



ERNEST ORLANDO LAWRENCE BERKELEY NATIONAL LABORATORY

Title: Complete genome of the cellylytic thermophile *Acidothermus cellulolyticus* 11B provides insights into its ecophysiological and evolutionary adaptations

Author(s): R. Barabote^{1,†}, G. Xie¹, D. Leu², P. Normand³, A. Necsulea⁴, V. Daubin⁴, C. Médigue⁵, W. Adney⁶, X. Xu², A. Lapidus⁷, C. Detter¹, P. Pujic³, D. Bruce¹, C. Lavire³, J. Challacombe¹, T. Brettin¹ and Alison M. Berry².

Author Affiliations: ¹ DOE Joint Genome Institute, Bioscience Division, Los Alamos National Laboratory, ² Department of Plant Sciences, University of California, Davis, ³ Centre National de la Recherche Scientifique (CNRS), UMR5557, Écologie Microbienne, Université Lyon I, Villeurbanne, ⁴ Centre National de la Recherche Scientifique (CNRS), UMR5558, Laboratoire de Biométrie et Biologie Évolutive, Université Lyon I, Villeurbanne, ⁵ Centre National de la Recherche Scientifique (CNRS), UMR8030 and CEA/DSV/IG/Genoscope, Laboratoire de Génomique Comparative, ⁶ National Renewable Energy Laboratory

⁷ DOE Joint Genome Institute

Date: 06/10/09

Funding: This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

Complete genome of the cellulolytic thermophile *Acidothermus cellulolyticus* 11B provides insights into its ecophysiological and evolutionary adaptations.

Ravi D. Barabote^{1,†}, Gary Xie¹, David H. Leu², Philippe Normand³, Anamaria Necsulea⁴, Vincent Daubin⁴, Claudine Médigue⁵, William S. Adney⁶, Xin Clare Xu², Alla Lapidus⁷, Chris Detter¹, Petar Pujic³, David Bruce¹, Celine Lavire³, Jean F. Challacombe¹, Thomas S. Brettin¹, and Alison M. Berry².

¹ DOE Joint Genome Institute, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545 USA.

² Department of Plant Sciences, University of California, Davis, CA 95616, USA.

³ Centre National de la Recherche Scientifique (CNRS), UMR5557, Écologie Microbienne, Université Lyon I, Villeurbanne F-69622, France.

⁴ Centre National de la Recherche Scientifique (CNRS), UMR5558, Laboratoire de Biométrie et Biologie Évolutive, Université Lyon I, Villeurbanne, F-69622, France.

⁵ Centre National de la Recherche Scientifique (CNRS), UMR8030 and CEA/DSV/IG/Genoscope, Laboratoire de Génomique Comparative, 2, rue Gaston Crémieux, 91057 Evry Cedex

⁶ National Renewable Energy Laboratory, 1617 Cole Blvd., Golden, CO 80401, USA.

⁷ DOE Joint Genome Institute, Walnut Creek, CA 94598, USA.

† Current address: Department of Plant Sciences, University of California, Davis, CA 95616, USA.

Corresponding Author:

Alison M. Berry, Department of Plant Sciences, Mailstop 1, PES 1210, University of California, Davis, CA 95616. USA.

Tel: 530-752-7683; Fax: 530-752-4361

e-mail: amberry@ucdavis.edu

Running title: Complete genome sequence of *Acidothermus cellulolyticus* 11B.

Keywords: Actinomycete, actinobacteria, comparative genomics, glycoside hydrolases, genomic islands, flagellar genes.

ABSTRACT

We present here the complete 2.4 Mb genome of the cellulolytic actinobacterial thermophile, *Acidothermus cellulolyticus* 11B. New secreted glycoside hydrolases and carbohydrate esterases were identified in the genome, revealing a diverse biomass-degrading enzyme repertoire far greater than previously characterized, and significantly elevating the industrial value of this organism. A sizable fraction of these hydrolytic enzymes break down plant cell walls and the remaining either degrade components in fungal cell walls or metabolize storage carbohydrates such as glycogen and trehalose, implicating the relative importance of these different carbon sources. A novel feature of the *A. cellulolyticus* secreted cellulolytic and xylanolytic enzymes is that they are fused to multiple tandemly arranged carbohydrate binding modules (CBM), from families 2 and 3. Interestingly, CBM3 was found to be always N-terminal to CBM2, suggesting a functional constraint driving this organization. While the catalytic domains of these modular enzymes are either diverse or unrelated, the CBMs were found to be highly conserved in sequence and may suggest selective substrate-binding interactions. For the most part, thermophilic patterns in the genome and proteome of *A. cellulolyticus* were weak, which may be reflective of the recent evolutionary history of *A. cellulolyticus* since its divergence from its closest phylogenetic neighbor *Frankia*, a mesophilic plant endosymbiont and soil dweller. However, ribosomal proteins and non-coding RNAs (rRNA and tRNAs) in *A. cellulolyticus* showed thermophilic traits suggesting the importance of adaptation of cellular translational machinery to environmental temperature. Elevated occurrence of IVYWREL amino acids in *A. cellulolyticus* orthologs compared to mesophiles, and inverse preferences for G and A at the first and third codon positions also point to its ongoing thermoadaptation. Additional interesting features in the genome of this cellulolytic, hot-springs dwelling prokaryote include a low occurrence of pseudogenes or mobile genetic elements, an unexpected complement of flagellar genes, and presence of three laterally-acquired genomic islands of likely ecophysiological value.

Supplementary material is included as a separate PDF file.

INTRODUCTION

Efforts are underway worldwide to develop renewable energy sources as alternatives to fossil fuels. Microorganisms capable of breaking down lignocellulosic plant matter, a bioenergy source, are of enormous interest in the global quest to identify enzymes that can convert biomass into biofuels. *Acidothermus cellulolyticus* was first isolated in enrichment cultures from acidic hot springs in Yellowstone National Park, in a screen for microorganisms that carry out efficient cellulose degradation at high temperature (Mohagheghi et al., 1986). *A. cellulolyticus* 11B is acid-tolerant (pH 4-6, with optimal pH 5.5) and thermophilic (growth between 37° and 70°C; optimal growth temperature [OGT] is 55°C). It produces many thermostable cellulose-degrading enzymes (Adney et al., 1995; Baker et al., 1994; Ding et al., 2003; Tucker et al., 1989). One of the endoglucanases, E1, which has been crystallized, is highly thermostable to 81°C and has very high specific activity on carboxymethylcellulose (Sakon et al., 1996; Thomas et al., 1995). E1 has been expressed in several plants and shows promise for generating genetically improved feedstock for the production of affordable cellulosic ethanol (Sticklen, 2008). Hydrolytic enzymes from *A. cellulolyticus* have great potential in the biofuels industry because of their thermostability and activity at low pH (Rubin, 2008).

A. cellulolyticus is a member of the Frankineae, a high G+C, primarily Gram-positive Actinobacterial group (Rainey and Stackebrandt, 1993). All of the characterized strains of *A. cellulolyticus* are thermophilic and do not grow below 37°C (Mohagheghi et al., 1986). This makes the evolutionary context of *A. cellulolyticus* interesting, because its closest known phylogenetic neighbor is the mesophilic actinobacterium, *Frankia*, based on the analysis of the 16S rRNA, *recA*, and *shc* nucleotide sequences (Supplementary Fig. S1; Alloisio et al., 2005; Marechal et al., 2000; Normand et al., 1996). *Frankia* is a mesophilic (OGT 25-28°C), nitrogen-fixing soil organism that forms symbiotic root nodule associations with plants (Benson, 1988). The genetic distance between *A. cellulolyticus* and three *Frankia* strains, ACN14a, CcI3 and EAN1pec, is very small and comparable to that found between certain strains within the *Frankia* species. Thus, although *Acidothermus* and *Frankia* share a close phylogenetic relationship at the DNA sequence level, they have evolved to live in dramatically diverse

environments over the last 200-250 million years since their last common ancestor (Normand et al., 2007). Complete genome sequences of three *Frankia* strains, ACN14a, CcI3 and EAN1pec, as well as those of other close relatives of *A. cellulolyticus* are now available, including the mesophilic *Streptomyces avermitilis*, *Streptomyces coelicolor*, and the terrestrial thermophilic *Thermobifida fusca* (Bentley et al., 2002; Ikeda et al., 2003; Lykidis et al., 2007; Normand et al., 2007; Omura et al., 2001). Genomic comparison of *A. cellulolyticus* with the mesophilic as well as thermophilic actinobacteria could provide insight into the nature of adaptation of this aquatic thermophile, and add to our understanding of evolution within the actinobacteria.

We present analysis of the complete genome of *Acidothermus cellulolyticus* 11B (ATCC 43068; Genbank accession NC_008578). Insights into the biomass degradation capabilities of the organism as well as thermophilic features of its genome and proteome are discussed. In addition, we discuss three laterally acquired genomic islands with genes of likely ecophysiological value, as well as the unexpected presence of flagellar genes in the genome.

