Comparative phenotypic analysis and genome sequence of *Clostridium beijerinckii* SA-1, an offspring of NCIMB 8052

Walter J. Sandoval-Espinola,¹ Satya T. Makwana,¹ Mari S. Chinn,² Michael R. Thon,³ M. Andrea Azcárate-Peril⁴ and José M. Bruno-Bárcena¹

Correspondence

José M. Bruno-Bárcena
jbbarcen@ncsu.edu

¹Department of Microbiology, North Carolina State University, Raleigh, NC 27695-7615, USA

²Department of Biological and Agricultural Engineering, North Carolina State University, Raleigh, NC 27695-7615, USA

³Centro Hispano-Luso de Investigaciones Agrarias (CIALE), Departamento de Microbiología y Genética, Universidad de Salamanca, Calle Del Duero 12, Villamayor 37185, Spain

⁴Department of Cell Biology and Physiology, School of Medicine and Microbiome Core Facility, Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, NC 27599-7545, USA

Production of butanol by solventogenic clostridia is controlled through metabolic regulation of the carbon flow and limited by its toxic effects. To overcome cell sensitivity to solvents, stress-directed evolution methodology was used three decades ago on *Clostridium beijerinckii* NCIMB 8052 that spawned the SA-1 strain. Here, we evaluated SA-1 solventogenic capabilities when growing on a previously validated medium containing, as carbon- and energy-limiting substrates, sucrose and the products of its hydrolysis D-glucose and D-fructose and only D-fructose. Comparative small-scale batch fermentations with controlled pH (pH 6.5) showed that SA-1 is a solvent hyper-producing strain capable of generating up to 16.1 g l⁻¹ of butanol and 26.3 g l⁻¹ of total solvents, 62.3% and 63% more than NCIMB 8052, respectively. This corresponds to butanol and solvent yields of 0.3 and 0.49 g g⁻¹, respectively (63% and 65% increase compared with NCIMB 8052). SA-1 showed a deficiency in D-fructose transport as suggested by its 7 h generation time compared with 1 h for NCIMB 8052. To potentially correlate physiological behaviour with genetic mutations, the whole genome of SA-1 was sequenced using the Illumina GA IIx platform. PCR and Sanger sequencing were performed to analyse the putative variations. As a result, four errors were confirmed and validated in the reference genome of NCIMB 8052 and a total of 10 genetic polymorphisms in SA-1. The genetic polymorphisms included eight single nucleotide variants, one small deletion and one large insertion that it is an additional copy of the insertion sequence ISCb1. Two of the genetic polymorphisms, the serine threonine phosphatase *cbs_4400* and the solute binding protein *cbs_0769*, may possibly explain some of the observed physiological behaviour, such as rerouting of the metabolic carbon flow, deregulation of the D-fructose phosphotransferase transport system and delayed sporulation.

Received 15 May 2013

Accepted 24 September 2013

INTRODUCTION

The most recent crisis in rising oil prices has reinvigorated interest in solvent-producing clostridia, well known for their ability to generate acetone/butanol/ethanol (ABE)

Abbreviations: ABE, acetone/butanol/ethanol; CRISPR, clustered regularly interspaced short palindromic repeat; DOE JGI, US Department of Energy Joint Genome Institute; PTS, phosphotransferase system; SNV, single nucleotide variation.

The GenBank/EMBL/DBJ accession number for the sequence read archive of *Clostridium beijerinckii* ATCC 35702 is SRR191792.

from plant biomass feedstock. Butanol has attracted historical attention as an energy commodity and solventogenic clostridia have been the traditional organisms selected for its biological production. This taxonomically complex group of organisms, when cultured under the right conditions, possesses a metabolic ability to convert a diverse group of carbohydrates, including pentoses, hexoses and even polysaccharides, into butanol and other subproducts (Mitchell, 1997; Qureshi *et al.*, 2007, 2010; Lee *et al.*, 2008; Jang *et al.*, 2012a). Additionally, a large number of studies on ABE fermentation have been carried out

demonstrating that among solventogenic strains, *Clostridium beijerinckii* exhibits unique cellulolytic and xylanolytic activities as well as desirable properties such as sugar co-fermentation (Lee *et al.*, 1985a, b; Wang & Blaschek, 2011; Heluane *et al.*, 2011). These properties facilitate the exploitation by fermentation of alternative sustainable sugar-based feedstock that also contains vitamins and nitrogenous compounds required for successful ABE fermentations. Therefore, ABE fermentation can be achieved using various agricultural residues, including sorghum juice, cane molasses and other byproducts of the sugar industry (Najafpour & Shan, 2003). Sucrose has been considered an economical and suitable substrate for ABE fermentation (Quratulain *et al.*, 1995; Ni *et al.*, 2012, 2013), and the genes involved in its utilization have been studied and constitute an operon (Reid *et al.*, 1999). Molecular data suggest that sucrose and glucose uptake by *C. beijerinckii* takes place through a phosphoenolpyruvate-dependent phosphotransferase system (PTS) (Mitchell *et al.*, 1991; Tangney *et al.*, 1998; Reid *et al.*, 1999). However, only a small number of studies have been published evaluating *C. beijerinckii* utilizing sucrose as the sole carbon and energy source (Wang & Blaschek, 2011).

The history and study of the ABE bioprocess dates back to the early twentieth century and, since then, taxonomic classification of solvent-producing clostridia has been a veritable quagmire that continues to be elucidated (Collins *et al.*, 1994; Johnson *et al.*, 1997; Stackebrandt & Rainey, 1997). Nonetheless, there are physiological characteristics shared by all solventogenic strains that limit the implementation of cost-effective industrial bioprocesses. Problems affecting strain performance include metabolic carbon flow regulation, low butanol tolerance and genetic degeneration. Butanol toxicity has been considered historically to be one of the main limiting factors restricting biological butanol accumulation (Ezeji *et al.*, 2010), whilst genetic degeneration relates to the cell failing to maintain the capacity to produce solvents – a common phenomenon reported in solvent-producing clostridia (Kutzenok & Aschner, 1952; Kashket & Cao, 1993, 1995; Gottschal & Morris, 1981). Historical strategies to overcome poor solvent tolerance and to improve total butanol titres have involved the screening of resistant butanol-producing species after stress-directed evolution (Lin & Blaschek, 1983; Soucaille *et al.*, 1987; Annous & Blaschek, 1991; Quratulain *et al.*, 1995), the isolation of asporogenic, butanol-resistant mutant strains (Jain *et al.*, 1993) and the genetic modification of existing strains (Dai *et al.*, 2012; Mann *et al.*, 2012; Xiao *et al.*, 2011). The stress-directed evolution strategy that emphasized butanol resistance was chosen in 1983 to select a butanol-tolerant mutant of *C. beijerinckii* NCIMB 8052, initially named SA-1 (Lin & Blaschek, 1983) and later catalogued as *C. beijerinckii* ATCC 35702. SA-1 has been overlooked the last three decades and only a few studies have evaluated this strain, including the work assessing its α -amylase and glucoamylase activities (Chojacki & Blaschek, 1986). Recently, we

evaluated the nutritional requirements of SA-1 and proposed a medium that facilitates its growth under butanol stress, making it possible to obtain reliable continuous butanol production processes when the pH was kept at 6.5 (Heluane *et al.*, 2011). Under these conditions SA-1 also exhibited characteristics of delayed lysis during sporulation similar to a *Clostridium acetobutylicum* *lyt-1* mutant – a desirable phenotype that improves butanol accumulation (Van Der Westhuizen *et al.*, 1982). However, the limited amounts of functional data on SA-1 require new physiological evaluations essential for the interpretation of the genotypic differences between SA-1 and its parent strain NCIMB 8052.

