Genome-scale resources for *Thermoanaerobacterium saccharolyticum*

Devin H. Currie[1], Babu Raman[2,3], Christopher M. Gowen[4,5], Timothy J. Tschaplinski[2], Miriam L. Land[2], Steven D. Brown[2], Sean F. Covalla[1], Dawn M. Klingeman[2], Zamin K. Yang[2], Nancy L. Engle[2], Courtney M. Johnson[2], Miguel Rodriguez[2],A. Joe Shaw[1,6], William R. Kenealy[1], Stephen S. Fong[4], Jonathan R. Mielenz[2], Brian H. Davison[2], David A. Hogsett[1], Christopher D. Herring[1,7,8]

1. Mascoma Corporation, 67 Etna Rd, Lebanon, NH 03766

2. BioEnergy Science Center, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831

3. Current affiliation: Dow AgroSciences, 9330 Zionsville Road, Indianapolis, IN 46268

4. Virginia Commonwealth University, P.O. Box 843028, Chemical and Life Science Engineering, Richmond, Virginia 23284

5. Current affiliation: Centre for Applied Bioscience and Bioengineering, Department of Chemical Engineering and Applied Chemistry, University of Toronto

6. Current affiliation: Novogy Inc., Cambridge, MA 02138

7. Current affiliation: Thayer School of Engineering, Dartmouth College, 14 Engineering Drive, Hanover, NH 03755

8. To whom correspondence should be addressed

Email Addresses

Devin H. Currie - Devin.H.Currie.GR@dartmouth.edu

Babu Raman – raman.babu@gmail.com

Christopher M. Gowen - chris.gowen@utoronto.ca

Timothy J. Tschaplinski - tschaplinstj@ornl.gov

Miriam L. Land - landml@ornl.gov

Steven D. Brown - brownsd@ornl.gov

Sean F. Covalla - scovalla@mascoma.com

Dawn M. Klingeman - klingemandm@ornl.gov

Zamin K. Yang - yangz@ornl.gov

Nancy L. Engle - englenl@ornl.gov

Courtney M. Johnson - Courtney.Johnson10@fsis.usda.gov

Miguel Rodriguez - rodriguezmjr@ornl.gov

A. Joe Shaw - joeshawiv@gmail.com

William R. Kenealy - wkenealy@mascoma.com

Stephen S. Fong - ssfong@vcu.edu

Jonathan R. Mielenz - mielenzjr@ornl.gov

Brian H. Davison - davisonbh@ornl.gov

David A. Hogsett - dhogsett@opxbio.com

 Christopher D. Herring - chrisherringfish@gmail.com

Abstract

Background: *Thermoanaerobacterium saccharolyticum* is a hemicellulose-degrading thermophilic anaerobe that was previously engineered to produce ethanol at high yield. A major project was undertaken to develop this organism into an industrial biocatalyst, but the lack of genome information and resources were recognized early on as a key limitation.

Results: Here we present a set of genome-scale resources to enable the systems level investigation and development of this potentially important industrial organism. Resources include a complete genome sequence for strain JW/SL-YS485, a genome-scale reconstruction of metabolism, tiled microarray data showing transcription units, mRNA expression data from 71 different growth conditions or timepoints and GC/MS-based metabolite analysis data from 42 different conditions or timepoints. Growth conditions include hemicellulose hydrolysate, the inhibitors HMF, furfural, diamide, and ethanol, as well as high levels of cellulose, xylose, cellobiose or maltodextrin. The genome consists of a 2.7 Mbp chromosome and a 110 Kbp megaplasmid. An active prophage was also detected, and the expression levels of CRISPR genes were observed to increase in association with those of the phage. Hemicellulose hydrolysate elicited a response of carbohydrate transport and catabolism genes, as well as poorly characterized genes suggesting a redox challenge. In some conditions, a time series of combined transcription and metabolite measurements were made to allow careful study of microbial physiology under process conditions. As a demonstration of the potential utility of the metabolic reconstruction, the OptKnock algorithm was used to predict a set

of gene knockouts that maximize growth-coupled ethanol production. The predictions validated intuitive strain designs and matched previous experimental results.

Conclusion: These data will be a useful asset for efforts to develop *T. saccharolyticum* for efficient industrial production of biofuels. The resources presented herein may also be useful on a comparative basis for development of other lignocellulose degrading microbes, such as *Clostridium thermocellum*.

Background

Whether biomass-derived fuels play a major role in the world's energy future depends on the development of technology to produce them at a cost that is competitive with petroleum and other alternatives [1]. Fermentation of lignocellulose is an attractive approach to this task [2, 3] and development of better fermenting organisms could achieve much of the necessary cost reductions [4-6]. This represents an opportunity to apply recent advances in metabolic engineering and systems biology to a problem of major importance: the need for carbon-neutral fuels [7].

The thermophilic anaerobic bacteria include species with natural abilities to digest and ferment the polysaccharides that make up the bulk of lignocellulosic biomass [8, 9]. Unfortunately, the lack of information and resources for these organisms has hindered their development. *Thermoanaerobacterium saccharolyticum* is a Gram positive, low G+C bacterium in the phylogenetic class "*Clostridia*" [10]. Members of the genus *Thermoanaerobacterium* are thermophilic, rod shaped, chemoorganotrophic and able to

reduce thiosulfate to elemental sulfur. The species *T. saccharolyticum* can ferment a wide array of carbohydrates, such as starch, xylan, glucose, cellobiose, xylose, arabinose, mannose, and galactose, but cannot degrade crystalline cellulose [10]. Most sugars are fermented to ethanol, acetic acid, lactic acid, carbon dioxide and hydrogen [4]. *T. saccharolyticum* has a temperature range of 45-70°C, and pH range between 4.5-7.0. The formation of endospores has not been observed in this species as they have in the related genus *Clostridium*.

A variety of thermophilic enzymes of industrial utility have been isolated from *T. saccharolyticum*, including endoxylanase, beta-xylosidase, amylopullanase and glucuronidase [11-17]. A system for genetic manipulation of *T. saccharolyticum* was first described by Mai et al. [18], which has been improved by the discovery of natural competence [19], and the development of methods for making unmarked mutations with negatively selectable markers [20]. The genes for lactate dehydrogenase, phosphate transacetylase and acetate kinase were knocked out using these methods [4, 20]. The result was a strain that produces ethanol at greater than 90% of theoretical yield, comparable to other ethanologens such as yeast, *E. coli* or *Z. mobilis* [21, 22]. The advantages that *T. saccharolyticum* has over these other biocatalysts are its elevated growth temperature (matching the temperature optimum of many cellulases [21, 23]), and its ability to hydrolyze hemicellulose and co-ferment the major sugars present in lignocellulose [24].