RESULTS

General Genome Characteristics. The 2.44 Mb genome of *A. cellulolyticus* is encoded on a single circular chromosome (Fig. 1) and is approximately 66.9% G+C-rich. The G+C content of the non-coding region (68.41%) is higher than the G+C content of the coding region (66.76%). The total GC-skew analysis revealed a potential origin of replication (OriC) upstream of the *dnaA* gene and a terminus at approximately 1.2 Mb from the origin. A single *rrn* operon containing the genes for the 16S, 23S, and 5S rRNAs is located towards the replication terminus, an unusual position. Forty-five tRNAs representing 43 different anticodons are encoded in the genome (Supplementary Table S1, and Supplementary Text). The *A. cellulolyticus* genome contains only four annotated pseudogenes (Acel_0124, Acel_0186, Acel_0477, Acel_1066) that do not encode any protein products. The protein coding sequence constitutes approximately 90% of the genome and encodes 2157 predicted proteins. No identifiable prophages or phage-related proteins were found in the genome and only two genes encoding fragments of a single

transposase (Acel_1666, Acel_1667) were found in the genome. One-fifth of all the predicted proteins have no decipherable function. Approximately 8% of the proteins (171 proteins) do not show sequence similarity to any sequences in the NCBI database and thus appear to be ORFans unique to *A. cellulolyticus* (Supplementary Fig. S2). Analysis of the phyletic distribution of BLAST hits of the remaining proteins revealed that approximately 80% of the *A. cellulolyticus* proteins show highest sequence similarity to proteins from other actinobacteria (Supplementary Fig. S2). Within the actinobacterial hits, the highest number of best BLAST hits, surprisingly, were to the phylogenetically more remote *Streptomyces* spp. (~18%), more so than to its closest phylogenetic neighbor *Frankia* spp. (~17%), and followed by *T. fusca* (~13%). Interestingly, 18 *A. cellulolyticus* proteins bear highest sequence similarity to archaeal proteins and 7 proteins show highest sequence similarity to eukaryotic proteins (Supplementary Table S2).

Based on the distribution of the top BLAST hits to *Frankia*, *Streptomyces* and *T. fusca*, sequenced genomes of these organisms were used for comparative genome analyses. An overview of the *A. cellulolyticus* genome features in comparison with the genomes of *Frankia*, *Streptomyces* and *T. fusca* is provided in Table 1.

Carbohydrate active enzymes. The genome of *A. cellulolyticus* contains at least forty-three genes encoding 35 glycoside hydrolase (GH) and 8 carbohydrate esterase (CE) enzymes (Table 2). Of these, 28 predicted enzymes break down structural or storage carbohydrates found in plant and fungal cells, including cellulose, xylan, starch and chitin. The GHs belong to 17 families while the CEs span 5 families as per the CAZy database (Henrissat, 1991; Coutinho and Henrissat, 1999; <http://www.cazy.org/>). At least 15 GHs belonging to families 1, 3, 5, 6, 9, 10, 12, 16, 48, and 74 and three CEs from families 1 and 7 may be important for plant biomass deconstruction in *A. cellulolyticus*. Two or more representatives of several of these enzyme families occur in the genome, except for GH 1, 16, 48 and 74; and CE 7 (Table 2).

Five previously described carbohydrate active enzymes (Ding et al., 2003) could be correctly mapped in the genome (Table 2). While these known cellulolytic enzymes are encoded in a large gene cluster (Ding et al., 2003), genes encoding many newly identified enzymes occur scattered throughout the genome (Fig. 1). The genome revealed six new cellulose-degrading enzymes including 4 endoglucanases and 2 beta-glucosidases. In addition, six enzymes for hemicellulose decomposition were identified including 2 xylanases, 3 xylan esterases, and a xylosidase. Except for the GH1 beta-glucosidase and the GH3 xylosidase that are predicted to be cytoplasmic as well as the CE7 esterase, the rest of the plant cell-wall-degrading enzymes are either predicted to be secreted or contain a signal peptide (Table 2).

In addition to the 17 plant cell-wall-degrading enzymes, the genome encodes 10 proteins potentially associated with the breakdown of fungal cell wall components. Two beta-N-acetylhexosaminidases and a chitooligosaccharide deacetylase were predicted to be cytoplasmic while the other seven proteins are either predicted to be secreted or have a signal sequence indicating that they are likely to be secreted. These include 4 chitinases, an N-acetylglucosaminidase, a GH16 endo-1,3-beta-glucanase, and a CenC domain containing putative chitin binding protein.

Sixteen enzymes are involved in either glycogen and trehalose biosynthesis and degradation (8 enzymes) or related cellular metabolic functions (Table 2). The GH13 alpha amylase (Acel_0679) may additionally participate in starch metabolism. None of these enzymes contain a signal sequence and are predicted to be cytoplasmic except the two GH23 lytic transglycosylases that may be cell-wall associated.

Carbohydrate Binding Modules (CBMs). Catalytic domains of two-thirds of the 21 secreted biomass-degrading enzymes in *A. cellulolyticus* were found fused to one or more CBM types (Table 2). Furthermore, members of the same GH families carry varying numbers and combinations of fused CBMs. Only one of the esterases (CE1) was fused to CBMs. The cellulose- and xylan-degrading *A. cellulolyticus* enzymes contain c-terminally fused CBM2 domains, a feature that was found to be similar to other actinobacterial homologs. However, many *A. cellulolyticus* enzymes additionally contain CBM3 domains. Curiously, CBM3 was always found to occur N-terminal relative to CBM2, but never C-

terminal to it. In general, the two CBM types were found to occur in tandem (as X-CBM3-CBM2, where X is GH, CE, or CBM3 domain), except in the case of the Gux1 exoglucanase and the GuxA cellulase where the two CBMs are separated by a GH domain (CBM3-X-CBM2). Although two endoglucanases, the previously characterized endoglucanase E1 (GH5) and a newly identified GH12 endoglucanase, contain just the CBM2, no enzymes with only the CBM3 module occur in the genome.

Overall, the *A. cellulolyticus* genome encodes about equal numbers of the two CBM types - 10 CBM2 and 9 CBM3 modules. Comparative genome analysis revealed that *Frankia alni* ACN14a and CcI3 lack either CBMs, while a single CBM2 fused to a chitinase was found in *Frankia* sp. EAN1pec. However, the three *Frankia* genomes also lack cellulolytic enzymes. The genomes of two close actinobacterial relatives with multiple cellulolytic enzymes, *Streptomyces* and *Thermobifida*, contain 11-14 CBM2 modules but just 1-2 CBM3 modules. In contrast, the genome of the anaerobic cellulosome-forming bacterium *Clostridium thermocellum* encodes about 24 CBM3 domains but no CBM2 homologs. Analysis of each of the two CBM types revealed that the sequences are highly conserved in *A. cellulolyticus*. In contrast, the different CBM2 domains in *Streptomyces* or *Thermobifida*, or the several CBM3 domains in *C. thermocellum*, exhibit sequence diversity.

In addition to the two CBM families, a single copy of CBM6 was found attached to a GH16 endo-1,3-beta-glucanase. Three of the secreted chitinases also contained CBM5 and/or CBM16 domains. A few of the cytoplasmic enzymes involved in glycogen/trehalose metabolism contain 1-2 CBM48 modules.

Genomic Islands. Three major genomic islands (GIs) with significantly lower G+C and deviant dinucleotide signature were identified (Fig. 2). Several proteins encoded in these islands have no recognizable orthologs in close relatives of *A. cellulolyticus*. *GI1* consists of 15 genes with an average G+C of 58% (Table 3). The first five genes likely constitute an operon that encodes fumarate reductase/succinate dehydrogenase, arylalkylphosphatase, a short-chain dehydrogenase, deoxyribose-

phosphate aldolase and a ROK-family protein, respectively. The second half of GI1 contains genes involved in sugar uptake and metabolism.

GI2 contains 18 genes (average G+C of 62.5%) flanked by tRNA genes (Table 3). Half of the genes do not have a recognizable function, while many of the remaining genes encode putative homologs of the *vrl* locus of *Dichelobacter nodosus*. The VrlI and J homologs in *A. cellulolyticus* have DNA-binding and ATPase domains, respectively, and the VrlK, P, and Q homologs do not have any identifiable domains. With respect to the four intervening proteins, one is a transcriptional regulator containing a helix-turn-helix motif, another shows weak homology to DNA methylases, a third is a hypothetical protein, and the fourth has a helicase domain and could be a VrlO homolog although the homology is undetectable at sequence level. Most proteins encoded in this island show highest similarity to proteins from low G+C Gram-positives, namely *Bacteroides*, *Nitrosococcus*, and *Thermoanaerobacter*.

GI3 carries 31 genes (average G+C of 61.7%) and is flanked by tRNA^{Arg} gene upstream and by tRNA^{His} gene downstream (Table 3). One-third of the proteins encoded on this island have no recognizable function. Of the remaining genes, 3 encode proteins involved in ABC transport, 2 of which may be involved in the uptake of amino acids. *Acel_1633 - Acel_1639* form an operon of 7 genes: the first two genes encode proteins with unknown function; the third and the last encode enzymes involved in amino acid metabolism; while the rest encode subunits of the carbon monoxide (CO) dehydrogenase family proteins. Another likely operon of 4 genes encodes an aldehyde oxidase, a coenzyme A transferase, glutamate coA-transferase and a luciferase family protein. Six genes in this GI (namely *Acel_1626*, *Acel_1628*, *Acel_1634*, *Acel_1639*, *Acel_1643*, and *Acel_1644*) encode proteins that bear highest sequence similarity to proteins from thermophilic bacteria and archaea. With the exception of *Acel_1626*, homologs of these six proteins do not occur in *Frankia*.

In addition to the three major islands, twenty-one smaller genomic regions (GR) were identified. Characteristics of the predicted regions are detailed in Supplementary Table S3.

Flagella and Motility. Mohagheghi et al (1986) reported that *A. cellulolyticus* cells were non-motile based on microscopic observations. Surprisingly, immediately downstream of GI2, we identified a stretch of 37 genes (Acel_0828 – Acel_0864) that did not have any homologs in *Frankia*, *Streptomyces* or *T. fusca*. This region encoded a complete set of genes coding for flagellar biosynthesis and motility. The genes are organized into two divergent gene clusters (Fig. 3). Most of the flagellar structural genes are organized in the larger cluster containing 31 genes on the leading strand. The regulatory gene *csrA*, recently shown to encode a regulator of flagellar biosynthesis (Yakhnin et al., 2007), is encoded by the last gene in the smaller cluster containing five genes. Thus far, only three other actinomycetes, *Nocardioides* sp. JS514, *Kineococcus radiotolerans*, and *Leifsonia xyli*, encode sequence homologs of the flagellar genes (Fig. 3). The gene content and order of the flagellar operon is highly conserved between *A. cellulolyticus* and *Nocardioides*, while minor differences in gene order are observed in *Kineococcus*. Several flagellar genes in *L. xyli* are pseudogenes, in agreement with the observation that the organism is non-motile and does not produce a flagellum (Monteiro-Vitorello et al., 2004); the presence of motility or flagella have not been well studied in the other two organisms. Although in the original study no motility was observed in *A. cellulolyticus* (Mohagheghi et al., 1986), the possibility of motility, perhaps under specific growth conditions, is being carefully re-examined.

