamed R. 1990. Relationship of cellulosomal and noncellulosomal xylanases of *Clostridium thermocellum* to cellulose-degrading enzymes. J Bacteriol **172**:6098–6105.

Moreira LRS, Milanezi NvG, Filho EXF. 2011. Enzymology of plant cell wall breakdown: an update. p 73-96, *In* Buckeridge MS, Goldman GH (ed), Routes to Cellulosic Ethanol, Springer, New York, NY.

Mori Y. 1990. Characterization of a symbiotic coculture of *Clostridium thermohydrosulfuricum* YM3 and *Clostridium thermocellum* YM4. Appl Environ Microbiol **56:**37–42.

Mori Y. 1995. Nutritional interdependence between *Thermoanaerobacter thermohydrosulfuricus* and *Clostridium thermocellum*. Arch Microbiol **164**:152–154.

Morikawa M, Izawa Y, Rashid N, Hoaki T, Imanaka T. 1994. Purification and characterization of a thermostable thiol protease from a newly isolated hyperthermophilic *Pyrococcus* sp. KOD1. Appl Environ Microbiol **60**:4559-4566.

Morisaka H, Matsui K, Tatsukami Y, Kuroda K, Miyake H, Tamaru Y, Ueda M. 2012. Profile of native cellulosomal proteins of *Clostridium cellulovorans* adapted to various carbon sources. AMB Express **2**:37.

Mosier N, Wyman C, Dale B, Elander R, Lee YY, Holtzapple M, Ladisch M. 2005. Features of promising technologies for pretreatment of lignocellulosic biomass. Bioresour Technol **96**:673–686.

Moukoumi J, Hynes RK, Dumonceaux TJ, Town J, Bélanger N. 2013. Characterization and genus identification of rhizobial symbionts from *Caragana arborescens* in western Canada. Can J Microbiol **59**:399–406.

Mukhopadhyay A, Redding AM, Rutherford BJ, Keasling JD. 2008. Importance of systems biology in engineering microbes for biofuel production. Curr Opin Biotechnol **19**:228–234.

Mulkidjanian AY, Dibrov P, Galperin MY. 2008. The past and present of sodium energetics: may the sodium-motive force be with you. Biochim Biophys Acta **1777**:985-992.

Mussatto SI, Dragone G, Guimarães PMR, Silva JPA, Carneiro LM, Roberto IC,

Vicente A, Domingues L, Teixeira JA. 2010. Technological trends, global market, and challenges of bio-ethanol production. Biotechnol Adv **28**:817–830.

Naik SN, Goud V V., Rout PK, Dalai AK. 2010. Production of first and second generation biofuels: a comprehensive review. Renewable Sustainable Energy Rev **14**:578–597.

Navdaeva V, Zurbriggen A, Waltersperger S, Schneider P, Oberholzer AE, Bähler P, Bächler C, Grieder A, Baumann U, Erni B. 2011 Phosphoenolpyruvate: sugar phosphotransferase system from the hyperthermophilic *Thermoanaerobacter tengcongensis*. Biochem **50**:1184-1193.

Nazina TN, Tourova TP, Poltaraus AB, Novikova E V, Grigoryan AA, Ivanova AE, Lysenko AM, Petrunyaka V V, Osipov GA, Belyaev SS, Ivanov M V. 2001. Taxonomic study of aerobic thermophilic bacilli: descriptions of *Geobacillus subterraneus* gen. nov., sp. nov. and *Geobacillus uzenensis* sp. nov. from petroleum reservoirs and transfer of *Bacillus stearothermophilus, Bacillus thermocatenulatus, Bacillus thermoleovorans, Bacillus kaustophilus, Bacillus thermoglucosidasius* and *Bacillus thermodenitrificans* to *Geobacillus* as the new combinations *G. stearothermophilus, G. thermocatenulatus, G. thermoleovorans, G. kaustophilus, G. thermoglucosidasius* and *G. thermodenitrificans*. Int J Syst Evol Microbiol **51**:433–446.

Nevoigt E. 2008. Progress in metabolic engineering of *Saccharomyces cerevisiae*. Microbiol Mol Biol Rev **72**:379–412.

Newcomb M, Chen C-Y, Wu JHD. 2007. Induction of the *celC* operon of *Clostridium thermocellum* by laminaribiose. Proc Natl Acad Sci U S A **104**:3747-3752.

Ng TK, Ben-Bassat A, Zeikus JG. 1981. Ethanol production by thermophilic bacteria : fermentation of cellulosic substrates by cocultures of *Clostridium thermocellum* and *Clostridium thermohydrosulfuricum*. Appl Environ Microbiol **41**:1337–1343.

Ng TK, Weimer PJ, Zeikus JG. 1977. Cellulolytic and physiological properties of *Clostridium thermocellum*. Arch Microbiol **114**:1–7.

Ng TK, Zeikus JG. 1982. Differential metabolism of cellobiose and glucose by *Clostridium thermocellum* and *Clostridium thermohydrosulfuricum*. J Bacteriol **150**:1391-1399.

Ohta K, Beall DS, Mejia JP, Shanmugam KT, Ingram L. 1991. Genetic improvemnt of *Escherichia coli* for ethanol production: chromosomal integration of *Zymomonas mobilis* genes encoding pyruvate decarboxylase and alcohol dehydrogenase Il. Appl Environ Microbiol **57**:893–900.

Okabe S, Satoh H, Watanabe Y. 1999. In situ analysis of nitrifying biofilms as determined by in situ hybridization and the use of microelectrodes. Appl Environ Microbiol **65**:3182–3191.

Olson DG, McBride JE, Shaw AJ, Lynd LR. 2012. Recent progress in consolidated bioprocessing. Curr Opin Biotechnol **23**:396–405.

Onyenwoke RU, Kevbrin V V, Lysenko AM, Wiegel J. 2007. *Thermoanaerobacter pseudethanolicus* sp. nov., a thermophilic heterotrophic anaerobe from Yellowstone National Park. Int J Syst Evol Microbiol **57**:2191–2193.