Biomass is prepared for hydrolysis and fermentation by various forms of pretreatment in order to expose the cellulose fibers and reduce particle size, though inhibitory compounds, such as furfural and hydroxymethyl furfural (HMF) are generated in the process [25]. Cost effective ethanol production requires ethanol concentrations > 40 g/L, which necessitates that substrates, and by the same token their inhibitors, be present at fairly high concentrations. The ability to reduce costs by increasing levels of pretreated substrate is limited by the levels of inhibitors in the fermentation. While there is great potential to reduce costs by developing organisms with greater tolerance to inhibitors, little is known about the effects of these compounds on microbial physiology. One of the goals of this project was to generate information about the effects of specific inhibitors and complex inhibitor extracts from pretreated material. The project was undertaken as part of a larger project to develop *T. saccharolyticum* for fermentation of pretreated hardwood [26].

Another goal was to compare the genome of *T. saccharolyticum* to the genomes of other bacteria potentially important for biofuel production, including *Clostridium thermocellum*, an organism highly specialized for the hydrolysis of cellulose and the focus of other OMICs, and systems biology efforts. This work supplements the knowledge about both these important organisms and presents a comprehensive resource for further investigation.

Results and Discussion

Genome Sequencing

As the sequence was being generated, there were early indications that contig 2 was in fact a megaplasmid. Furthermore, early draft sequences showed that the ends of contig 2 overlapped. When PCR primers were designed at the ends of contig 2 facing outwards, they amplified a product consistent with a circular DNA molecule. The gene *Tsac_2822* on the putative megaplasmid encodes a RepB DNA replication protein with high similarity to replication proteins from bacterial megaplasmids: *C. botulinum* plasmid pCLI, *B. methanolicus* plasmid pBM19, and *B. weihenstephanensis* plasmid pBWB402. Contig 2 was poorly represented in the initial Sanger sequence data and was observed to be completely absent in strain ALK2; its loss as a complete unit further supporting its identification as an extra chromosomal unit [4].

The genome contains 39 ORFs predicted to have transposase function, with 12 of these concentrated in a 50 kbp region. The tool Prophage Finder [27] was used to identify two regions containing genes with similarity to known phage genes. These two regions are 36 kbp and 42 kbp, located between ORFs Tsac_2404 – Tsac_2458 in contig 1 and between ORFs Tsac_2829-Tsac_2885 in contig 3 (the later listed under a separate accession number in GenBank, CP003186). Close examination of individual reads of CP003186 showed that some proceeded from the phage into contig 1 near position 2,009,359, suggesting a phage integration site. Contig 1 reads showed that in some, but not all of them the phage was absent. In those reads where the phage was missing, the sequence at 2009359-2009371 was duplicated. Primers were designed to the chromosome flanking this region and in contig 3 facing outwards. All combinations of

primers amplified, supporting the conclusion that contig 3 is a phage that exists in both integrated and circular forms at this locus (Supplemental Figure S1).

The chromosome contains a region containing 39 CRISPR repeats along with 8 CRISPR-associated genes. The CRISPR spacers were aligned with BLAST against the genome and two of them were found to match the two putative phage regions. This suggests that this strain of *T. saccharolyticum* has a history of infection and defense against these two phage [28]. Analysis of *C. thermocellum* also showed possible prophages and much more numerous and extensive CRISPR repeats and CRISPR-associated genes, possibly related to the low transformation efficiency of *C. thermocellum* [29]. Additional analysis across other Clostridia show further CRISPR features [30].

A high percentage of genes (11.2%) have predicted functions (i.e. COG category) related to carbohydrate transport and metabolism. For comparison, only 6.5% of the ORFs in *C. thermocellum* ATCC27405 are assigned to this functional group. Both ABC-type and phospotransferase transporters occur. The tool dbCAN [31] was used to compare all *T. saccharolyticum* protein sequences to hidden Markov models (HMMs) of all protein families in the CAZY database. The program identified 73 ORFs with similarity to glycosyl hydrolase HMMs, including 3 in glycosyl hydrolase family 5 with a predicted function of “cellulase.” It also identified 18 proteins with similarity to Cellulose Binding Module HMMs. It should be noted though that *T. saccharolyticum* does not grow on crystalline cellulose such as avicel [10].

Surprisingly, a total of 67 sporulation-associated genes were identified, including *spo0A*, but the strain is sporulation deficient. As with *Thermoanaerobacterium thermosaccharolyticum* [32], *T. saccharolyticum* contains the nitrogenase genes required for nitrogen fixation. Sequenced members of the related genus *Thermoanaerobacter* apparently do not.

The genome contains 5 ribosomal regions, all oriented in the same direction. Remarkably, the ribosomal sequences are not uniform, but rather of two types showing only 95% identity in the “universal” region of the 16s subunit (Figure 1). Similar, but less pronounced heterogeneity of ribosomal sequences has been noted in other firmicutes [33], but has yet to be explained.

It is possible that these additional sequences confer some advantage during growth at elevated temperatures. Another possibility is that these modifications decrease sensitivity to an environmentally prevalent antibiotic that targets the 16s rRNA. The 16s rRNA is a common target for antibiotic compounds, for example aminoglycosides [34]. That said, resistance-conferring mutations are frequently single base pair changes rather than large insertion events [34, 35]. In addition, at least for the aminoglycosides, the reported site of action is the A site near the 3’ end of the 16s rRNA [35, 36], whereas these insertions are very close to the 5’ end. However, the version that contains the inserts causes the 5’ and 3’ ends to no longer be located near one another,

as can be seen in Figure 1 panels B and C, and thus may be playing a role in resistance.

Effects of hemicellulose extract

Spotted microarrays were used to examine the effect that biomass-derived hydrolysate and the associated inhibitors have on *T. saccharolyticum*. In an initial experiment, cells were grown in fermenters containing rich medium and a mixture of xylose and glucose to mid-logarithmic phase, whereupon the cells were “shocked” by the addition of 10% volume of hemicellulose extract (“washate”). Control fermentations were conducted by shocking the cells with a mix of xylose and acetate at the same concentration and pH. The cells continued to grow, though growth was slightly slowed. Samples were analyzed before and up to 1 hour after the shock using spotted microarrays. Each mRNA sample was measured relative to a genomic DNA control, and all $\log_2$ ratios given below are relative to the gDNA control [37].

When comparing the results from control reactors to those treated with washate, an increasing number of genes were upregulated over time in response to washate (spots above the diagonal in Figure 2). An alternate way of analyzing the same data is by comparing expression levels at a given time point to those previous to the shock (Supplemental Figure S2). Such comparisons versus the pre-shock culture showed more scatter, most likely due to growth phase-related gene expression changes.

Most of the genes affected by the washate were upregulated, with 58 having $\log_2$ ratios > 1 in at least one time point (Figure 3). At 5 min post-shock, a cluster of 17 genes (Tsac_1270-1286) was upregulated. This cluster includes glycosyl hydrolases and carbohydrate transport and catabolism genes, including three genes required for arabinose utilization. At 15 min post-shock, additional genes were upregulated, including those responsible for the formation of bacterial microcompartments and rhamnose utilization.

cDNA samples from before and 1 hour post washate shock were also hybridized to tiled Nimblegen microarrays. Compared to data from spotted arrays, the tiled array data showed less noise in the lower dynamic range. Moreover, by examining the expression levels visually, the boundaries of transcription units can be determined (Figure 4).