Thermoadaptation. Principal component analysis (PCA) of global as well as synonymous codon usage revealed that *A. cellulolyticus*, surprisingly, did not contain patterns typically observed in thermophilic prokaryotes (Supplementary Figs. S3A and S3B). It was clearly positioned amidst mesophiles along the PC2 axis that correlated with OGT. Codon usage differences between *Acidothermus* and *Frankia* were very subtle (Supplementary Table S4). Differences in the codon usage of the six actinobacteria compared in our study did not always follow differences in G+C content in the coding region of their genomes (Supplementary Table S4), suggesting a physiological pressure influencing these differences. A detailed comparison of the relative abundances of the four nucleotides at each of the three codon positions showed that the relative proportion of G was higher and that of A was lower at the first codon position in the two

thermophiles as compared to the four mesophiles (Table 4). In addition, an opposite but slightly weaker trend was observed at the third codon position, i.e., the relative proportion of A was higher and that of G was lower in the two thermophiles as compared to the mesophiles (Table 4). Interesting differences were observed for the GNA and ANG codons (see Supplementary Table S4). Of the four GNA codons, the GAA codon (for glutamate) showed most prominent increase in the two thermophiles. Of the four ANG codons, the AGG codon (for arginine) was clearly less preferred in *A. cellulolyticus* and *T. fusca*.

Non-coding RNAs, ribosomal RNAs (rRNAs) and transfer RNA (tRNAs), in *A. cellulolyticus* had a higher G+C content than mesophilic species with similar genomic G+C (Fig. 4). Confidence intervals of the prediction of a linear model (RNA G+C content as a function of Genomic G+C content) for mesophilic species showed that *A. cellulolyticus* was clearly an outlier when compared to the mesophilic species in the study. The G+C content of functional RNAs has been shown to correlate positively with OGT (Galtier and Lobry, 1997).

Similar to the codon-usage PCA results (Supplementary Figs. S3A and S3B), PCA of the amino acids usage did not reveal thermophilic trends in the *A. cellulolyticus* proteome (Supplementary Fig. S4). Contrary to our expectation that it should segregate with other thermophiles, *A. cellulolyticus* was positioned near mesophiles along the PC2 axis that correlated with OGT. However, in a more detailed analysis of the amino-acid composition of ribosomal proteins, *A. cellulolyticus* was placed nearer to the thermophiles than *Frankia* or *Streptomyces*, and was at the same level as *T. fusca* (Fig. 5).

The total fractions of IVYWREL amino acids in *A. cellulolyticus* proteome and cytosolic sub-proteome were higher than those in *Frankia* sp. and *Streptomyces* sp. (Supplementary Table S5). Further, analysis of the amino acid composition of 478 conserved orthologous proteins in these six actinobacteria clearly revealed that both *A. cellulolyticus* and *T. fusca* orthologs contain a higher proportion of IVYWREL amino acids compared to the four mesophilic organisms (Supplementary Table S5). The values of IVYWREL fractions in the orthologs showed even greater linear correlation with OGT than those from the cytosolic subproteomes or whole proteomes. In addition, an extended analysis of 46 conserved orthologous proteins from several mesophilic and thermophilic actinobacteria with varying

G+C content showed a similar trend, namely that orthologs from the thermophilic actinobacteria contain increased representation of IVYWREL amino acids compared to the mesophiles (Supplementary Table S6). It is to be noted that there are exceptions to a strict increase in IVYWREL with OGT. Thus, the content of IVYWREL is a reasonable but not a perfect predictor of the OGT, as noted also by Zeldovitch et al (2007).

DISCUSSION

A. cellulolyticus has a small genome with very few pseudogenes or mobile genetic elements. The two transposase-encoding gene sequences in *A. cellulolyticus* encode frame-shifted fragments of an intact gene that is found in *Frankia* and other actinobacteria. As a result, *A. cellulolyticus* may not encode an active transposase. By contrast, many of the terrestrial as well as aquatic actinobacterial relatives of *A. cellulolyticus*, such as *Frankia* sp., *S. avermitilis*, *S. coelicolor*, and *T. fusca* (see Table 1) as well as *K. radiotolerans*, and *Nocardioide*s sp. (data not shown) possess multiple pseudogenes, as well as several transposase-encoding genes and IS elements in their genomes. With the exception of *T. fusca*, the other actinobacteria also possess large genomes, ranging from 5 to 9 Mb. It is conceivable that the presence and abundance of transposase-related genes in the larger genomes reflect the role of these mobile elements in their genome expansion, as described for *Frankia* (Normand et al., 2007), but also that genome reduction events accompanied by the loss of mobile elements may have resulted in a small genome size of *A. cellulolyticus*.

With the renewed interest and growing quest for microbes that efficiently deconstruct plant cell wall carbohydrates for conversion to biofuels, the sequenced genome of *A. cellulolyticus* offers substantial potential for the discovery of valuable thermostable enzymes. In addition to five previously described cellulolytic enzymes, the *A. cellulolyticus* genome revealed many additional possibilities for biomass degradation. The *A. cellulolyticus* genome encodes genes for several enzymes that break down cellulose and xylans, while the absence of pectin degradation genes corroborates the reported lack of growth on pectin (Mogagheghi et al., 1986). The organism devotes about equal numbers of enzymes to

the breakdown of cellulose (10 genes) and xylan (7 genes) in the plant cell wall, as well as chitin and other components in fungal cell walls (10 genes), and the metabolism of storage carbohydrates such as glycogen and trehalose (8 genes). This suggests that all these carbon sources are of comparable importance to the organism.

Complete enzymatic digestion of cellulose requires three types of glycosyl hydrolases, including cellulases (endoglucanases), cellobiohydrolases (exoglucanases), and cellobiosidases (beta-glucosidases). All three are present in multiple copies in the *A. cellulolyticus* genome. Specifically, there are 6 endoglucanases, 2 exoglucanases, and 2 beta-glucosidases. Efficient hydrolysis of crystalline cellulose requires the presence of at least one endoglucanase and two types of exoglucanases. The *Acidothermus* genome contains both a reducing-end specific GH48 exoglucanase and a non-reducing end specific GH6 exocellulase (Ding et al., 2003; Adney et al., 2003).

Based on sequence similarity of the *A. cellulolyticus* Acel_0129 protein to a characterized endo-1,3-beta-glucanase from *S. sioyaensis*, we predict that this protein binds to and hydrolyzes 1,3-β-D-glucan, a major constituent of fungal cell walls and laminarins of certain algal groups and diatoms (Hong et al., 2002). This enzyme likely helps the organism assimilate fungal cell walls as a food source. The functions of four putative chitinases remain to be confirmed experimentally. The capability to degrade chitin could permit degradation of fungal and insect biomass. After cellulose, chitin is the second most abundant structural cell wall polymer in nature. Unlike other eukaryotic cell-wall biopolymers, chitin contains nitrogen and hence could be used as a carbon and nitrogen source. Decaying plant matter as well as dead insects that fall into the thermal pools may provide sources of chitin and 1,3-β-D-glucan. The ability to utilize a range of carbon sources could offer a survival edge under limiting nutritional conditions in the thermal pool. Chitinases have received increased attention recently due to their wide applications in agricultural, medical and food industry. The potential for a source of thermostable chitinases elevates the industrial importance of *A. cellulolyticus* beyond its anticipated applications in cellulosic biofuel technologies.

The fact that secreted plant biomass degrading enzymes in *A. cellulolyticus* contain two different types of CBM domains, from families 2 and 3, is interesting functionally as well as evolutionarily. Only ten complete bacterial genomes, including *A. cellulolyticus*, encode both CBM types, of which six are Actinobacteria and one a Firmicute (<http://www.cazy.org>). This relatively low frequency suggests that the co-existence of both types of CBM domains is rare. Among these 10 genomes there is a clear preference for either CBM2 (in Actinobacteria) or CBM3 (in the Firmicutes) but not for both. The *A. cellulolyticus* genome with equal proportions of the two CBM types is clearly an exception to the pattern to date. The co-existence of CBM2 and CBM3 domains in a majority of the *A. cellulolyticus* modular enzymes as well as their restricted organization may suggest functional and/or thermostability constraints. It is possible that the presence of CBM3 alone or its location C-terminal to CBM2 may either destabilize the protein or affect the optimal activity of *A. cellulolyticus* enzymes. The high degree of sequence conservation within the two CBM families in *A. cellulolyticus* suggests duplication of each of these domains after speciation. Fusion of these duplicated domains to the GHs could indicate a selective pressure for localizing the secreted GHs on specific substrates. Both CBM2 and CBM3 bind predominantly to cellulose with experimental evidence for binding to chitin in a few cases (Boraston et al., 2004). A few CBM2 members have also been observed to bind xylan (Boraston et al., 2004). Whether the two families of CBM domains in *A. cellulolyticus* bind cellulose, xylan or chitin, or multiple substrates, remains to be determined functionally.