O'Sullivan CA, Burrell PC, Clarke WP, Blackall LL. 2005. Structure of a cellulose degrading bacterial community during anaerobic digestion. Biotechnol Bioeng **92**:871–878.

Overbeek R, Begley T, Butler RM, Choudhuri J V, Chuang H-Y, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res **33**:5691–5702.

Park Y-C, Choi J-H, Bennett GN, Seo J-H. 2006. Characterization of D-ribose biosynthesis in *Bacillus subtilis* JY200 deficient in transketolase gene. J Biotechnol **121**:508–516.

Patel G, Khan A, Agnew B, Colvin J. 1980. Isolation and characterization of an anaerobic, cellulolytic microorganism, *Acetivibrio cellulolyticus* gen. nov., sp. nov. Int J Syst Bacteriol **30**:179–185.

Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, Kyrpides NC. 2010. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. Nat Methods **7**:455–457.

Pearson WR, Wood T, Zhang Z, Miller W. 1997. Comparison of DNA sequences with protein sequences. Genomics **46**:24-36.

Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, Jin Z, Lee P, Yang L, Poles M, Brown SM, Sotero S, DeSantis T, Brodie E, Nelson K, Pei Z. 2010a. Diversity of 16S rRNA genes within individual prokaryotic genomes. Appl Environ Microbiol **76**:3886–3897.

Pei J, Zhou Q, Jian Y, Le Y, Li H, Shao W, Wiegel J. 2010b. *Thermoanaerobacter* spp. control ethanol pathway via transcriptional regulation and versatility of key enzymes. Metab Eng **12**:420-428.

Pei J, Zhou Q, Jing Q, Li L, Dai C, Li H, Wiegel J, Shao W. 2011. The mechanism for regulating ethanol fermentation by redox levels in *Thermoanaerobacter ethanolicus*. Metab Eng **13**:186–193.

Peña A, Teeling H, Huerta-Cepas J, Santos F, Yarza P, Brito-Echeverría J, Lucio M, Schmitt-Kopplin P, Meseguer I, Schenowitz C, Dossat C, Barbe V, Dopazo J, Rosselló-Mora R, Schüler M, Glöckner FO, Amann R, Gabaldón T, Antón J. 2010. Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting *Salinibacter ruber* strains. ISME J **4**:882–895.

Peng H, Fu B, Mao Z, Shao W. 2006. Electrotransformation of *Thermoanaerobacter ethanolicus* JW200. Biotechnol Lett **28**:1913–1917.

Peralta-Yahya PP, Zhang F, del Cardayre SB, Keasling JD. 2012. Microbial engineering for the production of advanced biofuels. Nature **488**:320–328.

Petitdemange E, Caillet F, Giallo J, Gaudin C. 1984. *Clostridium cellulolyticum* sp. nov., a cellulolytic, mesophilic species from decayed grass. Int J Syst Bacteriol **34**:155–159.

Piccolo C, Bezzo F. 2009. A techno-economic comparison between two technologies for bioethanol production from lignocellulose. Biomass Bioenergy **33**:478–491.

Pierce E, Xie G, Barabote RD, Saunders E, Han CS, Detter JC, Richardson P,

Brettin TS, Das A, Ljungdahl LG, Ragsdale SW. 2008. The complete genome sequence of *Moorella thermoacetica* (f. *Clostridium thermoaceticum*). Environ Microbiol **10:**2550-2573.

Podkaminer KK, Kenealy WR, Herring CD, Hogsett D A, Lynd LR. 2012. Ethanol and anaerobic conditions reversibly inhibit commercial cellulase activity in thermophilic simultaneous saccharification and fermentation (tSSF). Biotechnol Biofuels **5:**43.

Pontes DS, Lima-Bittencourt CI, Chartone-Souza E, Amaral Nascimento AM. 2007. Molecular approaches: advantages and artifacts in assessing bacterial diversity. J Ind Microbiol Biotechnol **34**:463–473.

Qian Z, Meng B, Wang Q, Wang Z, Zhou C, Wang Q, Tu S, Lin L, Ma Y, Liu S. 2009. Systematic characterization of a novel *gal* operon in *Thermoanaerobacter tengcongensis*. Microbiol **155**:1717–1725.

Rainey P, Buckling A, Kassen R, Travisano M. 2000. The emergence and maintenance of diversity: insights from experimental bacterial populations. Trends Ecol Evol **15**:243–247.

Rainey FA, Donnison AM, Janssen PH, Saul D, Rodrigo A, Bergquist PL, Daniel RM, Stackebrandt E, Morgan HW. 1994. Description of *Caldicellulosiruptor saccharolyticus* gen. nov., sp. nov: an obligately anaerobic, extremely thermophilic, cellulolytic bacterium. FEMS Microbiol Lett **120**:263–266.

Rainey FA, Stackebrandt E. 1993. Transfer of the type species of the genus *Thermobacteroides* to the genus *Thermoanaerobacter* as *Thermoanaerobacter acetoethylicus* (Ben-Bassat and Zeikus 1981) comb. nov., description of

Coprothermobacter gen. nov., and reclassification of *Thermobacteroides proteolyticus* as *Coprothermobacter proteolyticus* (Ollivier et al., 1985) comb. nov. Int J Syst Bacteriol **43**:857–859.

Rainey PB, Travisano M. 1998. Adaptive radiation in a heterogeneous environment. Nature **394**:69–72.

Rainey FA, Ward NL, Morgan HW, Toalster R, Stackebrandt E. 1993. Phylogenetic analysis of anaerobic thermophilic bacteria: Aid for their reclassification. J Bacteriol **175**:4772–4779.

Ramachandran U, Wrana N, Cicek N, Sparling R, Levin DB. 2011. Isolation and characterization of a hydrogen- and ethanol-producing *Clostridium* sp. strain URNW. Can J Microbiol **57**:236–243.

Raman B, McKeown CK, Rodriguez M, Brown SD, Mielenz JR. 2011. Transcriptomic analysis of *Clostridium thermocellum* ATCC 27405 cellulose fermentation. BMC Microbiol **11**:134.