Effects of HMF and furfural

Two of the major inhibitors in washate are furfural and hydroxymethylfurfural (HMF). To further investigate the specific effect these components have on *T. saccharolyticum*, we performed additional “shock” experiments in which HMF and furfural were added during logarithmic growth, while observing the cellular response by microarray and metabolite analysis. The levels of HMF and furfural in pretreated hardwood hydrolysates is approximately 0.1 g/L. We tested additions of HMF and furfural from 0.1 to 1.0 g/L and found that 0.5 g/L of each showed a clear effect on growth (data not shown). Notably, the effect was greatly diminished in medium containing yeast extract, so a defined medium was used in this experiment. Sample processing methods for metabolite

analysis were validated as described in Supplemental Figure S3 and Supplemental Table S1. Actively growing fermenters of *T. saccharolyticum* strain M700 at an O.D. of 0.6 were shocked with 0.5 g/L HMF and furfural. Samples were taken before the shock and at 15 minutes, 1, 2, and 4 hours after shock. These samples were analyzed via GC/MS and spotted microarrays.

A total of 40 different metabolites were tracked over the time course of the experiment (Figure 5, Supplemental Figure S4, Supplemental Table S2). Almost all metabolites decreased after 15 minutes of exposure to HMF and furfural, with the exception of hydroxymethylfurfurol and citric acid. Hydroxymethylfurfurol, presumably resulting from the reduction of HMF, increased steadily throughout the 4 hours that metabolites were tracked. HMF and furfural were almost entirely metabolized after 16 hours. It is notable that glucose-6-phosphate is among the many metabolites that decrease as the result of HMF and furfural addition. This suggests that the inhibition occurs very early in the glycolysis pathway, either at glucose transport or its phosphorylation.

Microarray analysis of the same fermentations showed large expression differences in the phage loci between replicates during growth in the presence of HMF and furfural (Supplemental Figure S5). Other non-phage genes were also observed to change sympathetically with the phage genes, including the aforementioned CRISPRs.

In order to determine if some of the same genes were affected by the addition of HMF / furfural as by washate, a comparison of the two datasets was performed. The $\log_2$ ratio

difference was calculated and analyzed via *t*-tests using the control from the same time point as reference for washate shock and using the pre-shock as reference for HMF / furfural shock. The 15 minute and 60 minute time points were considered for each, and the greater $\log_2$ ratio or significance value was used. In the washate shock experiment 502 genes were significantly affected (P value < 0.05) in either the 15 or 60 minute time points, and in the HMF / furfural shock experiment, 414 genes were affected in either the 15 or 60 minute time points. Between the two sets of significant genes, 88 were in common (Supplemental Table S3). Of these, 40 had a $\log_2$ difference in either experiment greater than 0.7, and 9 had a log2 difference greater than 0.7 in both. Among these notable genes upregulated after both types of shock are members of a gene cluster related to sulfur assimilation (Tsac_1655-1665), alanine dehydrogenase (Tsac_2175) and NADPH-dependent methylglyoxal reductase (Tsac_1406). It should be noted, however, that this comparison is less than ideal in that different media and strains were used and that phage activity was noted in half of the HMF+furfural shock samples.

A wealth of other microarray and metabolite data were generated (Table 1). Note that at each timepoint listed in Table 1, multiple biological replicates were usually generated. Data are available as Supplemental Files 1-5.

Genome Scale *in silico* metabolic model

Genome-scale constraint-based metabolic models are useful tools for exploring the metabolic capabilities of an organism and for integrating bioinformatics data sets with

the metabolic network. A genome-scale model for *Thermoanaerobacterium saccharolyticum* was built for this study based on its genomic content, current literature, and experimental data (Supplemental File 6). An initial reaction list was built by comparing its genetic content with that of the related bacterium *Clostridium thermocellum*, for which a curated metabolic model already exists [38]. To do this, a BLAST search was performed using the genes included in the *C. thermocellum* model versus the *T. saccharolyticum* ORF predictions. This resulted in an initial set of 425 reactions with gene-reaction mappings to serve as the foundation of the reconstruction. Additionally, metabolic pathways for xylose and sorbitol metabolism were added, and cellulose breakdown reactions were removed. The bifurcating ferredoxin:NAD oxidoreductase described by Wang *et al.* was added as well [39]. A number of other changes were made based on biochemical evidence, and additional gap filling was performed as described in the methods section to generate a working model. These changes are detailed in Supplementary File 7. The final model contains 528 metabolites and 516 reactions associated with 315 genes. A comparison of these statistics to the *C. thermocellum* model is shown in Table 2.

Model Validation

Although the metabolic network composition at this stage was consistent with available information based on genome annotation and experimental observations, the resulting flux space remained highly underdetermined. This is a consistent challenge facing all constraint-based models, because many thermodynamic and regulatory effects cannot be captured in the stoichiometric network. In particular for the *T.saccharolyticum* model,

the diversity of hydrogenase systems hosted by this organism, left unconstrained, provide the network with many ways to efficiently regenerate cofactors, allowing unrealistic levels of flux towards acetate and hydrogen. From a thermodynamic standpoint, actual allowable fluxes through these reactions are limited by many factors, including the intracellular concentrations of the cofactors, the concentration of hydrogen, intra- and extracellular pH, and the reduction potential of ferredoxin. This problem is complicated further by the kinetics and expression levels of the responsible enzymes. In the absence of the necessary parameters to formulate these constraints, we decided on a top-down approach to replicate experimental observations by making some of the hydrogenase reactions irreversible and by limiting the overall hydrogen production to observed yields. In a previous study [40], the four gene operon *hfs* coding for the reaction ferredoxin hydrogenase was found to be the primary hydrogen producer *in vivo*, whereas the other hydrogenase genes tested were found to contribute only slightly or not at all to hydrogen production. Reflecting this in the model, the energy-conserving hydrogenase (ECH) was blocked, and the bifurcating hydrogenase (BIFH2) and the NADH hydrogenase (NADH2) were forced to be irreversible in the direction of hydrogen uptake. Additionally, total hydrogen export was limited to a yield of 0.9 M H_2:M glucose to reflect the *in vivo* measurements [40]. These modifications had a dramatic impact on the predicted performance of the model by limiting the amount of reducing equivalents that could be sent to hydrogen production, thereby shifting some carbon flux from acetate to ethanol and other organic acids. Further experimentation with hydrogenase constraints may prove useful to help understand how electron and carbon flow are related in this and other mixed-acid fermentors.