The *A. cellulolyticus* genome revealed three laterally acquired GIs characterized by a lower G+C content and a deviation from the genomic signature. Regions that deviate significantly from the genomic signature are thought to have been laterally transferred (Karlin, 2001). In addition, the fact that the three islands are either flanked by tRNA genes and/or lack homologs in other actinobacteria strongly suggests that these DNA regions have been horizontally acquired in *A. cellulolyticus*. Several genes in these islands show highest sequence similarity to proteins from thermophilic organisms. Analysis of the genes encoded within the three GIs suggests a functional role for the acquired genes in the context of the organism's ecology. Aryldialkyl phosphatase (encoded on GI1) catalyzes the hydrolysis of an aryl-

dialkyl phosphate to form dialkyl phosphate and an aryl alcohol. In cellulolytic fungi aryl-alcohol dehydrogenase activity has been implicated in lignolysis (Reiser et al., 1994). GI2 carries homologs of the *vrl* genes found preferentially associated with more virulent isolates of *D. nodosus*, and which are proposed to have been acquired horizontally possibly from a bacteriophage or a plasmid (Billington et al., 1999). Although the precise function of the *vrl* locus is unclear, many of these genes could be involved in DNA restriction and modification, offering immunity to *A. cellulolyticus* against phage infection, similar to the *S. coelicolor* phage resistance Pgl system (Sumby and Smith, 2002), which bears sequence similarity to the Vrl proteins. GI3 contains genes that may be involved in amino acid transport and metabolism as well as genes for three subunits of the CO dehydrogenase family. Homologs also occur in other actinobacteria such as *Arthrobacter* and *Mycobacteria* that have been shown to grow chemolithotrophically on CO as the sole carbon and energy source under aerobic conditions (Meyer and Schlegel, 1983; Park et al., 2003), suggesting a similar potential may be present in *A. cellulolyticus*. Since CO dehydrogenases share high sequence similarity with xanthine dehydrogenases, it is difficult to predict whether the *A. cellulolyticus* homologs function in carbon fixation or in purine salvage. However, either of these possibilities would add eco-physiological value for *A. cellulolyticus*.

Thermophilic adaptations have not been systematically examined within the actinobacteria, an ecologically-diverse yet relatively under-studied bacterial group. *A. cellulolyticus* grows optimally at 55°C, while most of its closest phylogenetic relatives are mesophilic. The use of PCA, or the similar technique, correspondence analysis (CA), to study the genomes of hyperthermophilic, thermophilic and mesophilic prokaryotes has facilitated identification of their thermoadaptation characteristics (Lynn et al., 2002, Kreil and Ouzounis, 2001; Singer and Hickey, 2003, Suhre and Claverie, 2003). Contrary to our expectations based on these previous studies, in our PCA results, neither the genome nor the proteome of *A. cellulolyticus* segregate with other thermophiles. The degree of separation along PC2 axis that correlates with OGT may suggest how recently a thermophile has evolved. In that case, the lack of unambiguous separation of *A. cellulolyticus* from mesophiles along PC2 could reflect the relatively short history of *A. cellulolyticus* in thermal pools, as its genome and proteome still show meso-thermophilic

features. This pattern suggests a recent and ongoing adaptation to the thermophilic environment.

Alternatively, *A. cellulolyticus* may have evolved unique mechanisms of thermotolerance.

The subtle increase in the G and A nucleotides at the first and third codon positions, respectively, in the *A. cellulolyticus* genes could enhance thermostability of its mRNAs by probabilistically increasing the frequency of AG dinucleotides in its mRNAs, by a plausible increase in the frequency of NNA-GNN di-codons. The ApG dinucleotides are thought to stabilize DNA due to their low stacking energy and have been observed to occur at higher frequency in (hyper)thermophilic organisms compared to mesophiles (Zeldovich et al., 2007). The relatively lower frequency of AGG codons in *A. cellulolyticus* may in turn be due to the inverse purine preferences at the first and third codon positions and may explain the lack of separation of *A. cellulolyticus* from the mesophiles, along PC2 in our PCA (see Supplementary Figs. S3A and S3B). The AGG codon is known to strongly influence the separation between thermophiles and mesophiles (Lynn et al., 2002; Singer and Hickey, 2003). *A. cellulolyticus* is clearly an exception in the use of AGG codons compared to other thermophiles.

The *A. cellulolyticus* proteome contained elevated fraction of IVYWREL amino acids compared to both *Frankia* sp. and *Streptomyces* sp. A recently identified positive correlation between the total fraction of 7 amino acids (Ile, Val, Tyr, Trp, Arg, Glu, Leu) in prokaryotic proteomes and the OGT of the organisms is another measure for thermoadaptation (Zeldovich et al., 2007). Usage patterns of either the 20 individual amino acids (as studied using PCA) or the total fraction of IVYWREL amino acids likely represent alternative yet overlapping thermophilic signatures. This is because most hyperthermophiles and thermophiles separate well along the OGT axis in PCA and also contain relatively elevated content of IVYWREL residues in their proteomes. Interestingly, *A. cellulolyticus* appears to show the latter but not the former thermophilic signature. It is possible that the elevated IVYWREL content in the proteome represents an overarching adaptation to thermophily and that usages of individual amino acids get fine-tuned with evolutionary time. The higher IVYWREL content in conserved *A. cellulolyticus* proteins compared to their orthologs in mesophilic actinobacteria rules out the possibility that the differences in IVYWREL residues in the proteome and cytosolic subproteome are due to few proteins with skewed

amino acid composition. This suggests that this biased amino acid usage in *A. cellulolyticus* proteome may be reflective of its adaptation to the thermal environment. It is worth noting that there have been no findings of proteins unique to thermophiles that explain organismal adaptations to high temperature, and that proteins in thermophiles show biased amino acid compositions compared to orthologs in mesophiles (Takami et al., 2004).

Adaptation to thermophily is likely to be a slow and continuous process. Although the overall *A. cellulolyticus* proteome revealed no clear thermophilic tendency, a more detailed analysis revealed a preference for thermophilic amino acid usage in its ribosomal proteins. These results taken together with the fact that ribosomal proteins are essential for cellular viability, and that ribosomal RNAs and transfer RNAs in *A. cellulolyticus* contain distinct thermophilic features, suggest that evolution of a thermotolerant protein translation machinery may be an important early step in thermoadaptation. It has been reported that three characterized strains of *A. cellulolyticus* have different OGT (Mohagheghi et al., 1986). Conceivably, other strains of *A. cellulolyticus* that span a range of either lower or higher OGT exist in nature. Perhaps, the isolation of such strains in the future and availability of genome sequence from multiple *A. cellulolyticus* strains may shed further light on genomic evolutionary processes for thermophilic adaptation.

METHODS

Strains, Culture, and DNA Extraction: *Acidothermus cellulolyticus* 11B was grown at University of California, Davis, from DMSO stocks maintained and provided by National Renewable Energy Laboratory (NREL), Golden, CO, derived from the original isolate of Mohagheghi et al (1986). Cells were grown in shaking or rolling liquid cultures at 55°C, in LPBM medium (Mohagheghi et al., 1986; also called ATCC medium 1473), pH 5.5, modified such that the carbon source was 0.25 g/l cellobiose + 0.25 g/l glucose, without cellulose. For isolation of high-molecular-weight genomic DNA from *A. cellulolyticus*, a protocol was devised to reduce the extensive nuclease activity: cell pellets were suspended in 200 µl lysis buffer (50 mM Tris-HCl, pH 8.0, 10 mM EDTA, preheated to 37°C) with 10 µl

lysozyme (100 mg/ml, MP Biomedicals), and incubated at 37°C for 2 hours; 1200 µl of ATL solution (Qiagen) plus 200 µl of protease K (10 mg/ml, Qiagen) were added, followed by incubation at 55°C for 2.5 hours. The supernatant was extracted with phenol-chloroform and chloroform, and DNA was precipitated, air-dried and resuspended as in Sambrook et al (1989). Genomic DNA was stored at -20°C in the presence of 0.1 mg/ml RNaseI (Promega) and its integrity was verified on 0.5% agarose gel.

Sequencing, Gene Prediction, and Annotation. The *A. cellulolyticus* 11B genome (NCBI Record: NC_008578) was sequenced and annotated by the Joint Genomes Institute, U.S. Department of Energy. Large (40 kb), medium (8 kb) and small (3 kb) insert DNA libraries were sequenced using the random shotgun method with average success rate of 96% and average high-quality read lengths of 685 nucleotides. After the shotgun stage, reads were assembled with parallel phrap (High Performance Software, LLC). Possible mis-assemblies were corrected with Dupfinisher (unpublished, C. Han) or transposon bomb of bridging clones (EZ-Tn5 <P6Kyori/KAN-2> Tnp Transposome kit, Epicentre Biotechnologies). Gaps between the contigs were closed by editing, custom primer walks or PCR amplification. The completed genome sequence of *A. cellulolyticus* contains 59147 reads, achieving an average of 18-fold sequence coverage per base with error rate less than 1 in 100,000. Automated annotation steps were performed as described previously (Chain et al., 2003).

Data Acquisition. Genome sequence files, executable BLAST (Altschul et al., 1997) programs, and the ‘nr’ database were obtained from NCBI ftp site. In order to build a comprehensive dataset spanning the entire known range of OGTs for our PCA analyses, we extracted all complete prokaryotic genome sequences available in the NCBI genome database, without making any *a priori* choice of the species to be included in our analyses. OGT information was extracted from the American Tissue Culture Collection (ATCC) and the German Collection of Microorganisms and Cell Cultures (DSMZ). Organisms with unknown OGT were removed, and our final dataset contained 409 prokaryotes (Supplementary Table S7), including 17 hyperthermophilic species (OGT greater than or equal to 80°C),

19 thermophilic species (OGT between 55°C and 80°C), 369 mesophiles (OGT between 20°C and 55°C), 4 psychrophiles (OGT less than 20°C).

To extract ribosomal proteins, we scanned the annotations of the complete genomes listed in the NCBI ftp sites for the following terms: "ribosomal", "50S", "30S", "SSU", or "LSU". We then manually checked the annotations retrieved with this method, and we removed hits that did not correspond to ribosomal proteins *per se* (for example, "ribosomal large subunit pseudouridine synthase D").