Among solventogenic butanol-producing clostridia, there are two reference genomes publicly available: *C. acetobutylicum* ATCC 824 (Nolling *et al.*, 2001) and *C. beijerinckii* NCIMB 8052 (GenBank accession number NC_009617). In addition, a metabolic model has been proposed based on the NCIMB 8052 genome sequence (Milne *et al.*, 2011). The availability of these reference genomes facilitates comparative genome analysis and the discovery of new genetic polymorphisms involved potentially in the enhancement of butanol production (Jia *et al.*, 2012; Mann *et al.*, 2012). Thus, choosing closely related strains that show different phenotypic characteristics may facilitate a functional association with genomic polymorphisms.

The present study examined both physiological differences and genetic variations of *C. beijerinckii* SA-1 (offspring) compared with *C. beijerinckii* NCIMB 8052 (parent). We compared the solventogenic production capabilities of NCIMB 8052 and SA-1 growing in a carbon- and energy-limiting medium (Heluane *et al.*, 2011), and further identified and confirmed genetic polymorphisms in their genome sequences.

METHODS

Strains, media and cultivation methods. Strain SA-1/ATCC 35702 was deposited by Lin & Blaschek (1983) as an offspring strain derived from *C. acetobutylicum* ATCC 824. Later, Johnson *et al.* (1997) identified SA-1 as belonging to the *C. beijerinckii* species by DNA sequence similarity. After this finding, Blaschek recatalogued their ATCC 824 strain as *C. beijerinckii* NCIMB 8052 (Formanek *et al.*, 1997). For our study, *C. beijerinckii* NCIMB 8052 (ATCC 51743) and *C. beijerinckii* SA-1 (ATCC 35702) were obtained from the American Type Culture Collection. Species identity was verified by PCR amplification and sequencing of the 16S rRNA gene using the prokaryotic 16S rDNA universal primers 515F (5-GCGGATCC-TCTAGACTGCAGTGCCA-3) and 1492R (5-GGTTACCTTGTT-ACGACTT-3). The vegetative cells were routinely maintained at -80°C in M17 medium (Difco) containing 5% (v/w) glucose and glycerol (10%). The organisms were activated without heat shock in M17 containing the sugar or sugar combination to be tested. After activation, cultures were transferred into the same medium used in the reactors. The composition of the medium used for the batch experiments has been reported previously (Heluane *et al.*, 2011). The medium base components were autoclaved without the sugars and trace components, which were always sterilized by filtration and added aseptically to the medium reservoir. Combinations of sugars

(sucrose, D-glucose and D-fructose, and D-fructose) were aseptically added into the reactor to reach a 6% sugar final concentration.

Batch growth experiments were performed in a 2 l BIOSTAT Bplus bioreactor equipped with controllers for pH, temperature and agitation (Sartorius, BBI Systems). The temperature was set at 37 °C, the agitation speed was set at 250 r.p.m. and the pH was kept at 6.5 by the automated addition of 0.5 N KOH or 0.1 M H₃PO₄ into a working volume of 1400 ml. In order to maintain initial anaerobic conditions after inoculation, a continuous stream of sterile nitrogen gas at 0.1 v/v min⁻¹ was passed through the reactors, usually for between 5 and 10 min. Exact conditions for the fermentations are detailed when the different assays are presented in Results. The purity of the culture was confirmed daily by direct microscopic examination and by plating dilutions on M17 agar plates (Terzaghi & Sandine, 1975).

Preparation of total DNA and genome sequencing. Genomic DNA was isolated from *C. beijerinckii* SA-1 cells in exponential growth phase using the modified cetyltrimethylammonium bromide method (Wilson, 1994). DNA quantification was performed by comparison with DNA standards (15, 31, 63, 125, 250 and 500 ng) supplied by the US Department of Energy Joint Genome Institute (DOE JGI). Quantity One software (version 4.5.2) was used to create a calibration curve and gel imaging was performed using a Gel-Doc 2000 (Bio-Rad). Next-generation sequencing was performed on a Genome Analyser IIx (Illumina) by the DOE JGI. Sequences were mapped to the reference genome *C. beijerinckii* NCIMB 8052, and single nucleotides, small insertions and small deletions were analysed using MAQ version 0.7.1 (Li *et al.*, 2008). A genome-wide analysis of anomalous read pairs to identify large structural variations (inversions, translocations and large indels) was performed using BreakDancer (Chen *et al.*, 2009). Sequence data were analysed visually using the Integrative Genome Viewer (IGV) for confirmation of alignment assembly errors (Robinson *et al.*, 2011).

Nucleotide accession number. The Whole Genome Shotgun project has been deposited under GenBank sequence read archive accession number SRR191792.

Analytical Methods. Biomass proliferation in the fermentation tank was monitored using an in-line biomass sensor (Fundalux; Sartorius, BBI Systems) and also by measuring the OD₆₀₀ on a digital spectrophotometer (SmartSpec Plus; Bio-Rad). Sucrose, D-glucose and D-fructose were quantified by HPLC (Shimadzu) under isocratic conditions at 65 °C and a mobile phase of water at a 0.5 ml min⁻¹ flow rate using a Supelcogel Ca column (300 mm × 7.8 mm; Supelco Analytical) coupled to a refractive index detector. The products of the fermentations (butanol, butyric acid, acetone, acetic acid and ethanol) were quantified immediately after sample collection by a gas chromatograph (GC-8A) fitted with a flame ionization detector (Shimadzu). The products were separated in a GC SS Porapak Q 80/100 column (OV) using 200 kPa of nitrogen as the mobile phase with an injection temperature of 220 °C and a column temperature of 140 °C.

RESULTS

Batch cultivation revealed considerable physiological differences between *C. beijerinckii* NCIMB 8052 and *C. beijerinckii* SA-1

The literature reported that *C. beijerinckii* SA-1 was able to tolerate higher concentrations of butanol than NCIMB 8052 (Lin & Blaschek, 1983). Here, we confirmed that these two strains also display fundamental differences with

respect to growth, sugar utilization, solvent production and yields (Fig. 1, Table 1). The two strains were grown in pH (6.5)-controlled batch culture. Generation times of 1.5 h were achieved by both strains when growing with sucrose or glucose/fructose as the limiting carbon and energy sources. However, the sugar consumption and the co-fermentation profiles of the two strains differed; whilst NCIMB 8052 was able to co-ferment D-glucose/D-fructose and readily use D-fructose, SA-1 was unable to co-ferment both sugars.