Previous metabolic engineering efforts on *T. saccharolyticum* [4, 40] have explored two distinct strategies for improving ethanol yield: a carbon-centric approach that focuses on eliminating competing carbon fluxes at the pyruvate branch point, and an electron-centric approach that disrupts the cell's ability to produce hydrogen as a highly-reduced electron acceptor. Each of these strategies was shown to improve ethanol production to varying degrees. A phenotypic phase plane analysis was performed to explore the metabolic implications of these knockout strategies. Figure 6 shows the optimal growth surfaces for these knockouts over the complete ranges of possible carbon uptake and ethanol production rates. In the wild-type seen in Figure 6-A, optimal growth can occur across a wide range of ethanol flux values, limited by the maximum glucose uptake rate. Knocking out the lactate dehydrogenase (LDH) and phosphotransacetylase (PTA) reactions eliminates stoichiometrically equivalent solutions, leading to a maximum growth rate that is coupled to high ethanol production (Figure 6-B). The coupling of ethanol flux to growth rate was found to be much stronger, however, in the electron-centered strategy (Figure 6-C), which removed the reactions for LDH and ferredoxin hydrogenase (HFS). This knockout strategy greatly limits the available solution space and tightly dictates the ethanol yield at some penalty to the maximum growth rate. This finding is consistent with experimental results, which found a lower overall growth rate and cellobiose uptake rate in the *ldh-hfs* deletion strain when compared to the wild-type or *ldh-pta-ack* deletion strain [40]. However, the strong coupling of ethanol production to growth rate in the *ldh-hfs* knockout strain implies that it may be a good candidate for adaptive evolution to improve ethanol productivity.

We attempted to determine if any other knockout strain designs would maximize ethanol production at an optimal growth rate. The bilevel optimization algorithm OptKnock [41] was used to search for knockout strain designs that would improve production of ethanol by coupling it to improved growth rate. When OptKnock searches a maximum of 2 reaction knockouts, optimal ethanol production is predicted when knocking out LDH and HFS. When allowing three reactions knockouts, OptKnock finds a marginal improvement by deleting LDH, HFS, and glutamate dehydrogenase (GLUD). Removal of GLUD forces the cell to use the reactions glutamate synthase (GLUS) and glutamine synthetase (GLNS) in order to incorporate ammonium, spending an additional mole of ATP per mole of ammonium (Table 3). This inefficiency predicts only a marginal improvement in ethanol production of 0.3% over the ΔLDH-ΔHFS strain (Figure 7).

Conclusions

Here we report the first genome scale study of the industrially important bacterium *T. saccharolyticum*. This work informs and supports not only the study of fundamental microbial physiology, but also its potential applications in this organism. The resources presented herein will facilitate further efforts to engineer *T. saccharolyticum* for the production of biofuels. In addition, ongoing engineering efforts in other organisms to increase inhibitor tolerance and ethanol yield and titer may benefit from these data.

Materials and Methods

Strains

The tiled microarrays were designed based on wild-type *T. saccharolyticum* YS485 DSM 8691 (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Germany). The microarrays and metabolite profiling were performed using engineered and evolved ethanologenic strains of *T. saccharolyticum*, described previously [26].

Growth Media

MTC media [42] was supplemented with various concentrations of glucose, xylose, and mixtures of hemicellulose extract or acetic acid mixed with xylose. These concentrations are noted for each experiment. TS5 media was developed specifically for *T. saccharolyticum*. It is similar to the previously published media TSC1 [20] but with only 0.5 g/l KH_2PO_4 and with 0.5 g/l tryptone. The full media formulation per liter is: Solution I (yeast extract 8.5 g, 1.85 $(NH_4)_2SO_4$, 0.05 g $FeSO_4$, 0.5 g KH_2PO_4, 1 g $MgCl_2$ * 6 H2O, 0.05 g $CaCl_2$, 0.5 g Tryptone, 2 g Trisodium citrate * 2 H_2O, 800 ml H_2O) and Solution II (10 g Xylose, 200 ml H_2O). These are autoclaved separately to avoid burning the xylose, then mixed.

Hemicellulose extract

Hemicellulose extract, or 'washate', for the microarray and metabolite profiling experiments was prepared by suspending mixed hardwood pretreated with steam in an Andritz horiontal plug-flow reactor to severity 3.8 in water at 30% solids. It was then autoclaved for one hour, and the liquid was removed from the solids by vacuum filtration using Whatman Grade No. 1 Filter Paper (Whatman Ltd, Kent, UK). It was then brought to pH 6.0 using NH_4OH. The extract contained 11.52 g/L monomeric xylose, 0.89 g/L

glucose, 0.84 g/L lactate, 3.54 g/L acetate, 0.56 g/L HMF, 0.26 g/L furfural, and various other inhibitors. Additional carbohydrate was present but not analyzed due to its oligomeric state or lack of standards for analysis.

Genome Sequencing of T. saccharolyticum YS485

The genome of *T. saccharolyticum* JW/SL-YS485 was generated over a 4 year span by a variety of techniques. Initially, a clone library was constructed and Sanger sequenced to 8x coverage. Clones were selected for additional sequencing to close gaps, and additional sequence data was generated with the 454 platform. The assembled draft was then aligned to the complete genomes of *T. tengcongensis* and *T. pseudoethanolicus*, allowing the contigs to be ordered and oriented to each other. PCR primers were designed at the ends of contigs and used to amplify across gaps, consisting mostly of ribosomal regions. These PCR products were Sanger sequenced and used to manually close the genome. Finally, differences between the Sanger and 454 data were resolved by examining sequence data from various strains sequenced with Illumina technology. Open reading frames were predicted and functions annotated using an annotation pipeline at Oak Ridge National Laboratory. The genome sequence has been assigned GenBank accession numbers CP003184.1, CP003185.1 and CP003186.1 (the genome, the mega-pasmid, and the phage, respectively).

Phage Confirmation

Primers were designed to confirm the presence of a phage in contig 3 which is present in both integrated and circular forms (C17_near_endF: CTGCCCGTGGAACATCTAAT,

C17_near_startR: GTTGGTTCTGCCCTGTTTGT, C15_int_siteF: TTTGCACCGCCATTTAAGAG, C15_int_siteR: ACGGTGATGAAGAAGCGAAA, C18_near_startR: AATTCGGCATGTGTTGGAT). PCR was performed on genomic DNA from *T. saccharolyticum* strain YS485 and *T. saccharolyticum* M700 and products were separated on a 1% agarose gel.

Spotted Microarray Construction, Hybridization and Analysis

Spotted oligonucleotide microarrays were essentially constructed and hybridized as described previously [43, 44]. Briefly, DNA sequences that represented predicted-protein encoding sequences were obtained for the *T. saccharolyticum* YS485 genome (NCBI GenBank accession numbers CP003184.1, CP003185.1 and CP003186.1) and 70-mer oligonucleotide probes were designed using the CommOligo software [45]. The original genome sequence was in draft format and 2,627 oligonucleotide probes were designed for 2,667 putative CDS, representing 98.5% of the predicted-protein encoding sequences for the draft genome. Subsequently, refinements were made as the genome sequence was closed. Oligonucleotides commercially synthesized without modification (Integrated DNA Technologies, Coralville, Iowa) in 96-well stock plates and transferred to 384-well printing plates in a final concentration of 50% DMSO using a BioMek FX liquid handling robot (Beckman-Coulter, Fullerton, CA). Probes were then spotted onto UltraGAPS glass slides (Corning Life Sciences, Corning, NY) using a BioRobotics Microgrid II microarrayer (Genomic Solutions, Ann Arbor, MI) in a dust-free clean room maintained at 21ºC and 50% relative humidity. Spotted DNA was stabilized on slides by

ultraviolet cross-linking using a UV 1800 Stratalinker (Stratagene, La Jolla, CA) according to slide manufacturer's instructions (Corning Life Sciences).