Raman B, Pan C, Hurst GB, Rodriguez M, McKeown CK, Lankford PK, Samatova NF, Mielenz JR. 2009. Impact of pretreated Switchgrass and biomass carbohydrates on *Clostridium thermocellum* ATCC 27405 cellulosome composition: a quantitative proteomic analysis. PloS One 4:e5271.

Ramos LP, Breuil C, Saddler JN. 1993. The use of enzyme recycling and the influence of sugar accumulation on cellulose hydrolysis by *Trichoderma* cellulases. Enzyme Microb Technol **15**:19–25.

Ramsing NB, Ferris MJ, Ward DM. 2000. Highly ordered vertical structure of

Synechococcus populations within the one-millimeter-thick photic zone of a hot spring cyanobacterial mat. Appl Environ Microbiol **66**:1038–1049.

Rasko DA, Ravel J, Økstad OA, Helgason E, Cer RZ, Jiang L, Shores KA, Fouts DE, Tourasse NJ, Angiuoli S V, Kolonay J, Nelson WC, Kolstø A-B, Fraser CM, Read TD. 2004. The genome sequence of *Bacillus cereus* ATCC 10987 reveals metabolic adaptations and a large plasmid related to *Bacillus anthracis* pXO1. Nucleic Acids Res **32**:977–988.

Reddy AP, Allgaier M, Singer SW, Hazen TC, Simmons BA, Hugenholtz P, VanderGheynst JS. 2011. Bioenergy feedstock-specific enrichment of microbial populations during high-solids thermophilic deconstruction. Biotechnol Bioeng **108**:2088–2098.

Reddy N, Yang Y. 2005. Biofibers from agricultural byproducts for industrial applications. Trends Biotechnol **23**:22–27.

Ren N, Cao G, Wang A, Lee D-J, Guo W, Zhu Y. 2008. Dark fermentation of xylose and glucose mix using isolated *Thermoanaerobacterium thermosaccharolyticum* W16. Int J Hydrogen Energy **33**:6124–6132.

Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A **106**:19126–19131.

Riessen S, Antranikian G. 2001. Isolation of *Thermoanaerobacter keratinophilus* sp. nov., a novel thermophilic, anaerobic bacterium with keratinolytic activity. Extremophiles **5**:399–408.

Rigottier-Gois L, Rochet V, Garrec N, Suau A, Doré J. 2003. Enumeration of

Bacteroides species in human faeces by fluorescent *in situ* hybridisation combined with flow cytometry using 16S rRNA probes. Syst Appl Microbiol **26**:110–118.

Roberts MJ, Schlenker W. 2010. Identifying supply and demand elasticities of agricultural commodities: implications for the US ethanol mandate. National Bureau of Economic Research Working Paper Series. Working paper 15921: http://www.nber.org/papers/w15921.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature **424**:1042–1047.

Roh Y, Liu S V, Li G, Huang H, Phelps TJ, Zhou J. 2002. Isolation and characterization of metal-reducing *Thermoanaerobacter* strains from deep subsurface environments of the Piceance Basin, Colorado. Appl Environ Microbiol **68**:6013–6020.

Römling U, Schmidt KD, Tümmler B. 1997. Large genome rearrangements discovered by the detailed analysis of *21 Pseudomonas aeruginosa* clone C isolates found in environment and disease habitats. J Mol Biol **271**:386–404.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinforma **19**:1572–1574.

Rosegrant MW, Zhu T, Msangi S, Sulser T. 2008. Global scenarios for biofuels:

impacts and implications. Appl Econ Perspectives Policy **30**:495–505.

Rosillo-Calle F, Walter A. 2006. Global market for bioethanol: historical trends and future prospects. Energy Sustain Dev **10**:20–32.

Rubin JE, Harms NJ, Fernando C, Soos C, Detmer SE, Harding JCS, Hill JE. 2013. Isolation and characterization of *Brachyspira* spp. including "*Brachyspira hampsonii*" from lesser snow geese (*Chen caerulescens caerulescens*) in the Canadian Arctic. Microb Ecol. DOI 10.1007/s00248-013-0273-5.

Rydzak T, Levin DB, Cicek N, Sparling R. 2009. Growth phase-dependant enzyme profile of pyruvate catabolism and end-product formation in *Clostridium thermocellum* ATCC 27405. J Biotechnol **140**:169–75.

Rydzak T, McQueen PD, Krokhin O V, Spicer V, Ezzati P, Dwivedi RC, Shamshurin D, Levin DB, Wilkins J A, Sparling R. 2012. Proteomic analysis of *Clostridium thermocellum* core metabolism: relative protein expression profiles and growth phase-dependent changes in protein expression. BMC Microbiol **12**:214.

Saddler JN, Chan MK. 1984. Conversion of pretreated lignocellulosic substrates to ethanol by *Clostridium thermocellum* in mono- and co-culture with *Clostridium thermosaccharolyticum* and *Clostridium thermohydrosulphuricum*. Can J Microbiol **30**:212–220.

Saha BC. 2003. Hemicellulose bioconversion. J Ind Microbiol Biotechnol **30**:279-291.

Saha BC, Cotta MA. 2012. Ethanol production from lignocellulosic biomass by recombinant *Escherichia coli* strain FBR5. Bioengineered **3**:197–202.

Saier MH, Yen MR, Noto K, Tamang DG, Elkan C. 2009. The Transporter

Classification Database: recent advances. Nucleic Acids Res **37**:D274–278.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4**:406–425.

Sander R. 1999. Compilation of Henry's law constants for inorganic and organic species of potential important in environmental chemistry (Version 3). Air chemistry department, Max-Planck Institute of Chemistry. Mainz, Germany.

Sauer U, Lasko DR, Fiaux J, Glaser R, Szyperski T, Wüthrich K, Bailey JE, Hochuli M, Wu K. 1999. Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism. J Bacteriol **181**:6679–6688.

Savage N. 2011. The ideal biofuel. Nature **474**:S9–S11.

Schellenberg JJ, Verbeke TJ, McQueen P, Spicer V, Krokhin O, Zhang X, Fristensky B, Thallinger GG, Henrissat B, Wilkins JA, Levin DB, Sparling R. Whole genome sequence enhancement and molecular gene expression profile of *Clostridium stercorarium* DSM8532^{T} by RNA-seq transcriptomics and high-throughput proteomics. BMC Genomics, submitted for publication.