NCIMB 8052 was able to accumulate concentrations of butanol and overall solvents consistent with previously reported maximum levels of butanol >10 g l⁻¹ and solvents >15 g l⁻¹ (Formanek *et al.*, 1997; Wang & Blaschek, 2011). Moreover, while both strains reached comparable biomass values, SA-1 revealed a solventogenic hyper-production phenotype and was able to generate up to 16.1 g butanol l⁻¹ and 26.3 g total ABE l⁻¹, an accumulation improvement of 62.3% and 63% compared with NCIMB 8052, respectively. Also, butanol and ABE yield values were significantly higher as compared with NCIMB 8052 under all conditions tested (Table 1).

Sequencing of *C. beijerinckii* SA-1, comparative assembly to the reference genome *C. beijerinckii* NCIMB 8052 and validation of the genomic variations

In order to generate a solid basis for understanding the similarities and differences in cell biology and physiology of *C. beijerinckii* strains, we performed genome sequencing and a comparative analysis. Whole-genome sequencing of strain SA-1 was performed on the Illumina GA IIx platform using one flow cell lane with 36-cycle paired-end chemistry (GenBank sequence read archive accession number SRR191792). This generated 55 995 398 paired-end reads with a mean read length of 98 bp, resulting in a mean sequence depth of 695 reads that corresponds to a 914-fold mean coverage.

The genome sequence evaluated in this study confirms the taxonomic assignment of SA-1 to *C. beijerinckii* within the clostridial solventogenic group. The ORF identifiers in *C. beijerinckii* SA-1 also correspond to the same numbers in *C. beijerinckii* NCIMB 8052. For clarity, we have denoted the *C. beijerinckii* SA-1 genes as 'cbs_ORF number' (Fig. 2a).

The *C. beijerinckii* SA-1 genome is 6 000 627 bp in size, does not contain plasmids and is predicted to encode 5243 protein-coding genes. The G + C content of the genome is 29.9%, which is the same as the parent strain *C. beijerinckii* NCIMB 8052 (GenBank accession number NC_009617). The SA-1 strain contains 94 tRNA genes; this is the same number of tRNA genes and same genome size as *C. beijerinckii* NCIMB 8052.

Assembly of the short sequence reads (GenBank sequence read archive accession number SRR191792) and analysis of the results using MAQ and BreakDancer identified genetic

Fig. 1. Fermentation profile experiments for *C. beijerinckii* NCIMB 8052 (left) and SA-1/ATCC 35702 (right) at 37 °C, controlled pH (6.5) in a validated medium (Heluane *et al.*, 2011) using as limiting carbon and energy source: (a, b) sucrose (6%, w/w), (c, d) D-glucose (3%, w/v) and D-fructose (3%, w/v), and (e, f) D-fructose (6%, w/v). Biomass was monitored by two methods: by the in-line biomass sensor Fundalux (arbitrary units) and by optical density.

polymorphisms that occurred in the offspring strain. The list of polymorphisms includes eight single nucleotide variants (SNV), one small deletion and one large insertion that is an additional copy of the insertion sequence ISCb1, belonging to the IS4 family (Liyanage *et al.*, 2000) inserted between ORFs *cbi_1788* and *cbi_1789*. To validate all genomic variations, we performed PCR and Sanger sequencing using chromosomal DNA from the strains SA-1 and NCIMB 8052 as templates (Fig. 2b).

SNVs

Overall, most of the SNVs were found within ORFs. Out of the original 12 SNVs identified by MAQ, eight were confirmed by PCR amplification and Sanger sequencing (Eton Bioscience), whilst four SNVs corresponded to errors that were found in the reference genome of *C. beijerinckii* NCIMB 8052 (*cbi_1046*, *cbi_R0117*, *cbi_2885* and *cbi_4308*). The corrections to these loci are reported in Tables 2 and 3.

Table 1. Kinetic and yield parameters obtained after pH-controlled batch fermentations of *C. beijerinckii* NCIMB 8052 and SA-1 strains at 37 °C, pH 6.5 in a validated medium (Heluane *et al.*, 2011)

Carbon/energy source	Parameter	<i>C. beijerinckii</i>	
		NCIMB 8052	SA-1
Sucrose 6 % (w/v)	Biomass (OD ₆₀₀ ml ⁻¹)	4.72	5.7
	Butanol (g l ⁻¹)	8.09 ± 0.27	13.48 ± 0.01
	Acetone (g l ⁻¹)	6.5 ± 0.13	11.22 ± 0.04
	Ethanol (g l ⁻¹)	0.49 ± 0.12	0.59 ± 0.01
	Total solvents (g l ⁻¹)	15.09 ± 0.38	25.30 ± 0.04
	Sugar utilized (%)	72.57	100
	Specific growth rate μ (h ⁻¹)	0.41	0.39
	Butanol yield (g g ⁻¹)	0.21	0.24
	Solvents yield (g g ⁻¹)	0.39	0.45
	Butanol productivity (g l ⁻¹ h ⁻¹)	0.13	0.34
D-Glucose 3 % (w/v) and D-fructose 3 % (w/v)	Biomass (OD ₆₀₀ ml ⁻¹)	9.65	11
	Butanol (g l ⁻¹)	10.01 ± 0.48	16.05 ± 0.1
	Acetone (g l ⁻¹)	6.40 ± 0.15	9.92 ± 0.02
	Ethanol (g l ⁻¹)	0.28 ± 0.03	0.43 ± 0.03
	Total solvents (g l ⁻¹)	16.85 ± 0.33	26.4 ± 0.1
	Sugar utilized (%)	100	100 glucose/67 fructose
	Specific growth rate μ (h ⁻¹)	0.45	0.46
	Butanol yield (g g ⁻¹)	0.19	0.3
	Solvents yield (g g ⁻¹)	0.32	0.49
	Butanol productivity (g l ⁻¹ h ⁻¹)	0.62	1.14
D-Fructose 6 % (w/v)	Biomass (OD ₆₀₀ ml ⁻¹)	6.82	5.72
	Butanol (g l ⁻¹)	11.53 ± 0.12	12.6 ± 0.44
	Acetone (g l ⁻¹)	6.26 ± 0.14	11.33 ± 0.03
	Ethanol (g l ⁻¹)	0.80 ± 0.007	0.71 ± 0.09
	Total solvents (g l ⁻¹)	18.59 ± 0.26	24.64 ± 0.53
	Sugar utilized (%)	100	100
	Specific growth rate μ (h ⁻¹)	0.66	0.09
	Butanol yield (g g ⁻¹)	0.24	0.27
	Solvents yield (g g ⁻¹)	0.4	0.53
	Butanol productivity (g l ⁻¹ h ⁻¹)	0.54	0.17

cbs_0769 encodes a putative solute binding protein that is interrupted by a nonsense mutation at residue 51, potentially making this protein non-functional. A BLAST (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) analysis showed that this protein contains motifs similar to an ABC-type sugar transporter in *Clostridium thermocellum*. Previous work with *C. thermocellum* demonstrated that these ABC transport systems are able to transport cellodextrin and laminaribiose into the cell (Nataf *et al.*, 2009). We speculate that the mutation in *cbs_0769* may impact SA-1's ability to transport D-fructose.

In *cbs_1854*, a valine was substituted for alanine in the amino acid sequence for a putative peptidase S8/S53 subtilisin kexin sedolisin, a member of the serine-carboxyl peptidases (Wlodawer *et al.*, 2003).