Sequence Analyses. The %G+C of the genome and the non-coding RNAs was calculated from nucleotide sequences in the respective NCBI files. Short perl codes were written and utilized for specific computational tasks, such as for calculating G+C in DNA and RNA sequences, amino acid composition of proteins, codon usage, etc. The total fraction of IVYWREL residues was calculated by combining the fractions of the seven individual amino acids. The relative proportions of each nucleotide at each codon position were calculated from the codon usage tables. Genomic signature was calculated as described by Karlin (2001). The organization of flagellar genes in the different actinobacteria was obtained using the tools available on the Integrated Microbial Genomics (IMG) server (<http://img.jgi.doe.gov/cgi-bin/pub/main.cgi>; Markowitz et al., 2006).

All *A. cellulolyticus* proteins were searched against the nr database using the standalone blastp program and the distribution of organisms with the best hit was calculated from the BLAST results. Bi-directional top BLAST hits were used to identify the 478 conserved proteins (Supplementary Tables S8) in six organisms listed in Table 1. Similarly, 46 orthologous proteins (Supplementary Tables S9) were identified common to 45 completely sequenced actinobacteria.

Principal Component Analysis (PCA). The amino acid compositions of ribosomal proteins from 409 prokaryotes with known OGTs were subjected to PCA using the R statistical software (<http://www.r-project.org/>). Global and synonymous codon usage in the genomes, and amino acid usage in the whole proteomes of the 409 prokaryotes were also analyzed using PCA (see Supplementary Materials). All

statistical analyses were performed using the inbuilt functions in the R package (<http://www.r-project.org/>).

ACKNOWLEDGMENTS

This work was supported by a Microbial Sequencing Project, U.S. Department of Energy, proposed by AMB, and Experiment Station Project CA-D*-PLS-7688-H (AMB). We would like to thank Dr. Charlie Strauss and Dr. Chris Stubben at the Los Alamos National Laboratory for help with PCA and R software, respectively.

REFERENCES

- Adney, W.S., Tucker, M.P., Nieves, R.A., Thomas, S.R., and Himmel, M. E. 1995. Low molecular weight thermostable β -D-glucosidase from *Acidothermus cellulolyticus*. *Biotechnology Letters***17**: 49-54.
- Alloisio N., Marechal J., Heuvel B.V., Normand P., and Berry, A.M. 2005. Characterization of a gene locus containing squalene-hopene cyclase (*shc*) in *Frankia alni* ACN14a, and an *shc* homolog in *Acidothermus cellulolyticus*. *Symbiosis***39**: 83-90.
- Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.***25**: 3389-3402.
- Baker, J.O., Adney, W.S., Nieves, R.A., Thomas, S.R., Himmel, M.E., and Wilson, D.B. 1994. A new thermostable endoglucanase, *Acidothermus cellulolyticus* E1. *Appl. Biochem. Biotechnol.* **45-46**: 245-256.
- Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.***340**: 783-795.
- Benson, D.R. 1988. The genus *Frankia*: actinomycete symbionts of plants. *Microbiol Sci.* **5**: 9-12.

- Bentley, S.D., Chater, K.F., Cerdeno-Tarraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., et al. 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**: 141-147.
- Billington, S.J., Huggins, A.S., Johanesen, P.A., Crellin, P.K., Cheung, J.K., Katz, M.E., Wright, C.L., Haring, V., and Rood, J.I. 1999. Complete nucleotide sequence of the 27-kilobase virulence related locus (*vrl*) of *Dichelobacter nodosus*: evidence for extrachromosomal origin. *Infect Immun.* 67: 1277-1286.
- Boraston, A.B., Bolam, D.N., Gilbert, H.J., and Davies, G.J. 2004. Carbohydrate-binding modules: fine tuning polysaccharide recognition. *Biochem. J.* **382**: 769-81
- Chain, P., Lamerdin, J., Larimer, F., Regala, W., Lao, V., Land, M., Hauser, L., Hooper, A., Klotz, M., Norton, J., et al. 2003. Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosomonas europaea*. *J. Bacteriol.* **185**: 2759-2773.
- Coutinho, P.M. and Henrissat, B. 1999. Carbohydrate-active enzymes: an integrated database approach. In *Recent advances in carbohydrate bioengineering* (eds. H.J. Gilbert, G. Davies, B. Henrissat and B. Svensson), pp. 3-12. The Royal Society of Chemistry, Cambridge.
- Ding, S-Y, Adney, W.S., Vinzant, T.B., Decker, S.R., Baker, J.O., Thomas, S.R., and Himmel, M.E. Glycoside Hydrolase Gene Cluster of *Acidothermus cellulolytics*, In *Applications of Enzymes to Lignocellulosics*; Mansfield, S. D. and Saddler, J. N.; Eds ACS Symposium Series 855, American Chemical Society, Washington, D.C., 2003, pp. 332-360.
- Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783-791.
- Galtier, N., and Lobry, J.R. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol.* **44**: 632-636.
- Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M., and Brinkman, F.S.L. 2005. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis, *Bioinformatics* **21**: 617-623

- Henrissat, B. 1991. A classification of glycosyl hydrolases based on amino-acid sequence similarities. *Biochem. J.* **280**: 309-316.
- Hong, T.Y., Cheng, C.W., Huang, J.W., Meng, M. 2002. Isolation and biochemical characterization of an endo-1,3-beta-glucanase from *Streptomyces sioyaensis* containing a C-terminal family 6 carbohydrate-binding module that binds to 1,3-beta-glucan. *Microbiology.* 148:1151-1159.
- Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M., and Omura, S. 2003. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat Biotechnol.* **21**: 526-531.
- Karlin, S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol.* **9**: 335-343.
- Kreil, D.P., and Ouzounis, C.A. 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Research.* **29**: 1608-1615.
- Lykidis, A., Mavromatis, K., Ivanova, N., Anderson, I., Land, M., DiBartolo, G., Martinez, M., Lapidus, A., Lucas, S., Copeland, A., et al. (2007) Genome sequence and analysis of the soil cellulolytic actinomycete *Thermobifida fusca* YX. *J Bacteriol.* **189**: 2477-2486.
- Lynn, D.J., Singer, G.A., and Hickey, D.A. 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.* 30: 4272-4277.
- Marchler-Bauer, A., Anderson, J.B., Derbyshire, M.K., DeWeese-Scott, C., Gonzales, N.R., Gwadz, M., Hao, L., He, S., Hurwitz, D.I., Jackson, J.D., et al. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.* **35(D)**: 237-240.
- Marechal, J., Clement, B., Nalin, R., Gandon, C., Orso, S., Cvejic, J.H., Bruneteau, M., Berry, A., and Normand, P. 2000. A *recA* gene phylogenetic analysis confirms the close proximity of *Frankia* to *Acidothermus*. *Int J Syst Evol Microbiol.* **50**: 781-785.
- Markowitz, V.M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I., et al. 2006. The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34(D)**: D344-348.

- Mohagheghi, A., Grohmann, K., Himmel, M., Leighton, L., and Updegraff, D.M. 1986. Isolation and characterization of *Acidothermus cellulolyticus* gen. nov., sp. nov., a new genus of thermophilic, acidophilic, cellulolytic bacteria. *Int. J. Syst. Bacteriol.* **36**: 435-443.
- Meyer, O., and Schlegel, H.G. 1983. Biology of aerobic carbon monoxide-oxidizing bacteria. *Annu Rev Microbiol.* **37**: 277-310.
- Monteiro-Vitorello, C.B., Camargo, L.E., Van Sluys, M.A., Kitajima, J.P., Truffi, D., do Amaral, A.M., Harakava, R., de Oliveira, J.C., Wood, D., de Oliveira, M.C., et al. 2004. The genome sequence of the gram-positive sugarcane pathogen *Leifsonia xyli* subsp. *xyli*. *Mol Plant Microbe Interact.* **17**: 827-836.
- Normand, P., Lapierre, P., Tisa, L.S., Gogarten, J.P., Alloisio, N., Bagnarol, E., Bassi, C.A., Berry, A.M., Bickhart, D.M., Choise, N., et al. 2007. Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography. *Genome Res.* **17**: 7-15.
- Normand, P., Orso, S., Cournoyer, B., Jeannin, P., Chapelon, C., Dawson, J., Evtushenko, L., and Misra, A.K. 1996. Molecular phylogeny of the genus *Frankia* and related genera and emendation of the family Frankiaceae. *Int J Syst Bacteriol.* **46**: 1-9.
- Omura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., Takahashi, Y., Horikawa, H., Nakazawa, H., Osonoe, T., et al. 2001. Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. *Proc Natl Acad Sci U.S.A.* **98**: 12215-12220.
- Park, S.W., Hwang, E.H., Park, H., Kim, J.A., Heo, J., Lee, K.H., Song, T., Kim, E., Ro, Y.T., Kim, S.W., and Kim, Y.M. 2003. Growth of mycobacteria on carbon monoxide and methanol. *J Bacteriol.* **185**: 142-147.
- Rainey, F.A., and Stackebrandt, E. 1993. Phylogenetic evidence for the classification of *Acidothermus cellulolyticus* into the subphylum of actinomycetes. *FEMS Microbiol. Lett.* **108**: 27-30.
- Reiser, J., Muheim, A., Hardegger, M., Frank, G., and Fiechter, A. 1994 Aryl-alcohol dehydrogenase from the white-rot fungus *Phanerochaete chrysosporium*. Gene cloning, sequence analysis, expression, and purification of the recombinant enzyme. *J. Biol. Chem.* **269**: 28152-28159.