Scheller HV, Ulvskov P. 2010. Hemicelluloses. Ann Rev Plant Biol **61**:263–289.

Schloss PD, Handelsman J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol **71**:1501–1506.

Schloter M, Lebuhn M, Heulin T, Hartmann A. 2000. Ecology and evolution of bacterial microdiversity. FEMS Microbiol Rev **24**:647–660.

Schmid U, Giesel H, Schoberth SM, Sahm H. 1986. *Thermoanaerobacter finnii* spec.

nov., a new ethanologenic sporogeneous bacterium. Syst Appl Microbiol **8**:80-85.

Schmidt AS, Thomsen AB. 1998. Optimization of wet oxidation pretreatment of wheat straw. Bioresour Technol **64**:139–151.

Schübbe S, Kube M, Scheffel A, Wawer C, Heyen U, Meyerdierks A, Madkour MH, Mayer F, Reinhardt R, Schüler D. 2003. Characterization of a spontaneous nonmagnetic mutant of *Magnetospirillum gryphiswaldense* reveals a large deletion comprising a putative magnetosome island. J Bacteriol **185**:5779–5790.

Schumacher MA, Allen GS, Diel M, Seidel G, Hillen W, Brennan RG. 2004. Structural basis for allosteric control of the transcription regulator CcpA by the phosphoprotein HPr-Ser46-P. Cell **118**:731–741.

Schumacher MA, Seidel G, Hillen W, Brennan RG. 2006. Phosphoprotein Crh-Ser46-P displays altered binding to CcpA to effect carbon catabolite regulation. J Biol Chem **281**:6793–6800.

Schut GJ, Adams MWW. 2009. The iron-hydrogenase of *Thermotoga maritima* utilizes ferredoxin and NADH synergistically: a new perspective on anaerobic hydrogen production. J Bacteriol **191**:4451–7.

Schut GJ, Bridger SL, Adams MWW. 2007. Insights into the metabolism of elemental sulfur by the hyperthermophilic archaeon *Pyrococcus furiosus*: characterization of a coenzyme A-dependent NAD(P)H sulfur oxidoreductase. J Bacteriol **189**:4431-4441.

Shallom D, Shoham Y. 2003. Microbial hemicellulases. Curr Opin Microbiol **6**:219-228.

Sharma PK, Fu J, Cicek N, Sparling R, Levin DB. 2012. Kinetics of medium-chain-

length polyhydroxyalkanoate production by a novel isolate of *Pseudomonas putida* LS46. Can J Microbiol **989**:982–989.

Shao W, Blois SDE, Wiegel J. 1995. A high-molecular-weight, cell-associated xylanase isolated from exponentially growing *Thermoanaerobacterium* sp. strain JW/SL-YS485. Appl Environ Microbiol **61**:937–940.

Shaw AJ, Covalla SF, Hogsett DA, Herring CD. 2011a. Marker removal system for *Thermoanaerobacterium saccharolyticum* and development of a markerless ethanologen. Appl Environ Microbiol **77:**2534–2536.

Shaw AJ, Covalla SF, Miller BB, Firliet BT, Hogsett DA, Herring CD. 2012. Urease expression in a *Thermoanaerobacterium saccharolyticum* ethanologen allows high titer ethanol production. Metab Engin **14**:528–532.

Shaw AJ, Hogsett DA, Lynd LR. 2010. Natural competence in *Thermoanaerobacter* and *Thermoanaerobacterium* species. Appl Environ Microbiol **76**:4713–4719.

Shaw AJ, Jenney FE, Adams MWW, Lynd LR. 2008a. End-product pathways in the xylose fermenting bacterium, *Thermoanaerobacterium saccharolyticum*. Enzyme Microb Technol **42**:453–458.

Shaw AJ, Podkaminer KK, Desai SG, Bardsley JS, Rogers SR, Thorne PG, Hogsett D A, Lynd LR. 2008b. Metabolic engineering of a thermophilic bacterium to produce ethanol at high yield. Proc Natl Acad Sci U S A **105**:13769–13774.

Shaw AJ, Sillers WR, Miller BB, Folden J, Bhandiwad A, Rogers SR, Kenealy WR, Hogsett DA, Herring CD. 2011b. Abstr. SIM Ann Meet Exhibit., abstr. S124. Conversion of acetic acid to acetone and isopropanol in *Thermoanaerobacterium saccharolyticum*.

Shi W, Ding S-Y, Yuan JS. 2011. Comparison of insect gut cellulase and xylanase activity across different insect species with distinct food sources. BioEnergy Res **4**:1–10.

Shin H-D, McClendon S, Vo T, Chen RR. 2010. *Escherichia coli* binary culture engineered for direct fermentation of hemicellulose to a biofuel. Appl Environ Microbiol **76**:8150–8159.

Shockley KR, Scott KL, Pysz MA, Conners SB, Johnson MR, Clemente I, Wolfinger RD, Kelly RM, Shockley KR, Scott KL, Pysz MA, Conners SB, Johnson MR, Montero CI, Wolfinger RD, Kelly RM. 2005. Genome-wide transcriptional variation within and between steady states for continuous growth of the hyperthermophile *Thermotoga maritima*. Appl Environ Microbiol **71**:5572–5576.

Shong J, Diaz MRJ, Collins CH. 2012. Towards synthetic microbial consortia for bioprocessing. Curr Opin Biotechnol **23**:798–802.

Silverstein RA, Chen Y, Sharma-Shivappa RR, Boyette MD, Osborne J. 2007. A comparison of chemical pretreatment methods for improving saccharification of cotton stalks. Bioresour Technol **98**:3000–11.

Sjolander NO. 1937. Studies on anaerobic bacteria. XII. The fermentation products of *Clostridium thermosaccharolyticum*. J Bacteriol **34**:419–428.

Skinner KA, Leathers TD. 2004. Bacterial contaminants of fuel ethanol production. J Ind Microbiol Biotechnol **31**:401–408.