ORF *cbs_1935* encodes a putative PucR transcriptional regulator. PucR regulators are responsible for regulating purine catabolism (Schultz *et al.*, 2001). In SA-1, the SNV changed a histidine residue, positively charged, for a polar

uncharged asparagine residue. This missense mutation at residue 47 (H47N) occurred in a conserved region of the protein and may have an effect on the function of the protein in regulating purine utilization (Xi *et al.*, 2000; Heluane *et al.*, 2011).

cbs_1975 encodes an adenine deaminase, responsible for converting adenosine to inosine in the nucleotide salvage pathway (Kilstrup *et al.*, 2005). In SA-1, a SNV at position 1707 position in the coding sequence of a G→T produced an amino acid change from alanine to serine at position 11 (A11S). This missense mutation does not occur within the conserved protein functional region.

cbs_4400 encodes a serine threonine phosphatase. A SNV at position 1766 position in the coding sequence is a C→A transversion that caused an arginine to leucine substitution at position 589 (R589L) and is outside of a conserved domain of the protein. Multiple cellular processes are regulated by this type of protein, including sugar transport and endospore formation (Obuchowski, 2005).

Fig. 2. *C. beijerinckii* SA-1 genome atlas providing the physical locations of the putative and confirmed polymorphisms. (a) Polymorphisms discovered performing *in silico* analysis with MAQ and BreakDancer algorithms. (b) Polymorphisms validated after PCR and Sanger sequencing methods. Image produced using Geneious (<http://www.geneious.com>). Purple, insertions; navy blue, deletions; black, SNVs; maroon, putative insertions or deletions (indels); gold, putative small indels; green, inversions; cyan, internal translocations. The term 'Cbs' is used here to refer specifically to genes that are in the SA-1 strain.

ORF *cbs_4761*, encoding a cell wall binding motif, has a mutation at position 5556196 (A→G transition; V34A).

PCR and Sanger sequencing also confirmed two mutations occurring in non-coding regions of the genome. The SNV between ORFs *cbs_2247* and *cbs_2248* is a G→A transition, and the SNV between coding sequences *cbs_R0121* (16S RNA) and *cbs_R0122* (5S RNA) is an A→G change. In order to determine whether or not these non-coding regions contained regulatory RNAs, an alignment was performed against the Rfam RNA family database located at <http://rfam.sanger.ac.uk/> (Griffiths-Jones *et al.*, 2003). No conserved RNA families were found within these non-coding regions.

Small insertions and deletions

Eight small insertions and deletions (1–5 bp) were initially identified using MAQ and visualized with IGV software (Li *et al.*, 2008; Robinson *et al.*, 2011) (Table 4). Seven were found to be false positives, whilst a small deletion of the nucleotide sequence AAATA was confirmed using PCR and Sanger sequencing (primers 5'-TTGCACAAGGAAAA-CAAG-3' and 5'-TACAGAATCCCAACTATCAG-3'). At position 1099984 in the SA-1 genome, the deletion caused a frame-shift mutation that abolished the stop codon and extended a hypothetical protein (*cbs_0917*) by three amino acid residues. The detection of false mutations by the MAQ algorithm has been reported previously (Krawitz *et al.*,

Table 2. Single nucleotide corrections to the reference sequence (GenBank accession number CP000721.1) from *C. beijerinckii* NCIMB 8052

Table 3 shows physical locations and primer sequences used to verify these polymorphisms. CDS, coding sequence.

ORF number	Putative function	Position	Type	Reference sequence	Actual sequence
<i>cbei_1046</i>	Band 7 protein CDS	1 255 087	SNV	A	G
<i>cbei_R0117</i>	23S rRNA CDS	5 747 394	SNV	A	G
<i>cbei_2885</i>	L-lactate transport protein CDS	3 369 627	Indel	–A	+A
<i>cbei_4308</i>	Protein glutamate O-methyltransferase CDS	4 962 014	Indel	+C	–C

Table 3. List of point mutations detected and confirmed in the chromosome sequence of *C. beijerinckii* SA-1/ATCC 35702 using the MAQ algorithm (Li *et al.*, 2008)

Coverage, variant frequencies and reference frequencies determined through genome assembly using Geneious (www.geneious.com). PCR amplifications were performed using genomic DNA from both NCIMB 8052 (reference strain) and SA-1 as template. COG, Clusters of orthologous groups; CDS, coding sequence. IUPAC naming conventions are used to communicate base and amino acid changes. Primer names are indicated by their position and direction (F, forward; R, reverse). The term '*cbs*' here is used to emphasize the fact that the mutation occurs in the derivative strain, *C. beijerinckii* SA-1/ATCC 35702. Term, Termination codon.

ORF number and strand orientation	Associated COG	CDS size (bp)	Mutation position	Putative gene function	SNV change in CDS and amino acid change in CDS	Primers	Coverage	Variant frequencies (%)	Reference frequencies (%)
<i>cbs</i> _0769 +	COG1653G	1329	935 449	Solute binding protein	C151T Q51Term	934943F 5'-TCGGTTGATACTGATTATG-3' 935763R 5'-TTTACCTGGAAAAGTGATAC-3'	498	100.0	
<i>cbs</i> _1854 +	S (unknown function)	1737	2 149 286	Peptidase S8/S53 subtilisin kexin sedolisin	T698C V323A	2149110F 5'-GTAGTTTGGGAAATCATAAGG-3' 2149781R 5'-CTGCAACATCTATTCTATCAG-3'	403	100.0	
<i>cbs</i> _1935 +	COG2508TQ	1155	2 234 607	PucR transcriptional regulator	C139A H47N	2234281F 5'-CATCCGAATCATTTTATAGC-3' 2234964R 5'-ATGACTGTCTCTCAAATATC-3'	326	99.4	0.3
<i>cbs</i> _1975 +	COG1001F	1707	2 295 776	Adenine deaminase	G31T A11S	2295543F 5'-GCCTAAAACAAGCTATATC-3' 2296459R 5'-CTAAGCATCACATACATAC-3'	162	99.4	
Between <i>cbs</i> _2248 and <i>cbs</i> _2247	–	402 (interval length between coding regions)	2 611 044	Non-coding region between rRNA adenine dimethylase CDS and hypothetical protein CDS	G : A	2610380F 5'-AAACAAGATGGCATAAGC-3' 2611343R 5'-CCTCCAAGATTAATATAGG-3'			
<i>cbs</i> _4400 –	COG0631T	1839	5 075 403	Serine threonine protein phosphatase-like protein	C1766A R589L	5073615F 5'-GCATGATTCTGTCATTTAG-3' 5074072R 5'-AGCAGGTGATTATGTAAG-3'			
<i>cbs</i> _4761 –	COG5263R	1218	5 556 196	Cell wall binding repeat-containing protein	A101G V34A	5554176F 5'-CCCATTAACGTAATACTG-3' 5554911R 5'-GTCTATGGCTTATAACAC-3'	472	99.4	
Between <i>cbs</i> _R0121 and <i>cbs</i> _R0122	–	1112 (interval length between coding regions)	5 769 733	Non-coding region between 16S RNA CDS and 5S RNA CDS	A : G	5769433F 5'-GTTTCTGACAGCTTACTC-3' 5769939R 5'-TGTATCCAGGGGTAATTG-3'	738	57.2	42.5

Table 4. *In silico*-detected and not corroborated SNVs and small insertions and deletions (indels) from the sequence of *C. beijerinckii* SA-1/ATCC 35702

Coverage, variant frequencies, and reference frequencies determined through genome assembly using Geneious (<http://www.geneious.com>). PCR amplifications were performed using genomic DNA from both NCIMB 8052 (reference strain) and SA-1 as template. COG, Clusters of orthologous groups; CDS, coding sequence. Primer positions and directions are indicated by the number followed by F (forward) or R (reverse). NA, Amino acid sequence not available because base change is within a non-coding region.