- Rubin, E.M. (2008) Genomics of cellulosic biofuels. *Nature* 454:841-845
- Sakon, J., Adney, W.S., Himmel, M.E., Thomas, S.R., and Karplus, P.A. 1996. Crystal structure of thermostable family 5 endocellulase E1 from *Acidothermus cellulolyticus* in complex with cellotetraose. *Biochemistry*. **35**: 10648-10660.
- Sambrook, J., Fritsch, E.F., and Maniatis T. 1989. Molecular Cloning. A Laboratory Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
- Singer, G.A., and Hickey, D. A. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene*. **317**: 39-47.
- Suhre, K., and Claverie, J.M. 2003. Genomic Correlates of Hyperthermostability, an Update. *J. Biol. Chem*. **278**: 17198-17202.
- Sumby, P. and Smith, M.C. 2002. Genetics of the phage growth limitation (Pgl) system of *Streptomyces coelicolor* A3(2), *Mol Microbiol* **44**: 489–500.
- Takami, H., Takaki, Y., Chee, G.J., Nishi, S., Shimamura, S., Suzuki, H., Matsui, S., and Uchiyama, I. 2004. Thermoadaptation trait revealed by the genome sequence of thermophilic *Geobacillus kaustophilus*. *Nucleic Acids Res*. **32**: 6292-6303.
- Thomas, S.R., Laymon, R.A., Chou, Y.C., Tucker, M.P., Vinzant, T.B., Adney, W.S., Baker, J.O., Nieves, R.A., Mielenz, J.R., and Himmel, M.E. 1995. Initial approaches to artificial cellulase systems for conversion of biomass to ethanol. In *Enzymatic degradation of insoluble polysaccharides*. (eds. J.N. Saddler, and M.H. Penner). pp. 208-236. ACS Series 618, Washington, DC: American Chemical Society.
- Sticklen, M.B. (2008) Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. *Nature Reviews Genetics* 9: 433-443
- Tucker, M. P., Mohagheghi, A., Grohmann, K., and Himmel, M.E. 1989. Ultra-thermostable cellulases from *Acidothermus cellulolyticus*: comparison of temperature optima with previously reported cellulases. *Bio/Technology*. **7**: 817-820.

Yakhnin, H., Pandit, P., Petty, T.J., Baker, C.S., Romeo, T., and Babitzke, P. 2007. CsrA of *Bacillus subtilis* regulates translation initiation of the gene encoding the flagellin protein (hag) by blocking ribosome binding. *Mol Microbiol.* **64**: 1605-1620.

Zeldovich, K.B., Berezovsky, I.N., Shakhnovich, E.I. 2007. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol.* **3**: e5.

FIGURE LEGENDS

Figure 1. Schematic of the *A. cellulolyticus* 11B genome. The outermost circle gives the genome coordinates. The next two inner rings show the predicted genes on the leading (outer circle) and the lagging (inner circle) strands. Color scheme is as follows - dark grey: hypothetical proteins, light grey: conserved hypothetical and unknown function, brown: general function prediction, red: replication and repair, green: energy metabolism, blue: carbon and carbohydrate metabolism, cyan: lipid metabolism, magenta: transcription, yellow: translation, orange: amino acid metabolism, pink: metabolism of cofactors and vitamins, light red: purine and pyrimidine metabolism, lavender: signal transduction, sky blue: cellular processes, and pale green: structural RNAs. Ring 4 displays the positions of the glycoside hydrolases (black bars), the three GIs (triangles), the flagellar biosynthetic genes (red star), and the rRNA operon (blue star). Ring 5 shows the G+C content along the genome. The innermost ring, Ring 6, displays the GC-skew.

Figure 2. Genomic signature plot. A sliding window plot of the percent G+C content (top green line, y-axis on the left) as well as the deviation in genomic signature (Δ GS; bottom red line, secondary y-axis on right) along the chromosome. Regions 1, 2, and 3 on the plot indicate the location of the three GIs: GI1, GI2, and GI3, respectively. The arrow indicates the location of the flagellar and motility genes.

Figure 3. Synteny and gene organization of the flagellar biosynthetic genes in actinobacteria. The *A. cellulolyticus* locus Acel_0827-Acel_0864 is displayed; the syntenic region ranges from Acel_0829-Acel_0861. Ace, Krad, Lxy, and Noc denote *A. cellulolyticus*, *K. radiotolerans*, *L. xyli*, and *Nocardioides* sp. JS614, respectively. Chromosomal gene organization from each of the completely assembled genome is shown, except in the case of *K. radiotolerans* for which genes from two different contigs are shown. Therefore, the true order of the whole region in *K. radiotolerans* remains unclear. Synteny between the

different chromosomal regions is indicated by green lines (for genes on the same strand) and red lines (for genes on opposite strands). The gene sizes in the different organisms are not drawn to scale. Also, the *K. radiotolerans* genes are colored differently than the genes in the other three organisms.

Figure 4. Plot of the G+C content of non-coding RNAs (rRNA + tRNAs) versus the G+C of genome in prokaryotes. Red: hyperthermophiles, orange: thermophiles, green: mesophiles, blue: psychrophiles, magenta: *A. cellulolyticus*, maroon: *T. fusca*, yellow: *Frankia* sp. (ACN14a, CcI3), and dark green: *Streptomyces* sp. (*S. avermitilis*, *S. coelicolor*). Green lines represent the regression line and 95% confidence intervals, computed for the mesophiles.

Figure 5. Reduced dimensionality plot of PCA of amino acid usage in ribosomal proteins in 409 prokaryotes. Red: hyperthermophiles, orange: thermophiles, green: mesophiles, blue: psychrophiles, magenta: *A. cellulolyticus*, maroon: *T. fusca*, yellow: *Frankia* sp. (ACN14a, CcI3), and dark green: *Streptomyces* sp. (*S. avermitilis*, *S. coelicolor*).