Sleat R, Mah RA, Robinson R. 1984. Isolation and characterization of an anaerobic, cellulolytic bacterium, *Clostridium cellulovorans* sp. nov. Appl Environ Microbiol **48**:88–93.

Slobodkin AI, Tourova TP, Kuznetsov BB, Kostrikina NA, Chernyh NA, Bonch-Osmolovskaya EA. 1999. *Thermoanaerobacter siderophilus* sp., nov., a novel dissimilatory Fe(III)-reducing, anaerobic, thermophilic bacterium. Int J Syst Bacteriol **49**:1471–1478.

Soboh B, Linder D, Hedderich R. 2004. A multisubunit membrane-bound [NiFe] hydrogenase and an NADH-dependent Fe-only hydrogenase in the fermenting bacterium *Thermoanaerobacter tengcongensis*. Microbiol **150**:2451–2463.

Solomon BD, Barnes JR, Halvorsen KE. 2007. Grain and cellulosic ethanol: History, economics, and energy policy. Biomass Bioenergy **31**:416–425.

Songstad D, Lakshmanan P, Chen J, Gibbons W, Hughes S, Nelson R. 2009. Historical perspective of biofuels: learning from the past to rediscover the future. In Vitro Cell Dev Biol **45**:189-192.

Sparling R, Islam R, Cicek N, Carere C, Chow H, Levin DB. 2006. Formate synthesis by *Clostridium thermocellum* during anaerobic fermentation. Can J Microbiol **52**:681–688.

Sprenger GA. 1993. Nucleotide sequence of the *Escherichia coli* K-12 transketolase (*tkt*) gene. Biochim Biophys Acta **1216**:307–310.

Stackebrandt E, Frederiksen W, Garrity GM, Grimont PAD, Kampfer P, Maiden MCJ, Nesme X, Rosselló-Móra R, Swings J, Trüper HG, Vauterin L, Ward AC, Whitman WB. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol **52**:1043–1047.

Staley JT. 2006. The bacterial species dilemma and the genomic-phylogenetic species concept. Philos Trans R Soc Lond B Biol Sci **361**:1899–1909.

Steen EJ, Chan R, Prasad N, Myers S, Petzold CJ, Redding A, Ouellet M, Keasling JD. 2008. Metabolic engineering of *Saccharomyces cerevisiae* for the production of n-butanol. Microb Cell Factories **7**:36.

Stülke J, Arnaud M, Rapoport G, Martin-Verstraete, I. 1998. PRD – a protein domain involved in PTS-dependent induction and carbon catabolite repression of catabolic operons in bacteria. Mol Microbiol **28**:865–874

Sun Y, Cheng J. 2002. Hydrolysis of lignocellulosic materials for ethanol production: a review. Bioresour Technol **83**:1-11.

Suzuki Y, Kishigami T, Inoue K, Mizoguchi Y, Eto N, Takagi M, Abe S. 1983. *Bacillus thermoglucosidasius* sp. nov., a new species of obligately thermophilic bacilli. Syst Appl Microbiol **20**:487-495.

Svetlichnyi VA, Svetlichnaya TP, Chernykh NA, Zavarzin GA. 1990. *Anaerocellum thermophilum* gen. nov., sp. nov., an extremely thermophilic cellulolytic eubacterium isolated from hot-springs in the valley of Geyers. Microbiol **59**:598-604.

Svetlitchnyi VA, Kensch O, Falkenhan DA, Korseska SG, Lippert N, Prinz M, Sassi J, Schickor A, Curvers S. 2013. Single-step ethanol production from lignocellulose using novel extremely thermophilic bacteria. Biotechnol Biofuels **6**:31.

Tacão M, Alves A, Saavedra MJ, Correia A. 2005. BOX-PCR is an adequate tool for typing *Aeromonas* spp. Antonie Van Leeuwenhoek **88**:173–179.

Tai S-K, Lin H-PP, Kuo J, Liu J-K. 2004. Isolation and characterization of a

cellulolytic *Geobacillus thermoleovorans* T4 strain from sugar refinery wastewater. Extremophiles **8**:345–349.

Takahata Y, Hoaki T, Maruyama T. 2001. *Thermotoga petrophila* sp. nov. and *Thermotoga naphthophila* sp. nov., two hyperthermophilic bacteria from the Kubiki oil reservoir in Niigata , Japan. Int J Syst Evol Microbiol **51**:1901–1909.

Takasaki K, Miura T, Kanno M, Tamaki H, Hanada S, Kamagata Y, Kimura N. 2013. Discovery of glycoside hydrolase enzymes in an avicel-adapted forest soil fungal community by a metatranscriptomic approach. PloS One **8**:e55485.

Tamaru Y, Miyake H, Kuroda K, Nakanishi A, Matsushima C, Doi RH, Ueda M. 2011. Comparison of the mesophilic cellulosome-producing *Clostridium cellulovorans* genome with other cellulosome-related clostridial genomes. Microb Biotechnol **4**:64–73.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol **24**:1596-1599.

Tanaka K, Komiyama A, Sonomoto K, Ishizaki A, Hall SJ, Stanbury PF. 2002. Two different pathways for D-xylose metabolism and the effect of xylose concentration on the yield coefficient of L-lactate in mixed-acid fermentation by the lactic acid bacterium *Lactococcus lactis* IO-1. Appl Microbiol Biotechnol **60**:160–167.

Tartar A, Wheeler MM, Zhou X, Coy MR, Boucias DG, Scharf ME. 2009. Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite *Reticulitermes flavipes*. Biotechnol Biofuels **2**:25.

Tatusov RL, Galperin MY, Natale D a, Koonin E V. 2000. The COG database: a tool

for genome-scale analysis of protein functions and evolution. Nucleic Acids Res **28**:33–36.

Taylor BL, Zhulin IB. 1999. PAS Domains: internal sensors of oxygen, redox potential, and light. Microbiol Mol Biol Rev **63**:479-506.

Taylor MP, Eley KL, Martin S, Tuffin MI, Burton SG, Cowan D A. 2009. Thermophilic ethanologenesis: future prospects for second-generation bioethanol production. Trends Biotechnol **27**:398–405.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22**:4673–4680.