Type	ORF number	Associated COG	CDS size (bp)	Position	Gene function	Primer	Coverage	Variant frequency (%)	Reference frequency (%)
SNVs	<i>cbs_0433</i>	COG0653U	2568	524 879	Sec A	524413F 5'-TTCATCAAGCGATAGAAG-3' 525161R 5'-TTGAGCATTCTCAATAGC-3'	792	34.2	65.0
	<i>cbs_2250</i>	COG2205T	981	2 613 440	Histidine kinase	2612973F 5'-GAAAGGGATAATCAACAAAG-3' 2613589R 5'-CTTTATCCAGTTTCACATAG-3'	475	99.8	
	<i>cbs_1046</i>	COG0330O	948	1 255 087	Band 7 protein	1254781F 5'-TCAGCTACAATAATATGAG-3' 1255488R 5'-GATGCCTTCTAATTGATTC-3'			
	<i>cbs_R0117</i>	NA	2909	5 747 394	23S rRNA	5747160F 5'-GAGCTAATCACGCAATC-3' 5748001R 5'-CGCCTCCTAAAATGTAAC-3'			
	<i>cbs_2885</i>	COG1620C	1465	3 369 627	L-lactate transport protein CDS	3369256F 5'-CTAGAACTGCTGAAAATG-3' 3370068R 5'-GTAGCAATTGTTGGAAATG-3'			
Small indels	<i>cbs_4845</i>	S (unknown function)	1173	5 667 579	Mandelate racemase/ muconate lactonizing protein CDS	5667420F 5'-TATCATCATATGCTCCTAAG-3' 5668012R 5'-TAGAGCAGATGGATATACAG-3'	475	99.8	
	<i>cbs_4963</i>	COG2972T	1761	5 838 970	Integral membrane sensor signal transduction histidine kinase CDS	5838895F 5'-AATAATCCACTTCCCTTTAG-3' 5839486R 5'-ACAAATAAAGCCTCACTTTC-3'			
	<i>cbs_5024</i>	COG1506E	1167	5 907 883	α/β -fold family hydrolase CDS	5907769F 5'-CAAGCACGAATGTAACAG-3' 5908400R 5'-CATGGTGGAGAAATCTATG-3'			
	<i>cbs_4308</i>	COG1352T	713	4 962 014	Protein-glutamate O-methyltransferase CDS	4961567F 5'-CTAAGGCTTTCATTATTCC-3' 4962360R 5'-TGAGCCATATTCATTAGG-3'			
	Non-coding region	NA	392	1 314 429	Located between Cbei_1100 CDS and hypothetical protein CDS	1313971F 5'-GAGTGGTAATATACTAGGTAAG-3' 1314664R 5'-CTCTAGTTGGTAGTTCATATTC-3'			
	Non-coding region	NA	370	3 369 630	Located between L-lactate transport protein CDS and electron transfer flavoprotein, α subunit-like protein CDS	3369256F 5'-CTAGAACTGCTGAAAATG-3' 3370068R 5'-GTAGCAATTGTTGGAAATG-3'			

Table 5. *In silico*-detected and not corroborated large insertions, deletions, and structural variations

PCR amplifications were performed using genomic DNA from both NCIMB 8052 (reference strain) and SA-1 as template. CDS, Coding sequence; COG, clusters of orthologous groups. Primer positions are indicated by the number followed by the direction F (forward) or R (reverse).

Large indels				
Location		Size (bp)	Position	Primer
Between hypothetical protein CDS and hypothetical protein CDS		49	3 829 123–3 829 172	3828725F 5'-ATGCAGCCAAATAATGAC-3'
Between hypothetical protein CDS and hypothetical protein CDS		29	3 829 172–3 829 201	3829751R 5'-GAGCTCGATAATTACTAAG-3'
Between hydroxylamine reductase CDS and heavy metal transport detoxification protein CDS		37	1 685 323–1 685 360	1685136F 5'-TATATGCAGCAGAGCTAC-3'
				1685848R 5'-TCTTACCCCTGATGTTTC-3'
Between EmrB/QacA family drug resistance transporter CDS and CBS domain-containing protein CDS		35	1 797 550–1 797 585	1797141F 5'-GCTATGATGCCTATTAGC-3'
				1798121R 5'-CGCCTCTACTAATAACTCC-3'
Between MATE efflux family protein CDS and TPR repeat-containing protein CDS		30	2 039 733–2 039 763	2039609F 5'-TGGCATAAGGATATAATCG-3'
				2040035R 5'-AATGCCTTATCTAGCTTTC-3'
Between GCN5-related <i>N</i> -acetyltransferase CDS and hypothetical protein CDS		26	2 961 163–2 961 189	2960642F 5'-TCGGATTTATCTAGCAAAG-3'
				2961396R 5'-CTTTTATCGCACGAATATG-3'
Between glycosyltransferase family protein CDS and triple helix repeat-containing collagen CDS		26	3 001 786–3 001 812	3001353F 5'-CCAGTATGTAGCTAAGTTTC-3'
				3002219R 5'- CTTCTAAATAGGTCATTAGG-3'
Between thioredoxin reductase CDS and hypothetical protein CDS		28	3 117 231–3 117 259	3117006F 5'-CCAGGATAATTTTCCACTTC-3'
				3002219R 5'-AGACCTAGGCTATAGTTTTG-3'
Between metal-dependent phosphohydrolase CDS and sulfate adenylyltransferase large subunit CDS		37	4 819 865–4 819 902	4819453F 5'-GTTAGGCTCATCTAAATAC-3'
				4820145R 5'-GAAACTCTAGGAGATTTC-3'
Structural variations				
ORF number	Associated COG/type	Size (bp)	Position	Primer
<i>cbs</i> _4437	COG4974L/prophage inversion	27	5 120 090–5 120 520	5119799F 5'-CACCCCTTTCTACAATTAAC-3'
				5120810R 5'-TGCCATTGAGAAATATTGC-3'
<i>cbs</i> _4182	COG1352NT/deletion	483	4 812 971–4 814 517	4812830F 5'-AAGTTATCTGTGGTTCTATC-3'
<i>cbs</i> _4182	COG1352NT/deletion	529	4 814 951–4 814 980	4813942R 5'-ATTGAAGAAATGGGTGTATC-3'
<i>cbs</i> _0060	S (unknown function)/insertion	149	79 389–79 831	79067F 5'-TGCTATATGTAGGGTAAAAG-3'
				80063R 5'-CACGTCATTCAATTTATCC-3'
<i>cbs</i> _0878– <i>cbs</i> _0932	COG1476K/internal translocation (prophage ITX1)	39 411	1 071 995–1 111 510	1071556F 5'-GACCAGAACAACTATTTAC-3'
				1072303R 5'-TAAGGCTACAGATGATAAAG-3'
				1111245F 5'-AGATGGCACATTGTTTTG-3'
				1111872R 5'-ACTGCTGACCTTTATAGATAG-3'
<i>cbs</i> _1605– <i>cbs</i> _1680	COG1961L/internal translocation (prophage ITX2)	75 644	1 881 845–1 957 557	1881067F 5'-AAAATTGGACGAGGTAATAG-3'
				1958001R 5'-CAGTGAATAGTGCTATATAC-3'
				1957306F 5'-TATGAAAAGGAGCATATCAG-3'
				1883070R 5'-ATGAACTTCAGGTCTATGG-3'