Total RNA was purified using an RNeasy Plus Mini Kit (Qiagen), which was used as template to generate cDNA copies labeled with Cy5-dUTP (Amersham Biosciences, Piscataway, NJ). Labeled genomic DNA (Cy3) was used as a control and as the common reference to co-hybridize with labeled RNA (Cy5) samples for each slide and microarray hybridization and washing conditions have been described elsewhere [43, 44, 46]. Microarray images were scanned using a ScanArray Express (PerkinElmer) scanner, and spot signal, quality, and background fluorescent intensities were quantified using ImaGene version 6.0 (Biodiscovery, Marina Del Rey, CA). Outlier detection, background correction, normalizations and log ratios were generated as described previously [46], except that the workflow was conducted using JMP Genomics (SAS Institute Inc.) with custom scripts.

Washate Shock Microarrays

T. saccharolyticum ALK2 was grown overnight shaking in bottles with 50 ml MTC + 5 g/L glucose and xylose to an optical density of 0.5 at 600 nm. 25 ml from these bottles were inoculated into 1 L of MTC media + 2.7 g/L xylose + 4.6 g/L glucose. Fermentations were performed in duplicate at 1 L volume in Sartorius BiostatA+ reactors maintained at pH 5.8, 55°C, stirring 150 rpm, and purged with N2/CO2 prior to inoculation. Upon reaching an OD of 0.6, 100 ml of either a control solution (11.5 g/L xylose, 3.5 g/L acetic acid, with pH adjusted to 6.0 with NH4OH) or washate was added.

Samples were taken at time 0, 5, 15, and 60 minutes after shock. The samples were mixed with 30 ml RNAprotect bacteria reagent (QIAGEN Corp, Valencia CA) and left at room temperature for 5 minutes. The samples were then centrifuged at 4000 rpm for 10 min at 4°C. The pellets were then resuspended in 1 ml SET buffer and stored at -80°C.

HMF and Furfural Shock Microarrays

T. saccharolyticum M700 was grown overnight in bottles with 50 ml Defined TS5 media shaking at 55°C. 25 ml from these bottles were inoculated into 4 reactors containing 1 L of Defined TS5 and maintained at pH 5.8, 55°C, stirring 150 rpm, and under constant purging with N2/CO2. The reactors were grown to an O.D. of 0.06 at which point 0.5 g/l each of HMF and furfural were added to two of the reactors, leaving the second two as controls. Samples were taken at times 0, 15 minutes, 1, 2, and 4 hours after addition. Two sets of samples were taken at each time point, one for microarrays and one for proteomics. The samples for microarray analysis were mixed with 30 ml RNAprotect bacteria reagent (QIAGEN Corp, Valencia CA) and left at room temperature for 5 minutes. The samples were then centrifuged at 4000 rpm for 10 min at 4°C. The pellets were then resuspended in 1 ml SET buffer and stored at -80°C. The samples for the metabolite profiling assays were centrifuged at 4°C at 4000 rpm for 10 minutes, supernatants were poured off and the pellets were frozen at -80°C

Tiled Microarrays

Tiled microarrays were performed by Nimblegen Corporation (Madison, WI).

Metabolite profiling

Metabolites from *T. saccharolyticum* culture pellets and hydrolysates were analyzed as trimethylsilyl (TMS) derivatives by gas chromatography-mass spectrometry (GC/MS) using electron impact (EI) ionization, as described previously (Yang et al. 2009). Briefly, aliquots of culture supernatants (50 µL to 2 mL) and sorbitol (aqueous internal standard added to yield 10 – 60 ng per µL injected) were transferred by pipette to a vial and stored at -20^{o}C until analyzed. Microbial pellets were fast-frozen in liquid nitrogen and stored at -80 ºC until analyzed. Frozen pellets were weighed and added to 10 mL 80% ethanol containing sorbitol as internal standard. Cell pellets were ruptured by sonication with temperature maintained below 0 °C, and cell debris separated from the extract by centrifugation at 4°C, and 2 mL were dried in a stream of N_2 prior to silylation. The hydrolysate samples were thawed and also concentrated to dryness under a stream of N_2. The internal standard was added to correct for subsequent differences in derivatization efficiency and changes in sample volume during heating. Dried extracts were dissolved in 500 µL of silylation–grade acetonitrile followed by the addition of 500 µL N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) (Thermo Scientific, Bellefonte, PA), and samples then heated for 1 h at 70 °C to generate TMS derivatives. After 2-3 days, 1-µL aliquots were injected into an Agilent Technologies Inc. (Santa Clara, CA) 5975C inert XL gas chromatograph-mass spectrometer, fitted with an Rtx-5MS with Integra-guard (5% diphenyl/95% dimethyl polysiloxane) 30 m x 250 µm x 0.25 µm film thickness capillary column. The standard quadrupole GC/MS was operated in the EI (70 eV) ionization mode, with 6 full-spectrum (50-650 Da) scans per second. Gas (helium) flow was 1.3 mL per minute with the

injection port configured in the splitless mode. The injection port, MS Source, and MS Quad temperatures were 250 ºC, 230 ºC, and 150 ºC, respectively. The initial oven temperature was held at 50 ºC for 2 min and was programmed to increase at 20 ºC per min to 325 ºC and held for another 11 min, before cycling back to the initial conditions. A large user-created database (>1800 spectra) of mass spectral electron ionization (EI) fragmentation patterns of TMS-derivatized compounds, as well as the Wiley Registry 8th Edition combined with NIST 05 mass spectral database, were used to identify the metabolites of interest to be quantified. Peaks were reintegrated and reanalyzed using a key selected ion, characteristic m/z fragment, rather than the total ion chromatogram, to minimize integrating co-eluting metabolites. The extracted peaks of known metabolites were scaled back up to the total ion current using predetermined scaling factors. The scaling factor for the internal standard was used for unidentified metabolites. Peaks were quantified by area integration and the concentrations were normalized to the quantity of the internal standard recovered, volume of sample processed, derivatized, and injected. Three to six replicate samples were analyzed per time point, and the metabolite data were averaged at a given time point. Unidentified metabolites were denoted by their retention time as well as key mass-to-charge (m/z) ratios.