Toivari MH, Aristidou A, Ruohonen L, Penttilä M. 2001. Conversion of xylose to ethanol by recombinant *Saccharomyces cerevisiae*: importance of xylulokinase (XKS1) and oxygen availability. Metab Eng **3**:236–249.

Tolonen AC, Haas W, Chilaka AC, Aach J, Gygi SP, Church GM. 2011. Proteome-wide systems analysis of a cellulosic biofuel-producing microbe. Mol Syst Biol **7**:461.

Tong W, Chen Z, Cao Z, Wang Q, Zhang J, Bai X, Wang R, Liu S. 2013. Robustness analysis of a constraint-based metabolic model links cell growth and proteomics of *Thermoanaerobacter tengcongensis* under temperature perturbation. Mol Biosyst **9**:713–722.

Tracy BP, Jones SW, Fast AG, Indurthi DC, Papoutsakis ET. 2012. Clostridia: the

importance of their exceptional substrate and metabolite diversity for biofuel and biorefinery applications. Curr Opin Biotechnol **23**:364–381.

Tripathi SA, Olson DG, Argyros DA, Miller BB, Barrett TF, Murphy DM, McCool JD, Warner AK, Rajgarhia VB, Lynd LR, Hogsett DA, Caiazza NC. 2010. Development of *pyrF*-based genetic system for targeted gene deletion in *Clostridium thermocellum* and creation of a *pta* mutant. Appl Environ Microbiol **76**:6591–6599.

Truscott KN, Høj PB, Scopes RK. 1994. Purification and characterization of chaperonin 60 and chaperonin 10 from the anaerobic thermophile *Thermoanaerobacter brockii*. Eur J Biochem **222**:277–284.

Truscott KN, Scopes RK. 1998. Sequence analysis and heterologous expression of the *groE* genes from *Thermoanaerobacter* sp. Rt8.G4. Gene **217**:15–23.

Tsakraklides V, Shaw AJ, Miller BB, Hogsett DA, Herring CD. 2012. Carbon catabolite repression in *Thermoanaerobacterium saccharolyticum*. Biotechnol Biofuels **5**:85.

Turner P, Mamo G, Karlsson EN. 2007. Potential and utilization of thermophiles and thermostable enzymes in biorefining. Microb cell factories **6**:9.

Turner PE, Souza V, Lenski RE. 1996. Tests of ecological mechanisms promoting the stable coexistence of two bacterial genotypes. Ecol **77**:2119-2129.

Tyner WE, Viteri D. 2010. Implications of blending limits on the US ethanol and biofuels markets. Biofuels **1**:251-253.

Untergrasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3 – new capabilities and interfaces. Nucleic Acids Res **40**:e115.

Urzì C, Brusetti L, Salamone P, Sorlini C, Stackebrandt E, Daffonchio D. 2001. Biodiversity of *Geodermatophilaceae* isolated from altered stones and monuments in the Mediterranean basin. Environ Microbiol **3**:471–479.

Valdez-Vazquez I, Sparling R, Risbey D, Rinderknecht-Seijas N, Poggi-Varaldo HM. 2005. Hydrogen generation via anaerobic fermentation of paper mill wastes. Bioresour Technol **96**:1907–1913.

van de Werken HJG, Verhaart MR a, VanFossen AL, Willquist K, Lewis DL, Nichols JD, Goorissen HP, Mongodin EF, Nelson KE, van Niel EWJ, Stams AJM, Ward DE, de Vos WM, van der Oost J, Kelly RM, Kengen SWM. 2008. Hydrogenomics of the extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus*. Appl Environ Microbiol **74**:6720–6729.

van der Veen D, Lo J, Brown SD, Johnson CM, Tschaplinski TJ, Martin M, Engle NL, van den Berg RA, Argyros AD, Caiazza NC, Guss AM, Lynd LR. 2013. Characterization of *Clostridium thermocellum* strains with disrupted fermentation end-product pathways. J Ind Microbiol Biotechnol **40**:725–734.

Van Ham RCHJ, González-Candelas F, Silva FJ, Sabater B, Moya A, Latorre A. 2000. Postsymbiotic plasmid acquisition and evolution of the *repA1*-replicon in *Buchnera aphidicola*. Proc Natl Acad Sci U S A **97**:10855–10860.

van Niel EWJ, Budde MAW, Haas GG De, Wal FJ Van Der, Claassen PAM, Stams AJM. 2002. Distinctive properties of high hydrogen producing extreme thermophiles, *Caldicellulosiruptor saccharolyticus* and *Thermotoga elfii*. Int J Hydrogen Energy 27:1391–1398.

van Zyl WH, Lynd LR, den Haan R, McBride JE. 2007. Consolidated bioprocessing

for bioethanol production using *Saccharomyces cerevisiae*. Adv Biochem Eng Biotechnol **108**:205–235.

VanFossen AL, Ozdemir I, Zelin SL, Kelly RM. 2011. Glycoside hydrolase inventory drives plant polysaccharide deconstruction by the extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus*. Biotechnol Bioeng **108**:1559–1569.

Vanfossen AL, Verhaart MRA, Kengen SMW, Kelly RM. 2009. Carbohydrate utilization patterns for the extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus* reveal broad growth substrate preferences. Appl Environ Microbiol **75**:7718–24.

Verhaart MRA, Bielen AAM, van der Oost J, Stams AJM, Kengen SWM. 2010. Hydrogen production by hyperthermophilic and extremely thermophilic bacteria and archaea: mechanisms for reductant disposal. Environ Technol **31**:993–1003.

Vermette CJ, Russell AH, Desai AR, Hill JE. 2010. Resolution of phenotypically distinct strains of *Enterococcus* spp. in a complex microbial community using *cpn60* universal target sequencing. Microb Ecol **59**:14–24.

Versalovic J, Schneider M, de Bruijn FJ, Lupski JR. 1994. Genomic fingerprinting of bacteria using repetitive sequence-based polymerase chain reaction. Methods Mol Cell Biol **5**:25-40.