2010; Hamada *et al.*, 2011). With regard to DNA repeats in the sequence, it became apparent that MAQ and BreakDancer had difficulty in identifying mutations in these regions; this limitation of the alignment software in the *in silico* analysis has been reported elsewhere and remains to be resolved (Treangen & Salzberg, 2012).

Large insertions, deletions and structural variations

Larger insertions and deletions (20–40 bp) as well as structural variations (e.g. chromosomal translocations and inversions) were analysed using BreakDancer version 1.0 (Chen *et al.*, 2009). The initial *in silico* analysis produced 10 large putative indels (Table 5). One was found to be an additional copy of the insertion sequence from the IS4 family located between a hypothetical protein (*cb*s₁₇₈₈) and a polysaccharide deacetylase (*cb*s₁₇₈₉) at position 2 073 281 in the SA-1 genome. Using forward primer 5'-TTGCACCTAATGTGTAATC-3' and reverse primer 5'-CTATCTCCTCTTTTCATTCTATC-3', this insertion sequence was confirmed by PCR and Sanger sequencing. Previous work has shown that *C. beijerinckii* NCIMB 8052 contains at least nine copies of an insertion sequence from the IS4 family (Liyanage *et al.*, 2000) (Fig. 3).

Two overlapping putative indels were also confirmed in a non-coding region between 3 829 123 and 3 829 201, which had AT-rich regions where runs and repeats were common (64 repeats $\geq$ 4 bp in length).

Six structural variations were suggested by the *in silico* analysis; however, through PCR and Sanger sequencing methods we have determined these structural variations were false positives (Table 5). One of these structural variations included a potential inversion at ORF *cb*s₄₄₃₇. An attempt to amplify this structural variation with PCR proved to be difficult and upon further analysis it was found that this region contained 93 repeats $\geq$ 6 bp in

length. In order to determine whether or not these repeats were CRISPR (clustered regularly interspaced short palindromic repeat) elements, a basic local alignment with the SA-1 genome was performed against the CRISPR database available at <http://crispr.u-psud.fr/> (Grissa *et al.*, 2007a, b, 2008; Garrett *et al.*, 2011). It was found that there were no previously characterized CRISPR elements present within ORF *cb*s₄₄₃₇ or the non-coding region between positions 3 829 123 and 3 829 201.

DISCUSSION

The development and characterization of butanol-producing strains is essential for the continued advancement of renewable energy technology. In this study, we reveal the solventogenic hyper-production phenotype of *C. beijerinckii* SA-1/ATCC 35702 (a derivative strain of *C. beijerinckii* NCIMB 8052). Batch cultivations of SA-1 on a previously validated medium (Heluane *et al.*, 2011) compared with NCIMB 8052 showed higher physiological differences in solvent accumulation and sugar utilization than those observed initially in extruded broth (Lin & Blaschek, 1983). SA-1 displays a metabolic carbon flow that is redirected mainly to the production of increased amounts of butanol and acetone. Redirecting the metabolic flow of carbon towards solvents to reach higher final butanol titres and yields has been shown previously by genetically modified *C. acetobutylicum* strains with disrupted genes responsible for butyric acid production (Harris *et al.*, 2000; Jang *et al.*, 2012b). Moreover, the overall butanol accumulation, ABE yields and solvent productivity of SA-1 are comparable to values previously reported for *C. beijerinckii* BA101, another butanol hyper-producing offspring of NCIMB 8052 (Formanek *et al.*, 1997; Ezeji *et al.*, 2007a, b; Qureshi & Blaschek, 2000; Lienhardt *et al.*, 2002). BA101 solventogenic carbon flow redirection has been correlated with a 50 % lower PTS system activity, including reduced fructose PTS activity when growing in D-glucose (Lee & Blaschek, 2001; Lee *et al.*, 2005).

The comparative analysis of the genomes of SA-1 and *C. beijerinckii* NCIMB 8052 enabled us to detect four errors in the reference genome, and a relatively low number of SNVs, indels and chromosomal rearrangements in SA-1, that suggest relative chromosome stability in these important solventogenic *Clostridium* strains (Kutzenok & Aschner, 1952; Gottschal & Morris, 1981; Kashket & Cao, 1993, 1995). Butanol concentration and tolerance limits are not correlated directly (Harris *et al.*, 2000) and no differences were found to explain previously reported butanol tolerance of SA-1 (Lin & Blaschek, 1983). However, the differences in D-fructose utilization and the SNV found in *cb*s₀₇₆₉ could be evidence that this putative solute binding protein may be involved in D-fructose transport. The understanding of sucrose utilization and its regulation has obvious industrial significance as it is possible that D-fructose may repress the uptake of sucrose

Fig. 3. Agarose gel (1 %) showing PCR amplification of the novel insertion sequence within SA-1 (lane 2) compared with the WT NCIMB 8052 (lane 3). Lane 1, 10 kb Logic DNA Marker (Lambda Biotech). Lane 4, PCR with no template added (negative control).

(Tangney *et al.*, 1998). Additionally, the polymorphism identified in a serine threonine phosphatase (*cbs_4400*) may be responsible for deregulation of a SA-1 PTS transport system, for delayed sporulation and/or involved in the increased production of butanol (Mitchell *et al.*, 1991).

Once genetic modification tools are more advanced for *C. beijerinckii*, the insights unveiled here will support future functional and metabolic engineering work. Consequently, genetic modification efforts targeting these polymorphisms will allow us to correlate the proposed gene functions and the phenotypic differences, including metabolic carbon flow redirection and sugar transport in *C. beijerinckii*.

ACKNOWLEDGEMENTS

This work was partially supported by the College of Life Sciences at NC State University, the North Carolina Agricultural Research Service and the Biofuels Center of North Carolina. The work conducted by the US DOE JGI was supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231.