Constraint-based modeling of Thermoanaerobacterium saccharolyticum

Initial construction of the *Thermoanaerobacterium saccharolyticum* reaction list was based on the previously published model of the closely related species *Clostridium thermocellum* [38, 47]. This was accomplished by using BLAST to search for genes in *T. saccharolyticum* that were homologous to the genes represented in the *C.*

thermocellum model. Further refinement to the model was done by manual curation, incorporating available biochemical and genetic information. The resulting reaction list was not yet able to produce flux through the biomass reaction using appropriate exchange boundary conditions, so additional gap filling was required. This was accomplished through the use of a novel gap filling algorithm called FBA-gap [48] which proposes a minimal set of reaction additions necessary to support biomass production. These reactions are sourced from a reaction database populated using the reaction lists of available stoichiometric models.

Flux balance analysis (FBA) [49], was used throughout the reconstruction and analysis of the *T. saccharolyticum* model to simulate optimal growth. Modeling work was done using the COBRA toolbox for Matlab [50, 51] along with custom methods and the Gurobi Optimizer. OptKnock [41] was used to search for knockout strains that would putatively couple ethanol production with an improved growth rate. An implementation of OptKnock is available in the COBRA toolbox for MATLAB.

Statistical analysis

All statistical analyses were performed using R statistical software [52] and the package gplots [53].

Availability of supporting data

The data sets supporting the results of this article are included within the article (and its additional files).

List of abbreviations

HMF- hydroxymethyl furfural

ORF - Open Reading Frame

CRISPR - Clustered Regularly Interspaced Short Palindromic Repeats

HMM - Hidden Markov Model

ECH - Energy-Conserving Hydrogenase

BIFH2 - Bifurcating Hydrogenase

NADH2 - NADH Hydrogenase

LDH - Lactate Dehydrogenase

PTA - Phosphotransacetylase

HFS - Ferredoxin Hydrogenase

GLUD - Glutamate Dehydrogenase

GLUS - Reactions Glutamate Synthase

GC/MS - Gas Chromatography-Mass Spectrometry

MSTFA - N-methyl-N-trimethylsilyltrifluoroacetamide

BLAST - Basic Local Alignment Search Tool

FBA - Flux Balance Analysis

Competing Interests

DHC was partially supported by funding from Mascoma Corporation during his doctoral work at Dartmouth College. SFC, AJS and CDH were salaried employees of Mascoma Corporation and are listed as inventors on several patent filings related to

Thermoanaerobacterium saccharolyticum by Mascoma Corporation. WRK was a salaried employee of Mascoma Corporation. DAH is a stockholder of Mascoma Corporation and was a salaried employee of Mascoma Corp. and Dartmouth College. He is listed as an inventor on several patent filings related to *T. saccharolyticum* by Mascoma Corporation and Dartmouth College, each of which owns or has applied for patents related to *T. sacchraolyticum*. JRM DAH is a stockholder of Mascoma Corporation. All other authors have no competing interests.

Author Contributions

DH Currie analyzed data and drafted the manuscript, B Raman performed microarray work and was the main point of contact at ONRL, CM Gowen performed the metabolic reconstruction and drafted the manuscript, TJ Tschaplinski performed metabolite profiling analyses, ML Land performed genome analysis and annotation, SD Brown designed the microarrays and assisted with analysis, SF Covalla performed fermentations and processed samples, DM Klingeman processed microarray samples, ZK Yang processed microarray samples, NL Engle processed metabolite sample extracts, CM Johnson processed microarray samples, M Rodriguez assisted in performing fermentations, AJ Shaw assisted in metabolic analysis, WR Kenealy supervised fermentation work, SS Fong supervised metabolic reconstruction and analysis, JR Mielenz coordinated work at ORNL and helped design experiments, BH Davison coordinated work at ORNL and helped design experiments, DA Hogsett supervised the project, CD Herring directed the project and drafted the manuscript.

Acknowledgements

ORNL annotated the genome sequence with funding from Mascoma Corp. The U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy Office in the BioEnergy Technologies Office provided support for the experimental work and analysis under award # GO17057. Manuscript preparation was supported by the BioEnergy Science Center. The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the DOE under Contract DE-AC05-00OR22725.

References

1. Sims REH, Mabee W, Saddler JN, Taylor M: **An overview of second generation biofuel technologies**. *Bioresource Technol* 2010, **101**(6):1570-1580.
2. Olson DG, McBride JE, Joe Shaw A, Lynd LR: **Recent progress in consolidated bioprocessing**. *Curr Opin Biotechnol* 2012, **23**(3):396-405.
3. Lynd LR, van Zyl WH, McBride JE, Laser M: **Consolidated bioprocessing of cellulosic biomass: an update**. *Curr Opin Biotechnol* 2005, **16**(5):577-583.
4. Shaw AJ, Podkaminer KK, Desai SG, Bardsley JS, Rogers SR, Thorne PG, Hogsett DA, Lynd LR: **Metabolic engineering of a thermophilic bacterium to**

produce ethanol at high yield. *Proc Natl Acad Sci U S A* 2008, **105**(37):13769-13774.

5. Shaw JA, Covalla SF, Miller BB, Firliet BT, Hogsett DA, Herring CD: **Urease expression in a *Thermoanaerobacterium saccharolyticum* ethanologen allows high titer ethanol production**. *Metab Eng* 2012, **14**:528–532.

6. Parawira W, Tekere M: **Biotechnological strategies to overcome inhibitors in lignocellulose hydrolysates for ethanol production: review**. *Crit Rev Biotechnol* 2011, **31**(1):20-31.

7. Ragauskas AJ, Williams CK, Davison BH, Britovsek G, Cairney J, Eckert CA, Frederick WJ, Hallett JP, Leak DJ, Liotta CL *et al*: **The path forward for biofuels and biomaterials**. *Science* 2006, **311**(5760):484-489.

8. Lee J: **Biological conversion of lignocellulosic biomass to ethanol**. *J Biotechnol* 1997, **56**(1):1-24.

9. Chang T, Yao S: **Thermophilic, lignocellulolytic bacteria for ethanol production: current state and perspectives**. *Appl Microbiol Biotechnol* 2011, **92**(1):13-27.

10. Lee YE, Jain MK, Lee CY, Lowe SE, Zeikus JG: **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* 1993, **43**(1):41-51.

11. Podkaminer KK, Guss AM, Trajano HL, Hogsett DA, Lynd LR: **Characterization of xylan utilization and discovery of a new endoxylanase in Thermoanaerobacterium saccharolyticum through targeted gene deletions**. *Appl Environ Microbiol* 2012.

12. Vocadlo DJ, Wicki J, Rupitz K, Withers SG: **Mechanism of Thermoanaerobacterium saccharolyticum beta-xylosidase: kinetic studies**. *Biochemistry* 2002, **41**(31):9727-9735.

13. Bronnenmeier K, Meissner H, Stocker S, Staudenbauer WL: **alpha-D-glucuronidases from the xylanolytic thermophiles Clostridium stercorarium and Thermoanaerobacterium saccharolyticum**. *Microbiology* 1995, **141 (Pt 9)**:2033-2040.

14. Ramesh MV, Podkovyrov SM, Lowe SE, Zeikus JG: **Cloning and sequencing of the Thermoanaerobacterium saccharolyticum B6A-RI apu gene and purification and characterization of the amylopullulanase from Escherichia coli**. *Appl Environ Microbiol* 1994, **60**(1):94-101.