Viana R, Monedero V, Dossonnet V, Vadeboncoeur C, Pérez-Martínez G, Deutscher J. 2000. Enzyme I and HPr *from Lactobacillus casei*: their role in sugar transport, carbon catabolite repression and inducer exclusion. Mol Microbiol **36**:570–584.

Viikari L, Alapuranen M, Puranen T, Vehmaanperä J, Siika-Aho M. 2007.

Thermostable Enzymes in Lignocellulose Hydrolysis. Adv Biochem Eng Biotechnol **108**:121–145.

Wang A, Ren N, Shi Y, Duu-Jong L. 2008. Bioaugmented hydrogen production from microcrystalline cellulose using co-culture—*Clostridium acetobutylicum* X9 and *Ethanoligenens harbinense* B49. Int J Hydrogen Energy **33**:912–917.

Wang H, Lehtomäki A, Tolvanen K, Puhakka J, Rintala J. 2009. Impact of crop species on bacterial community structure during anaerobic co-digestion of crops and cow manure. Bioresour Technol **100**:2311–2315.

Wang J, Xue Y, Feng X, Li X, Wang H, Li W, Zhao C, Cheng X, Ma Y, Zhou P, Yin J, Bhatnagar A, Wang R, Liu S. 2004. An analysis of the proteomic profile for *Thermoanaerobacter tengcongensis* under optimal culture conditions. Proteomics **4**:136–150.

Wang J, Zhao C, Meng B, Xie J, Zhou C, Chen X, Zhao K, Shao J, Xue Y, Xu N, Ma Y, Liu S. 2007a. The proteomic alterations of *Thermoanaerobacter tengcongensis* cultured at different temperatures. Proteomics **7**:1409–1419.

Wang L-T, Lee F-L, Tai C-J, Kasai H. 2007b. Comparison *of gyrB* gene sequences, 16S rRNA gene sequences and DNA-DNA hybridization in the *Bacillus subtilis* group. Int J Syst Evol Microbiol **57**:1846–1850.

Wang S, Huang H, Moll J, Thauer RK. 2010. $NADP^+$ reduction with reduced ferredoxin and $NADP^+$ reduction with NADH are coupled via an electron-bifurcating enzyme complex in *Clostridium kluyveri*. J Bacteriol **192**:5115–5123.

Warnecke F, Luginbühl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT,

Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, Sorek R, Tringe SG, Podar M, Martin HG, Kunin V, Dalevi D, Madejska J, Kirton E, Platt D, Szeto E, Salamov A, Barry K, Mikhailova N, Kyrpides NC, Matson EG, Ottesen EA, Zhang X, Hernández M, Murillo C, Acosta LG, Rigoutsos I, Tamayo G, Green BD, Chang C, Rubin EM, Mathur EJ, Robertson DE, Hugenholtz P, Leadbetter JR. 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature **450**:560–565.

Warner JB, Lolkema JS. 2003. CcpA-dependent carbon catabolite repression in Bacteria. Microbiol Mol Biol Rev **67**:475–490.

Warnick TA, Methé BA, Leschine SB. 2002. *Clostridium phytofermentans* sp. nov., a cellulolytic mesophile from forest soil. Int J Syst Evol Microbiol **52**:1155–1160.

Weber C, Farwick A, Benisch F, Brat D, Dietz H, Subtil T, Boles E. 2010. Trends and challenges in the microbial production of lignocellulosic bioalcohol fuels. App Microbiol Biotechnol **87**:1303–1315.

Weimer PJ. 2011. End product yields from the extraruminal fermentation of various polysaccharide, protein and nucleic acid components of biofuels feedstocks. Bioresour Technol **102**:3254–3259.

Weng FY, Chiou CS, Lin PHP, Yang SS. 2009. Application of *recA* and *rpoB* sequence analysis on phylogeny and molecular identification of *Geobacillus* species. J Appl Microbiol **107**:452–464.

Wiegel J. 2009. Family I. *Thermoanaerobacteraceae* fam. nov. p 1225-1267. *In* De Vos P, Garrity GM, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, Whitmas WB (ed), Phylum XIII. *Firmicutes* Gibbons and Murray 1978, 5 (*Firmicutes* [sic]

Gibbons and Murray 1978, 5); Bergey's Manual of Systematic Bacteriology, vol. 3, Springer, New York, NY.

Wiegel J, Ljungdahl LG, Rawson JR. 1979. Isolation from soil and properties of the extreme thermophile *Clostridium thermohydrosulfuricum*. J Bacteriol **139**:800–810.

Wiegel J, Ljungdahl LG. 1981. *Thermoanaerobacter ethanolicus* gen. nov., spec. nov., a new, extreme thermophilic, anaerobic bacterium. Arch Microbiol **128**:343–348.

Wiegel J, Mothershed CP, Puls J. 1985. Differences in xylan degradation by various noncellulolytic thermophilic anaerobes and *Clostridium thermocellum*. App Environ Microbiol **49**:656–659.

Wilson M, Lindow SE. 1994. Ecological similarity and coexistence of epiphytic ice-nucleating (Ice^+) *Pseudomonas syringae* strains and a non-ice-nucleating (Ice^-) biological control agent. Appl Environ Microbiol **60**:3128–3137.

Wiśniewski JR, Zougman A, Nagaraj N, Mann M. 2009. Universal sample preparation method for proteome analysis. Nat Methods **6**:359–362.

Woese CR. 1987. Bacterial evolution. Microbiol Rev **51**:221–271.

Wong CCL, Cociorva D, Miller CA, Schmidt A, Monell C, Aebersold R, Yates JR. 2013. Proteomics *of Pyrococcus furiosus* (*Pfu*): identification of extracted proteins by three independent methods. J Proteome Res **12**:763-770.

Wongwilaiwalin S, Rattanachomsri U, Laothanachareon T, Eurwilaichitr L,

Igarashi Y, Champreda V. 2010. Analysis of a thermophilic lignocellulose degrading microbial consortium and multi-species lignocellulolytic enzyme system. Enzyme Microb Technol **47**:283–290.

Wyman CE, Dale BE, Elander RT, Holtzapple M, Ladisch MR, Lee YY. 2005. Coordinated development of leading biomass pretreatment technologies. Bioresour Technol **96**:1959–1966.