REFERENCES

- Annous, B. A. & Blaschek, H. P. (1991). Isolation and characterization of *Clostridium acetobutylicum* mutants with enhanced amyolytic activity. *Appl Environ Microbiol* **57**, 2544–2548.
- Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q. & other authors (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677–681.
- Chojacki, A. & Blaschek, H. P. (1986). Effect of carbohydrate source on alpha-amylase and glucoamylase formation by *Clostridium acetobutylicum* SA-1. *J Ind Microbiol* **1**, 63–67.
- Collins, M. D., Lawson, P. A., Willems, A., Cordoba, J. J., Fernandez-Garayzabal, J., Garcia, P., Cai, J., Hippe, H. & Farrow, J. A. E. (1994). The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. *Int J Syst Bacteriol* **44**, 812–826.
- Dai, Z., Dong, H., Zhu, Y., Zhang, Y., Li, Y. & Ma, Y. (2012). Introducing a single secondary alcohol dehydrogenase into butanol-tolerant *Clostridium acetobutylicum* Rh8 switches ABE fermentation to high level IBE fermentation. *Biotechnol Biofuels* **5**, 44.
- Ezeji, T., Qureshi, N. & Blaschek, H. P. (2007a). Butanol production from agricultural residues: impact of degradation products on *Clostridium beijerinckii* growth and butanol fermentation. *Biotechnol Bioeng* **97**, 1460–1469.
- Ezeji, T. C., Qureshi, N. & Blaschek, H. P. (2007b). Production of acetone butanol (AB) from liquefied corn starch, a commercial substrate, using *Clostridium beijerinckii* coupled with product recovery by gas stripping. *J Ind Microbiol Biotechnol* **34**, 771–777.
- Ezeji, T., Milne, C., Price, N. D. & Blaschek, H. P. (2010). Achievements and perspectives to overcome the poor solvent resistance in acetone and butanol-producing microorganisms. *Appl Microbiol Biotechnol* **85**, 1697–1712.
- Formanek, J., Mackie, R. & Blaschek, H. P. (1997). Enhanced butanol production by *Clostridium beijerinckii* BA101 grown in semidefined P2 medium containing 6 percent maltodextrin or glucose. *Appl Environ Microbiol* **63**, 2306–2310.
- Garrett, R. A., Vestergaard, G. & Shah, S. A. (2011). Archaeal CRISPR-based immune systems: exchangeable functional modules. *Trends Microbiol* **19**, 549–556.
- Gottschal, J. C. & Morris, J. G. (1981). Non-production of acetone and butanol by *Clostridium acetobutylicum* during glucose and ammonium limitation in continuous culture. *Biotechnol Lett* **3**, 525–530.
- Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. (2003). Rfam: an RNA family database. *Nucleic Acids Res* **31**, 439–441.
- Grissa, I., Vergnaud, G. & Pourcel, C. (2007a). CRISPRfinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35** (Web Server), W52–W57.
- Grissa, I., Vergnaud, G. & Pourcel, C. (2007b). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172.
- Grissa, I., Vergnaud, G. & Pourcel, C. (2008). CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **36** (Web Server), W145–W148.
- Hamada, M., Wijaya, E., Frith, M. C. & Asai, K. (2011). Probabilistic alignments with quality scores: an application to short-read mapping toward accurate SNP/indel detection. *Bioinformatics* **27**, 3085–3092.
- Harris, L. M., Desai, R. P., Welker, N. E. & Papoutsakis, E. T. (2000). Characterization of recombinant strains of the *Clostridium acetobutylicum* butyrate kinase inactivation mutant: need for new phenomenological models for solventogenesis and butanol inhibition? *Biotechnol Bioeng* **67**, 1–11.
- Heluane, H., Evans, M. R., Dagher, S. F. & Bruno-Bárcena, J. M. (2011). Meta-analysis and functional validation of nutritional requirements of solventogenic clostridia growing under butanol stress conditions and cointilization of D-glucose and D-xylose. *Appl Environ Microbiol* **77**, 4473–4485.
- Jain, M. K., Beacom, D. & Datta, R. (1993). Mutant strain of *C. acetobutylicum* and process for making butanol. US Patent US 5192673A.
- Jang, Y. S., Lee, J., Malaviya, A., Seung, Y., Cho, J. H. & Lee, S. Y. (2012a). Butanol production from renewable biomass: rediscovery of metabolic pathways and metabolic engineering. *Biotechnol J* **7**, 186–198.
- Jang, Y. S., Lee, J. Y., Lee, J., Park, J. H., Im, J. A., Eom, M. H., Lee, J., Lee, S. H., Song, H. & other authors (2012b). Enhanced butanol production obtained by reinforcing the direct butanol-forming route in *Clostridium acetobutylicum*. *MBio* **3**, e00314-12.
- Jia, K., Zhang, Y. & Li, Y. (2012). Identification and characterization of two functionally unknown genes involved in butanol tolerance of *Clostridium acetobutylicum*. *PLoS ONE* **7**, e38815.
- Johnson, J. L., Toth, J., Santiwatanakul, S. & Chen, J. S. (1997). Cultures of “*Clostridium acetobutylicum*” from various collections comprise *Clostridium acetobutylicum*, *Clostridium beijerinckii*, and two other distinct types based on DNA-DNA reassociation. *Int J Syst Bacteriol* **47**, 420–424.
- Kashket, E. R. & Cao, Z. Y. (1993). Isolation of a degeneration resistant mutant of *Clostridium acetobutylicum* NCIMB 8052. *Appl Environ Microbiol* **59**, 4198–4202.
- Kashket, E. R. & Cao, Z. Y. (1995). Clostridial strain degeneration. *FEMS Microbiol Rev* **17**, 307–315.
- Kilstrup, M., Hammer, K., Ruhdal Jensen, P. & Martinussen, J. (2005). Nucleotide metabolism and its control in lactic acid bacteria. *FEMS Microbiol Rev* **29**, 555–590.
- Krawitz, P., Rödelberger, C., Jäger, M., Jostins, L., Bauer, S. & Robinson, P. N. (2010). Microindel detection in short-read sequence data. *Bioinformatics* **26**, 722–729.