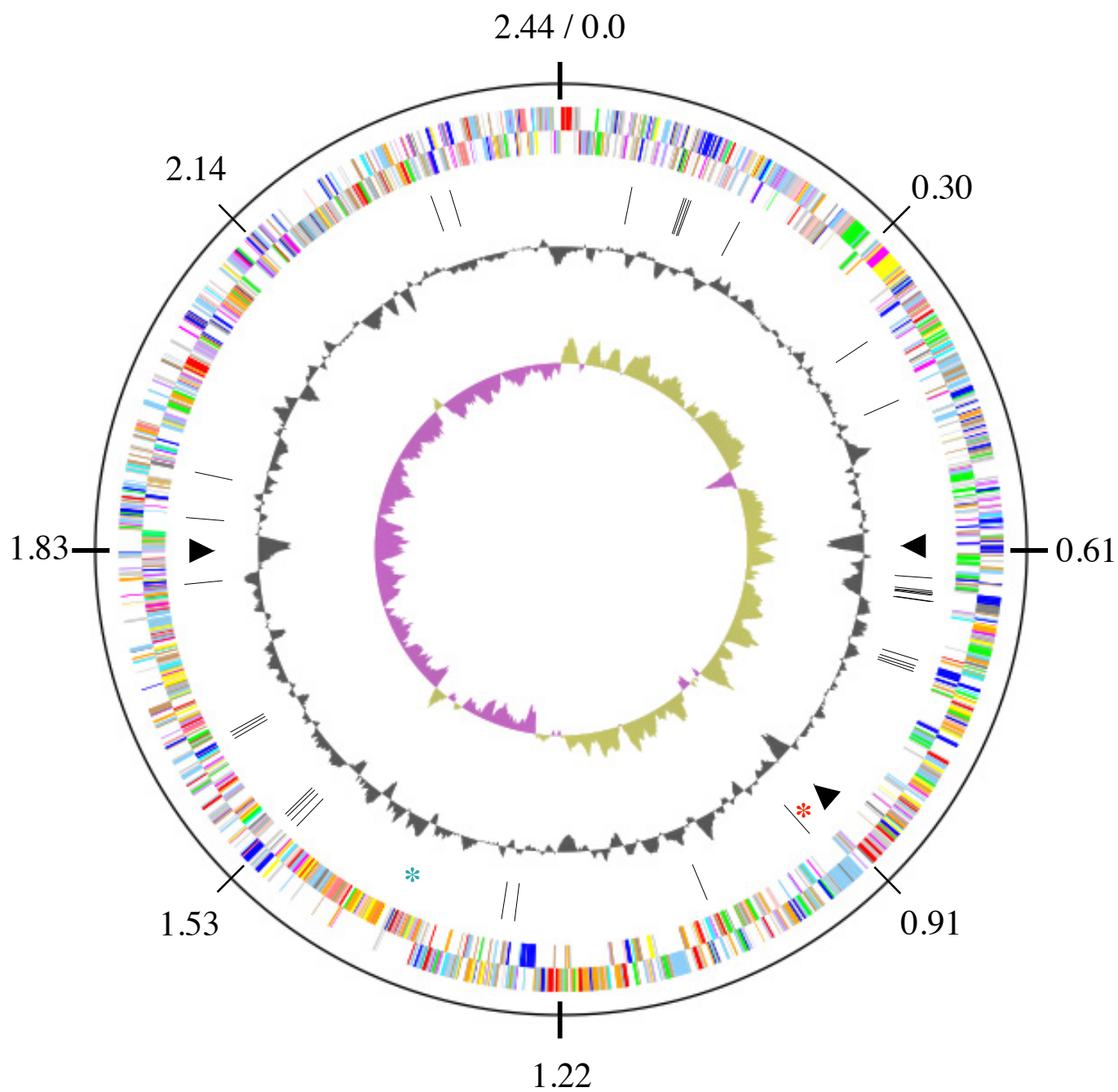


Fig. 1

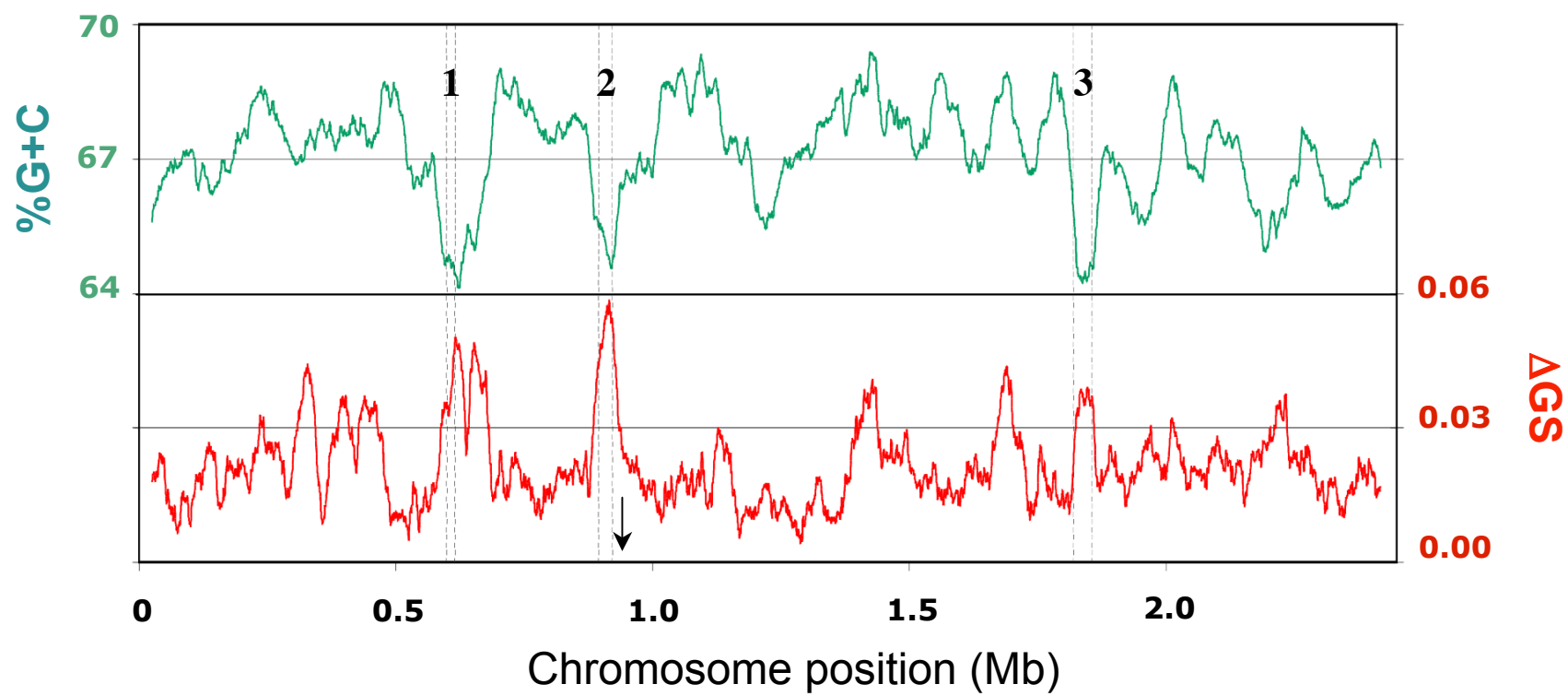


Fig. 2

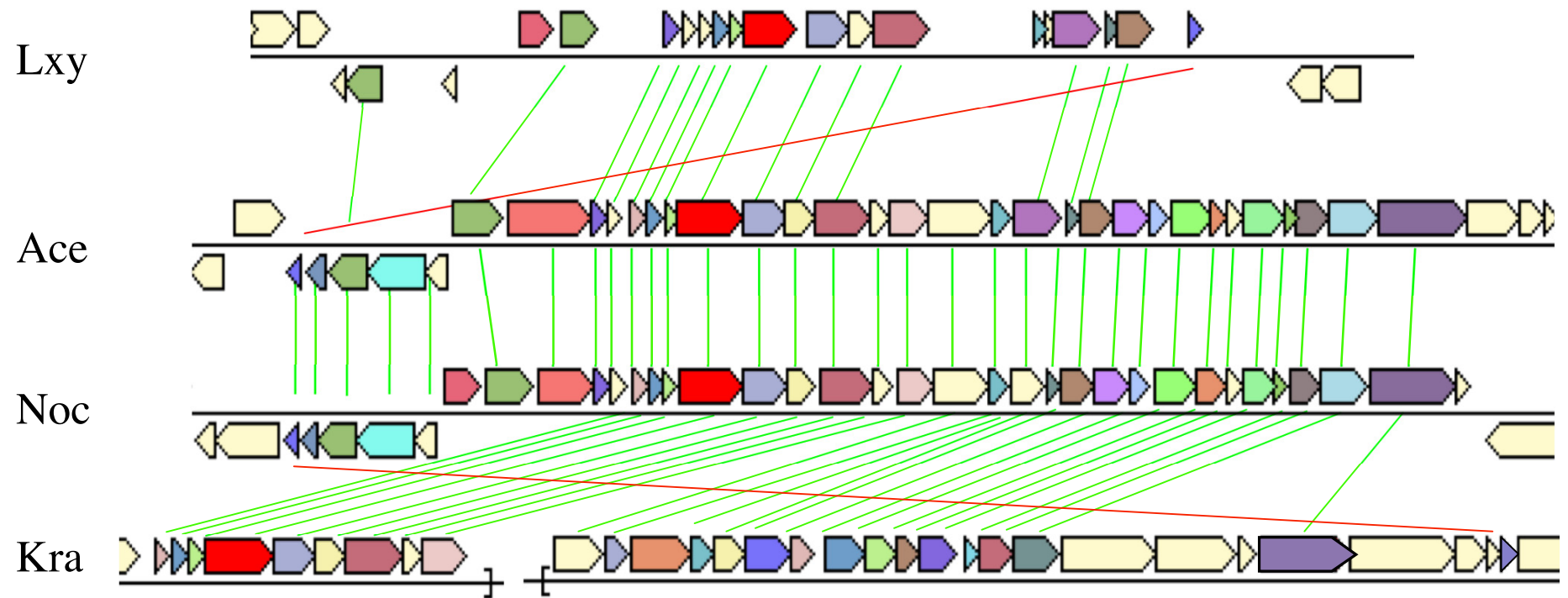
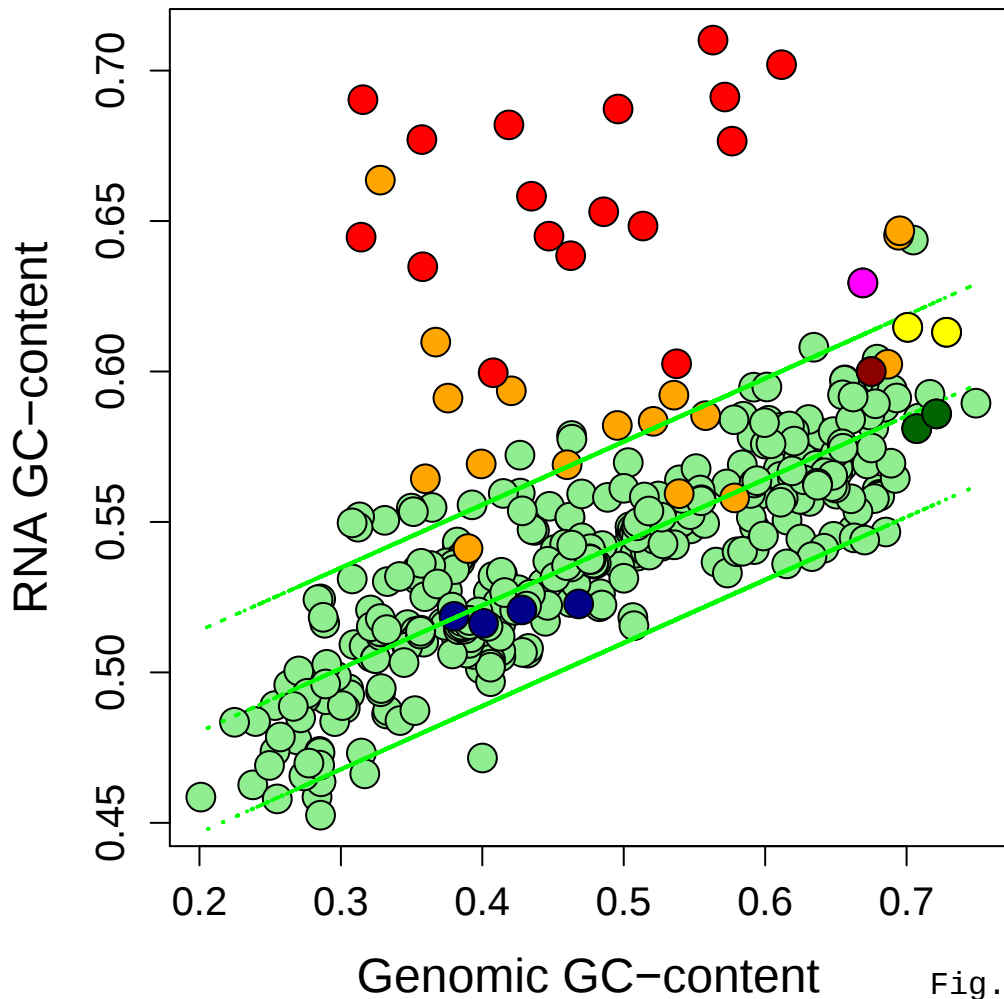