Xiao Z, Zhang X, Gregg DJ, Saddler JN. 2004. Effects of sugar inhibition on cellulases and β-glucosidase during enzymatic hydrolysis of softwood substrates. Appl Biochem Biotechnol **113-116**:1115–1126.

Xing D, Ren N, Li Q, Lin M, Wang A, Zhao L. 2006. *Ethanoligenens harbinense* gen. nov., sp. nov., isolated from molasses wastewater. Int J Syst Evol Microbiol **56**:755–760.

Xu C, Qin Y, Li Y, Ji Y, Huang J, Song H, Xu J. 2010. Factors influencing cellulosome activity in consolidated bioprocessing of cellulosic ethanol. Bioresour Technol **101**:9560–9569.

Xu L, Tschirner U. 2011. Improved ethanol production from various carbohydrates through anaerobic thermophilic co-culture. Bioresour Technol **102**:10065–10071.

Xue Y, Xu Y, Liu Y, Ma Y, Zhou P. 2001. *Thermoanaerobacter tengcongensis* sp. nov., a novel anaerobic, saccharolytic, thermophilic bacterium isolated from a hot spring in Tengcong, China. Int J Syst Evol Microbiol **51**:1335–1341.

Yang S-J, Kataeva I, Hamilton-Brehm SD, Engle NL, Tschaplinski TJ, Doeppke C, Davis M, Westpheling J, Adams MWW. 2009. Efficient degradation of lignocellulosic plant biomass, without pretreatment, by the thermophilic anaerobe

"*Anaerocellum thermophilum*" DSM 6725. Appl Environ Microbiol **75**:4762–4769.

Yang S-J, Kataeva I, Wiegel J, Yin Y, Dam P, Xu Y, Westpheling J, Adams MWW. 2010. Classification of '*Anaerocellum thermophilum'* strain DSM 6725 as *Caldicellulosiruptor bescii* sp. nov. Int J Syst Evol Microbiol **60**:2011–2015.

Yao S, Mikkelsen MJ. 2010. Metabolic engineering to improve ethanol production in *Thermoanaerobacter mathranii*. Appl Microbiol Biotechnol **88**:199–208.

Ye JJ, Reizer J, Cui X, Saier MH. 1994. ATP-dependent phosphorylation of serine-46 in the phosphocarrier protein HPr regulates lactose/H^+ symport in *Lactobacillus brevis*. Proc Natl Acad Sci U S A **91**:3102–3106.

Ye JJ, Saier MH. 1995. Cooperative binding of lactose and the phosphorylated phosphocarrier protein HPr (Ser-P) to the lactose/H^+ symport permease of *Lactobacillus brevis*. Proc Natl Acad Sci U S A **92**:417–421.

Yeoman CJ, Han Y, Dodd D, Schroeder CM, Mackie RI, Cann IKO. 2010. Thermostable enzymes as biocatalysts in the biofuel industry. Adv Appl Microbiol. **70**:1–55.

Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FSL. 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinforma **26**:1608–15.

Zeigler DR. 2003. Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int J Syst Evol Microbiol **53**:1893–1900.

Zeigler DR. 2005. Application of a *recN* sequence similarity analysis to the identification

of species within the bacterial genus *Geobacillus*. Int J Syst Evol Microbiol **55**:1171–1179.

Zeikus JG, Ben-Bassat A, Hegge PW. 1980. Microbiology of methanogenesis in thermal, volcanic environments. J Bacteriol **143**:432–440.

Zeikus JG, Hegge PW, Anderson MA. 1979. *Thermoanaerobium brockii* gen. nov. and sp. nov., a new chemoorganotrophic, caldoactive, anaerobic bacterium. Arch Microbiol **122**:41–48.

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

Zhang Y-HP, Lynd LR. 2005. Cellulose utilization by *Clostridium thermocellum*: bioenergetics and hydrolysis product assimilation. Proc Natl Acad Sci U S A **102**:7321–7325.

Zhou Q, Shao W-L. 2010. Molecular genetic characterization of the thermostable L-Lactate dehydrogenase gene (*ldhL*) of *Thermoanaerobacter ethanolicus* JW200 and biochemical characterization of the enzyme. Biochem (Mosc) **75**:526–530.

Zoetendal EG, Ben-Amor K, Hermie JM, Schut F, Akkermans ADL, de Vos M. 2002. Quantification of uncultured *Ruminococcus obeum*-like bacteria in human fecal samples by fluorescent in situ hybridization and flow cytometry using 16S rRNA-targeted probes. Appl Environ Microbiol **68:**4225–4232.

Zuroff TR, Curtis WR. 2012. Developing symbiotic consortia for lignocellulosic biofuel production. Appl Microbiol Biotechnol **93**:1423–35.

Zuroff TR, Xiques SB, Curtis WR. 2013. Consortia-mediated bioprocessing of

cellulose to ethanol with a symbiotic *Clostridium phytofermentans*/yeast co-culture. Biotechnol Biofuels **6**:59.

Zverlov VV, Kellermann J, Schwarz WH. 2005. Functional subgenomics of *Clostridium thermocellum* cellulosomal genes: identification of the major catalytic components in the extracellular complex and detection of three new enzymes. Proteomics **5**:3646–3653.

Zverlov VV, Klupp M, Krauss J, Schwarz WH. 2008. Mutations in the scaffoldin gene, *cipA*, of *Clostridium thermocellum* with impaired cellulosome formation and cellulose hydrolysis: Insertions of a new transposable element, IS1447, and implications for cellulase synergism on crystalline cellulose. J Bacteriol **190**:4321–4327.

Zverlov VV, Schantz N, Schmitt-Kopplin P, Schwarz WH. 2005. Two new major subunits in the cellulosome of *Clostridium thermocellum*: xyloglucanase Xgh74A and endoxylanase Xyn10D. Microbiol **151**:3395–3401.

Zverlov VV, Schwarz WH. 2008. Bacterial cellulose hydrolysis in anaerobic environmental subsystems-*Clostridium thermocellum* and *Clostridium stercorarium*, thermophilic plant-fiber degraders. Ann N Y Acad Sci **1125**:298–307.