- Kutzenok, A. & Aschner, M. (1952). Degenerative processes in a strain of *Clostridium butylicum*. *J Bacteriol* **64**, 829–836.
- Lee, J. & Blaschek, H. P. (2001). Glucose uptake in *Clostridium beijerinckii* NCIMB 8052 and the solvent-hyperproducing mutant BA101. *Appl Environ Microbiol* **67**, 5025–5031.
- Lee, S. F., Forsberg, C. W. & Gibbins, L. N. (1985a). Cellulolytic activity of *Clostridium acetobutylicum*. *Appl Environ Microbiol* **50**, 220–228.
- Lee, S. F., Forsberg, C. W. & Gibbins, L. N. (1985b). Xylanolytic activity of *Clostridium acetobutylicum*. *Appl Environ Microbiol* **50**, 1068–1076.
- Lee, J., Mitchell, W. J., Tangney, M. & Blaschek, H. P. (2005). Evidence for the presence of an alternative glucose transport system in *Clostridium beijerinckii* NCIMB 8052 and the solvent-hyperproducing mutant BA101. *Appl Environ Microbiol* **71**, 3384–3387.
- Lee, S. Y., Park, J. H., Jang, S. H., Nielsen, L. K., Kim, J. & Jung, K. S. (2008). Fermentative butanol production by clostridia. *Biotechnol Bioeng* **101**, 209–228.
- Li, H., Ruan, J. & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851–1858.
- Lienhardt, J., Schripsema, J., Qureshi, N. & Blaschek, H. P. (2002). Butanol production by *Clostridium beijerinckii* BA101 in an immobilized cell biofilm reactor: increase in sugar utilization. *Appl Biochem Biotechnol* **98–100**, 591–598.
- Lin, Y. L. & Blaschek, H. P. (1983). Butanol production by a butanol-tolerant strain of *Clostridium acetobutylicum* in extruded corn broth. *Appl Environ Microbiol* **45**, 966–973.
- Liyanage, H., Holcroft, P., Evans, V. J., Keis, S., Wilkinson, S. R., Kashket, E. R. & Young, M. (2000). A new insertion sequence, ISCb1, from *Clostridium beijerinckii* NCIMB 8052. *J Mol Microbiol Biotechnol* **2**, 107–113.
- Mann, M. S., Dragovic, Z., Schirmacher, G. & Lütke-Eversloh, T. (2012). Over-expression of stress protein-encoding genes helps *Clostridium acetobutylicum* to rapidly adapt to butanol stress. *Biotechnol Lett* **34**, 1643–1649.
- Milne, C. B., Eddy, J. A., Raju, R., Ardekani, S., Kim, P. J., Senger, R. S., Jin, Y. S., Blaschek, H. P. & Price, N. D. (2011). Metabolic network reconstruction and genome-scale model of butanol-producing strain *Clostridium beijerinckii* NCIMB 8052. *BMC Syst Biol* **5**, 130.
- Mitchell, W. J. (1998). Physiology of carbohydrate to solvent conversion by clostridia. *Adv Microb Physiol* **39**, 31–130.
- Mitchell, W. J., Shaw, J. E. & Andrews, L. (1991). Properties of the glucose phosphotransferase system of *Clostridium acetobutylicum* NCIB 8052. *Appl Environ Microbiol* **57**, 2534–2539.
- Najafpour, G. D. & Shan, C. P. (2003). Enzymatic hydrolysis of molasses. *Bioresour Technol* **86**, 91–94.
- Nataf, Y., Yaron, S., Stahl, F., Lamed, R., Bayer, E. A., Scheper, T. H., Sonenshein, A. L. & Shoham, Y. (2009). Cellodextrin and laminaribiose ABC transporters in *Clostridium thermocellum*. *J Bacteriol* **191**, 203–209.
- Ni, Y., Wang, Y. & Sun, Z. (2012). Butanol production from cane molasses by *Clostridium saccharobutylicum* DSM 13864: batch and semicontinuous fermentation. *Appl Biochem Biotechnol* **166**, 1896–1907.
- Ni, Y., Xia, Z., Wang, Y. & Sun, Z. (2013). Continuous butanol fermentation from inexpensive sugar-based feedstocks by *Clostridium saccharobutylicum* DSM 13864. *Bioresour Technol* **129**, 680–685.
- Nölling, J., Breton, G., Omelchenko, M. V., Makarova, K. S., Zeng, Q., Gibson, R., Lee, H. M., Dubois, J., Qiu, D. & other authors (2001). Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *J Bacteriol* **183**, 4823–4838.
- Obuchowski, M. (2005). [Serine-threonine protein phosphatases from *Bacillus subtilis*]. *Postepy Biochem* **51**, 95–104 (in Polish).
- Quratulain, S., Qadeer, M. A., Chaudhry, M. Y. & Kausar, A. R. (1995). Development and characterization of butanol-resistant strain of *Clostridium acetobutylicum* in molasses medium. *Folia Microbiol (Praha)* **40**, 467–471.
- Qureshi, N. & Blaschek, H. P. (2000). Butanol production using *Clostridium beijerinckii* BA101 hyper-butanol producing mutant strain and recovery by pervaporation. *Appl Biochem Biotechnol* **84–86**, 225–235.
- Qureshi, N., Saha, B. C. & Cotta, M. A. (2007). Butanol production from wheat straw hydrolysate using *Clostridium beijerinckii*. *Bioprocess Biosyst Eng* **30**, 419–427.
- Qureshi, N., Saha, B. C., Dien, B., Hector, R. E. & Cotta, M. A. (2010). Production of butanol (a biofuel) from agricultural residues: Part I – Use of barley straw hydrolysate. *Biomass Bioenergy* **34**, 559–565.
- Reid, S. J., Rafudeen, M. S. & Leat, N. G. (1999). The genes controlling sucrose utilization in *Clostridium beijerinckii* NCIMB 8052 constitute an operon. *Microbiology* **145**, 1461–1472.
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26.
- Schultz, A. C., Nygaard, P. & Saxild, H. H. (2001). Functional analysis of 14 genes that constitute the purine catabolic pathway in *Bacillus subtilis* and evidence for a novel regulon controlled by the PucR transcription activator. *J Bacteriol* **183**, 3293–3302.
- Soucaille, P., Joliff, G., Izard, A. & Goma, G. (1987). Butanol tolerance and autobacteriocin production by *Clostridium acetobutylicum*. *Curr Microbiol* **14**, 295–299.
- Stackebrandt, E. & Rainey, F. A. (1997). Phylogenetic relationships. In *The Clostridia: Molecular Biology and Pathogenesis*, pp. 3–19. Edited by I. R. Julian, A. M. Bruce, J. G. Songer, W. T. Richard & W. T. Richard. San Diego, CA: Academic Press.
- Tangney, M., Rousse, C., Yazdani, M. & Mitchell, W. J. (1998). Note: sucrose transport and metabolism in *Clostridium beijerinckii* NCIMB 8052. *J Appl Microbiol* **84**, 914–919.
- Terzaghi, B. E. & Sandine, W. E. (1975). Improved medium for lactic streptococci and their bacteriophages. *Appl Microbiol* **29**, 807–813.
- Treangen, T. J. & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36–46.
- Van Der Westhuizen, A., Jones, D. T. & Woods, D. R. (1982). Autolytic activity and butanol tolerance of *Clostridium acetobutylicum*. *Appl Environ Microbiol* **44**, 1277–1281.
- Wang, Y. & Blaschek, H. P. (2011). Optimization of butanol production from tropical maize stalk juice by fermentation with *Clostridium beijerinckii* NCIMB 8052. *Bioresour Technol* **102**, 9985–9990.
- Wilson, K. (1994). Preparation of genomic DNA from bacteria. In: *Current Protocols in Molecular Biology*, pp. 2.4.1–2.4.5. Edited by F. A. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith & K. Struhl. New York: Wiley.
- Wlodawer, A., Li, M., Gustchina, A., Oyama, H., Dunn, B. M. & Oda, K. (2003). Structural and enzymatic properties of the sedolisin family of serine-carboxyl peptidases. *Acta Biochim Pol* **50**, 81–102.
- Xi, H., Schneider, B. L. & Reitzer, L. (2000). Purine catabolism in *Escherichia coli* and function of xanthine dehydrogenase in purine salvage. *J Bacteriol* **182**, 5332–5341.

Xiao, H., Gu, Y., Ning, Y., Yang, Y., Mitchell, W. J., Jiang, W. & Yang, S. (2011). Confirmation and elimination of xylose metabolism bottlenecks in glucose phosphoenolpyruvate-dependent phosphotransferase system-deficient *Clostridium acetobutylicum* for simultaneous

utilization of glucose, xylose, and arabinose. *Appl Environ Microbiol* **77**, 7886–7895.

Edited by: S. Kengen