15. Lee YE, Lowe SE, Zeikus JG: **Gene cloning, sequencing, and biochemical characterization of endoxylanase from Thermoanaerobacterium saccharolyticum B6A-RI**. *Appl Environ Microbiol* 1993, **59**(9):3134-3137.

16. Lee YE, Zeikus JG: **Genetic organization, sequence and biochemical characterization of recombinant beta-xylosidase from Thermoanaerobacterium saccharolyticum strain B6A-RI**. *J Gen Microbiol* 1993, **139 Pt 6**:1235-1243.

17. Lee YE, Lowe SE, Zeikus JG: **Regulation and Characterization of Xylanolytic Enzymes of Thermoanaerobacterium-Saccharolyticum B6a-Ri**. *Appl Environ Microbiol* 1993, **59**(3):763-771.

18. Mai V, Wiegel J: **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* 2000, **66**(11):4817-4821.

19. Shaw AJ, Hogsett DA, Lynd LR: **Natural competence in *Thermoanaerobacter* and *Thermoanaerobacterium* species**. *Appl Environ Microbiol* 2010, **76**(14):4713-4719.

20. Shaw AJ, Covalla SF, Hogsett DA, Herring CD: **Marker removal system for *Thermoanaerobacterium saccharolyticum* and development of a markerless ethanologen**. *Appl Environ Microbiol* 2011, **77**(7):2534-2536.

21. Dien BS, Cotta MA, Jeffries TW: **Bacteria engineered for fuel ethanol production: current status**. *Appl Microbiol Biotechnol* 2003, **63**(3):258-266.

22. Jarboe LR, Grabar TB, Yomano LP, Shanmugan KT, Ingram LO: **Development of ethanologenic bacteria**. *Adv Biochem Eng Biotechnol* 2007, **108**:237-261.

23. Olofsson K, Bertilsson M, Liden G: **A short review on SSF - an interesting process option for ethanol production from lignocellulosic feedstocks**. *Biotechnol Biofuels* 2008, **1**(1):7.

24. Tsakraklides V, Shaw AJ, Miller BB, Hogsett DA, Herring CD: **Carbon catabolite repression in Thermoanaerobacterium saccharolyticum**. *Biotechnol Biofuels* 2012, **5**(1):85.

25. Liu S, Lu H, Hu R, Shupe A, Lin L, Liang B: **A sustainable woody biomass biorefinery**. *Biotechnol Adv* 2012, **30**(4):785-810.

26. **Final Report on Development of Thermoanaerobacterium saccharolyticum for the conversion of lignocellulose to ethanol.** [http://www.osti.gov/servlets/purl/1033560/]

27. Bose M, Barber RD: **Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences**. *In Silico Biol* 2006, **6**(3):223-227.

28. Bhaya D, Davison M, Barrangou R: **CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation**. *Annu Rev Genet* 2011, **45**:273-297.

29. Olson DG, Lynd LR: **Transformation of Clostridium thermocellum by electroporation**. *Methods Enzymol* 2012, **510**:317-330.

30. Brown SD, Nagaraju S, Utturkar S, De Tissera S, Segovia S, Mitchell W, Land ML, Dassanayake A, Koepke M: **Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of Clostridium autoethanogenum and analysis of CRISPR systems in industrial relevant Clostridia**. *Biotechnology for Biofuels* 2014, **7**.

31. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y: **dbCAN: a web resource for automated carbohydrate-active enzyme annotation**. *Nucleic Acids Res* 2012, **40**(Web Server issue):W445-451.

32. Bogdahn M, Kleiner D: **N2 fixation and NH4+ assimilation in the thermophilic anaerobes Clostridium thermosaccharolyticum and Clostridium thermoautotrophicum**. *Arch Microbiol* 1986, **144**(1):102-104.

33. Shimizu T, Ohshima S, Ohtani K, Hoshino K, Honjo K, Hayashi H: **Sequence heterogeneity of the ten rRNA operons in Clostridium perfringens**. *Syst Appl Microbiol* 2001, **24**(2):149-156.

34. Beauclerk AA, Cundliffe E: **Sites of action of two ribosomal RNA methylases responsible for resistance to aminoglycosides**. *J Mol Biol* 1987, **193**(4):661-671.

35. Recht MI, Fourmy D, Blanchard SC, Dahlquist KD, Puglisi JD: **RNA sequence determinants for aminoglycoside binding to an A-site rRNA model oligonucleotide**. *J Mol Biol* 1996, **262**(4):421-436.

36. Kotra LP, Haddad J, Mobashery S: **Aminoglycosides: perspectives on mechanisms of action and resistance and strategies to counter resistance**. *Antimicrob Agents Chemother* 2000, **44**(12):3249-3256.

37. Yang Y, Zhu M, Wu L, Zhou J: **Assessment of data processing to improve reliability of microarray experiments using genomic DNA reference**. *Bmc Genomics* 2008, **9 Suppl 2**:S5.

38. Roberts SB, Gowen CM, Brooks JP, Fong SS: **Genome-scale metabolic analysis of Clostridium thermocellum for bioethanol production**. *BMC Syst Biol* 2010, **4**:31.

39. Wang S, Huang H, Moll J, Thauer RK: **NADP+ reduction with reduced ferredoxin and NADP+ reduction with NADH are coupled via an electron-bifurcating enzyme complex in Clostridium kluyveri**. *J Bacteriol* 2010, **192**(19):5115-5123.

40. Shaw AJ, Hogsett DA, Lynd LR: **Identification of the [FeFe]-hydrogenase responsible for hydrogen generation in *Thermoanaerobacterium saccharolyticum* and demonstration of increased ethanol yield via hydrogenase knockout**. *J Bacteriol* 2009, **191**(20):6457-6464.

41. Burgard AP, Pharkya P, Maranas CD: **Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization**. *Biotechnol Bioeng* 2003, **84**(6):647-657.

42. Hogsett DA: **Cellulose hydrolysis and fermentation by *Clostridium thermocellum* for the production of ethanol**. Hanover: Dartmouth College; 1995.

43. Brown SD, Raman B, McKeown CK, Kale SP, He ZL, Mielenz JR: **Construction and evaluation of a *Clostridium thermocellum* ATCC 27405 whole-genome oligonucleotide microarray**. *Appl Biochem Biotechnol* 2007, **137**:663-674.

44. Chhabra SR, He Q, Huang KH, Gaucher SP, Alm EJ, He Z, Hadi MZ, Hazen TC, Wall JD, Zhou J *et al*: **Global analysis of heat shock response in *Desulfovibrio vulgaris* Hildenborough**. *J Bacteriol* 2006, **188**:1817–1828.

45. Li X, He Z, Zhou J: **Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation**. *Nucleic Acids Res* 2005, **33**:6114-6123.

46. Mukhopadhyay A, He Z, Alm EJ, Arkin AP, Baidoo EE, Borglin SC, Chen W, Hazen TC, He Q, Holman H-Y *et al*: **Salt Stress in Desulfovibrio vulgaris Hildenborough: an Integrated Genomics Approach**. *Journal of Bacteriology* 2006, **188**(11):4068-4078.

47. Gowen CM, Fong SS: **Genome-scale metabolic model integrated with RNAseq data to identify metabolic states of Clostridium thermocellum**. *Biotechnol J* 2010, **5**(7):759-767.

48. Brooks JP, Burns WP, Fong SS, Gowen CM, Roberts SB: **Gap detection for genome-scale constraint-based models**. *Adv Bioinformatics* 2012, **2012**:323472.

49. Edwards JS, Ramakrishna R, Palsson BO: **Characterizing the metabolic phenotype: a phenotype phase plane analysis**. *Biotechnol Bioeng* 2002, **77**(1):27-36.

50. Schellenberger J, Que R, Fleming RM, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S *et al*: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0**. *Nat Protoc* 2011, **6**(9):1290-1307.

51. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox**. *Nat Protoc* 2007, **2**(3):727-738.

52. R Core Team: **R: A language and environment for statistical computing**. In. Vienna, Austria; 2012.

53. Gregory R. Warnes BB, Lodewijk Bonebakker, Robert Gentleman,, Wolfgang Huber Andy Liaw TL, Martin Maechler, Arni, Magnusson SM, Marc Schwartz and Bill Venables: **gplots: Various R programming tools for plotting data**. In., 2.11.0 edn; 2012.

54. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction**. *Nucleic Acids Res* 2003, **31**(13):3406-3415.

Figure 1. A comparison between the two versions of the 16s mRNA found in *T. saccharolyticum.* A) an alignment and consensus sequence for a heterogeneous segment of the five 16S ribosomal components found in *T. saccharolyticum.* B) Mfold prediction of the structure of the shorter 16S mRNA. [54]. C) Mfold prediction of the structure of the longer 16S mRNA.

Figure 2. Time points between 5 and 60 minutes post-shock with hemicellulose extract. The horizontal axis represents $\log_2$ of the control xylose+acetate expression level (mRNA:gDNA ratio), while the vertical axis represents the hemicellulose extract-treated expression level. All data are the average of duplicate experiments with the exception of the 5 minutes post hemicellulose extract shock which is in triplicate.

Figure 3. Heat map of hierarchical clustering of genes that change in expression level upon the addition of washate with a P value of <0.01 and with a $\log_2$ ratio >1.0 in at least one time point. The range of log2 mRNA:gDNA ratios is given in the color key.

Figure 4. Example of data from Nimbegen tiled microarrays (bottom) showing transcription units correlated to open reading frames (top).

Figure 5. Inhibitor shock. A) Plot showing the addition of HMF and furfural in culture supernatants and the temporary disruption of growth. B) Plot showing the levels of intracellular citric acid and hydroxymethylfurfurol, as well as the average of all other

metabolites. C) A heat map of a hierarchical clustering of the concentration of all monitored intracellular metabolites over the course of the 4 hour experiment.

Figure 6. Phenotypic phase planes for *T. saccharolyticum* high-ethanol knock out strains. The maximum growth rate is shown as a surface over a range of fluxes for glucose uptake and ethanol production. The wild-type surface (A) shows the maximum growth rate occurring equally across a wide range of ethanol production rates, while the phase planes for the Δ*ldh-pta* strain (B) and the *Δldh-hfs* strain (C) demonstrate that the potential solution space is trimmed in a way that couples maximum growth to high ethanol yield.

Figure 7. Growth envelope for various ethanol strain designs during growth on glucose. ΔLDH-ΔHFS and ΔHFS-ΔLDH-ΔGLUD were both identified by OptKnock as being optimal designs for ethanol production.

Table 1. Summary of microarray and metabolomics data sets.

Table 2. A comparison between the number of components in the models generated for *T. saccharolyticum* and *C. thermocellum*.

Table 3. Relevant reactions in ethanol producing knockout strain designs.

844

845 **Table 1**.

Strain	Medium	Condition	Microarray timepoints analyzed	Metabolite timepoints analyzed
ALK2	MTC	arabinose	1	
ALK2	MTC	cellobiose	2	5
ALK2	MTC	cellobiose: nitrogen-limited	4	4
ALK2	MTC	cellobiose fed-batch: N-limited	4	5
ALK2	MTC	enzymatic hydrolysate	1	
ALK2	MTC	glucose-arabinose-mannose	4	
ALK2	MTC	glucose-xylose	7	
ALK2	MTC	glucose-xylose-cellobiose	1	
ALK2	MTC	glucose-xylose-ethanol	1	
ALK2	MTC	glucose-xylose-acetate shock	3	
ALK2	MTC	glucose-xylose-washate shock	3	
ALK2	MTC	pretreated hardwood SSF	3	
ALK2	MTC	xylose fed-batch	4	5
M0355	MTC	glucose-xylose	1	

M0355	MTC	glucose-xylose-ethanol	1	
M0355	MTC	pretreated hardwood SSF	1	
M0521	MTC	pretreated hardwood SSF	2	
M0700	MTC	glucose-xylose	2	
M0700	MTC	glucose-xylose-ethanol	2	
M0700	TSD	glucose-xylose HMF+furfural shock	5	5
M1151	TSC3	cellobiose-maltodextrin	4	4
M1151	TSC4	xylose-detoxified washate shock	7	3
M1291/1442	TSC4	sigmacell SSF	2	6
M1732	TSC7	xylose-diamide shock	6	5

846

847

848

849

850 **Table 2.**

	T. saccharolyticum	*C. thermocellum*[a]
Genes	315	432
Metabolites	503	525
Reactions	515	577
- Gene-associated	461 (90%)	463 (80%)
- Biomass	1	1
- Non-gene associated [cytosolic]	43	60
- Non-gene associated [transport]	11	54

851 a[38]

852

853

Table 3.

ID	Reaction Name	Formula	Gene Association
GLNS	Glutamine synthetase	glu-L[c] + ATP[c] + NH4[c] -> ADP[c] + Pi[c] + H[c] + gln-L[c]	Tsac_2029
GLUDy	Glutamate dehydrogenase (NADP)	NADP[c] + H2O[c] + glu-L[c] <=> H[c] + NADPH[c] + NH4[c] + akg[c]	Tsac_2172
GLUSy	Glutamate synthase(NADPH)	H[c] + NADPH[c] + gln-L[c] + akg[c] -> NADP[c] + 2 glu-L[c]	Tsac_1234
LDH_L	L-lactate dehydrogenase	lac-L[c] + NAD[c] <=> NADH[c] + H[c] + pyr[c]	Tsac_0179
PTAr	Phosphotransacetylase	Pi[c] + AcCoA[c] <=> CoA[c] + actp[c]	Tsac_1744
HFS	Ferredoxin hydrogenase	Fdred[c] + 2 h[c] <==> Fdox + H2[c]	Tsac_1550 & Tsac_1551 & Tsac_1552 & Tsac_1553