Fig. 3



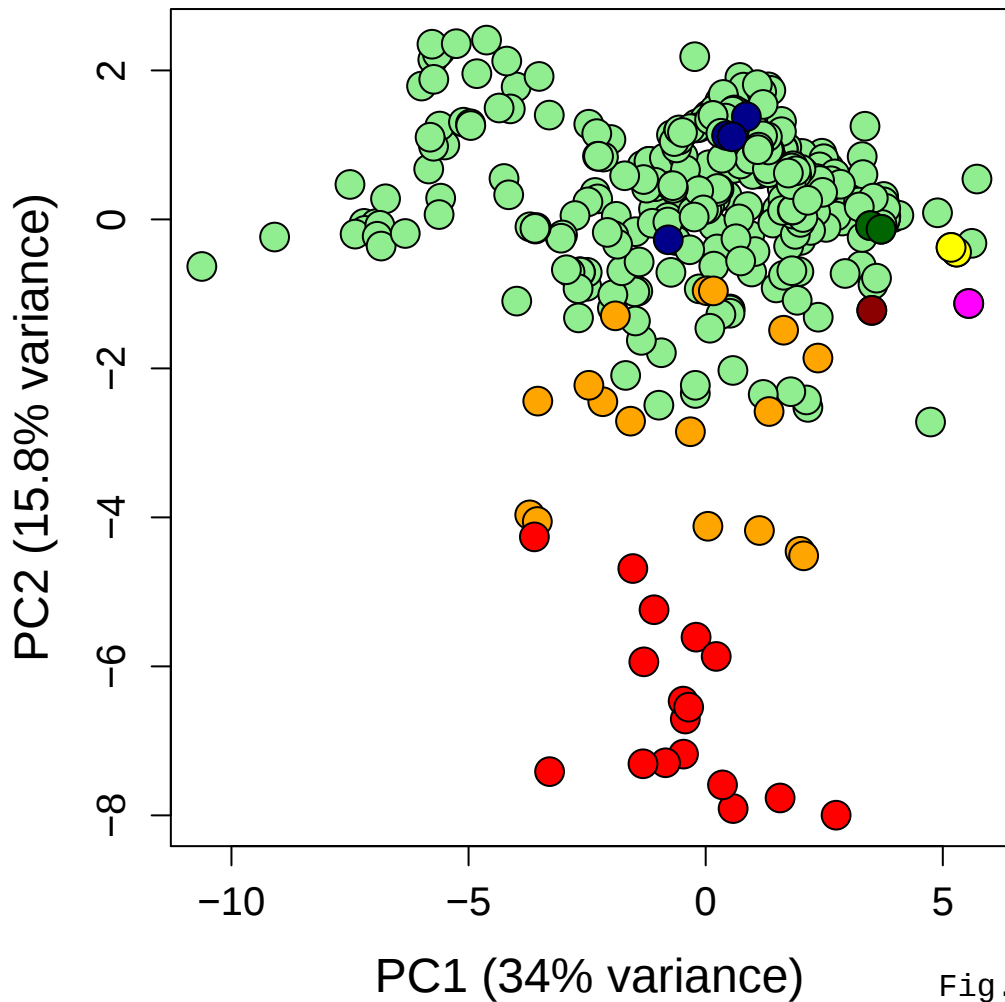


Fig. 5

Table 1. Comparative features of *Acidothermus cellulolyticus* 11B and close actinobacterial relatives.^a

Features	<i>Acidothermus cellulolyticus</i> 11B	<i>Frankia</i> sp. CcI3	<i>Frankia</i> <i>alni</i> ACN14a	<i>Streptomyces</i> <i>avermitilis</i> MA-4680	<i>Streptomyces</i> <i>coelicolor</i> A3(2)	<i>Thermobifida</i> <i>fusca</i> YX
OGT	55°C	27°C	28°C	28°C	30°C	50°C
Genome size (Mb)	2.4	5.4	7.5	9.0	8.7	3.6
G+C of the genome	66.9%	70.1%	72.8%	70.7%	72.1%	67.5%
Coding DNA fraction	89%	84%	86%	86%	88%	85%
Predicted proteins	2157	4499	6711	7577	7769	3110
rRNA operons	1	2	2	6	5	4
tRNA genes	46	46	46	68	64	52
Pseudogenes	4	50	12	0	56	7
Transposase/IS elements	2 ^b	145	33	110	55	5
Phage/viral proteins	0	6	24	20	8	3

^aThe genomes chosen for comparison were based on two attributes: (i) majority of the top BLAST hits of *A. cellulolyticus* proteins were from these species (see Supplementary Fig. S2), and (ii) both mesophilic and thermotolerant species were represented.

^b The 2 transposase genes are frame-shifted fragments of an intact gene found in *Frankia*, and thus, are unlikely to encode a functional transposase in *A. cellulolyticus*.

Table 2. Carbohydrate active enzymes encoded in the *A. cellulolyticus* 11B genome.

Locus tag	MW ^a	Domains ^b	Known or Predicted function	Role ^c	Sig ^d	Loc ^e	Ref ^f
Acel_0072	60	GH20	Beta-N-acetylhexosaminidase (EC 3.2.1.52)	Fun	Y	Cyt	
Acel_0128	50	GH3	Beta-N-acetylhexosaminidase	Fun	N	Cyt	
Acel_0129	49	GH16-CBM6	Endo-1,3-beta-glucanase	Fun	Y	U	
Acel_0133	53	GH1	Beta-glucosidase (EC 3.2.1.21)	Cel	N	Cyt	
Acel_0135	51	GH6	Beta-1,4-endoglucanase (CelB; EC 3.2.1.4)	Cel	Y	U	
Acel_0179	68	CE1-CBM3-CBM2	Acetyl-xylan esterase	Hem	Y	Sec	
Acel_0180	71	GH10-CBM3-CBM2	Beta-1,4-xylanase	Hem	Y	Sec	
Acel_0372	43	GH10	Endo-1,4-beta-xylanase (EC 3.2.1.8)	Hem	Y	Sec	
Acel_0374	27	CE14	Putative deacetylase	M	N	Cyt	
Acel_0424	83	GH18	N-acetylglucosaminidase	Fun	Y	Sec	
Acel_0557	40	CE9	N-acetylglucosamine 6-phosphate deacetylase	M	N	Cyt	
Acel_0603	51	GH18	Chitinase	Fun	Y	U	
Acel_0614	61	GH5-CBM2	Endo-1,4-glucanase E1 (Cel5A; EC 3.2.1.4)	Cel	Y	Sec	1,2
Acel_0615	125	GH6-CBM3-GH12-CBM2	Cellulase (GuxA; EC 3.2.1.4)	Cel	Y	Sec	2
Acel_0616	80	GH5-CBM3-CBM2	Mannanase (ManA)	Hem	Y	Sec	2
Acel_0617	119	CBM3-GH48-CBM2	Exoglucanase (Gux1)	Cel	Y	Sec	2
Acel_0618	134	GH74-CBM3-CBM2	Avicelase (Cel74A)	Cel	Y	Sec	2
Acel_0619	41	GH12-CBM2	Endoglucanase	Cel	Y	Sec	
Acel_0676	82	CBM48-CBM48-GH13	1,4-Alpha-glucan branching enzyme	G/T	N	Cyt	
Acel_0678	65	GH13	Trehalose synthase	G/T	N	Cyt	
Acel_0679	73	GH13	Alpha amylase	G/T	N	Cyt	
Acel_0681	78	CBM48-GH13	Glycogen debranching enzyme GlgX	G/T	N	Cyt	
Acel_0767	41	CE1	Putative esterase	Hem	Y	Sec	
Acel_0846	33	NLPC_P60-GH23	Lytic transglycosylase	M	N	Sec	
Acel_0970	95	GH9-CBM3-CBM2	Beta-1,4-endoglucanase	Cel	N	Sec	
Acel_1143	71	GH15	Trehalase/glucoamylase/glucoextranase	G/T	N	Cyt	
Acel_1157	41	GH23	Lytic transglycosylase	M	N	U	
Acel_1363	38	GH32	Putative beta-fructosidase	M	N	Cyt	
Acel_1372	80	CBM48-GH13	Glycogen debranching enzyme GlgX	G/T	N	Cyt	
Acel_1373	85	GH13	Malto-oligosyltrehalose synthase	G/T	N	Cyt	
Acel_1374	64	CBM48-GH13	Malto-oligosyltrehalose trehalohydrolase	G/T	N	Cyt	
Acel_1458	47	GH18-CBM16	Chitinase (EC 3.2.1.14)	Fun	N	SW	
Acel_1459	26	CBM16	Carbohydrate-binding CenC domain protein	Fun	Y	U	
Acel_1460	80	GH18-CBM5-CBM16	Chitinase (EC 3.2.1.14)	Fun	Y	SW	
Acel_1601	83	GH77	4-Alpha-glucanotransferase (EC 2.4.1.25)	M	N	Cyt	
Acel_1659	93	GH3-GH3C-PA14-GH3C	Beta-glucosidase (EC 3.2.1.21)	Cel	Y	Sec	
Acel_1701	120	GH9-CBM3-CBM3-CBM2	Endoglucanase	Cel	Y	Sec	
Acel_1868	31	CE14	Putative deacetylase	M	N	Cyt	
Acel_1886	36	CE14	Putative deacetylase	M	N	Cyt	
Acel_1917	27	CE4	Putative chitooligosaccharide deacetylase	Fun	N	Cyt	
Acel_2033	61	GH18-CBM5-CBM16	Chitinase (EC 3.2.1.14)	Fun	Y	SW	
Acel_2045	35	CE7	Acetyl xylan esterase	Hem	N	U	
Acel_2050	88	GH3-GH3C	Beta-D-xylosidase (EC 3.2.1.37)	Hem	N	Cyt	

^a MW = calculated molecular weight of the protein in kilo Daltons, rounded to a whole number.

^b Domain architecture was deciphered using the CAZy database (Coutinho and Henrissat, 1999; Henrissat, 1991; <http://www.cazy.org/>) and the Conserved Domains Search tool (<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>; Marchler-Bauer et al. 2007). The GH (glycoside hydrolase), CBM (carbohydrate binding module), and CE (carbohydrate esterase) family numbers are based on the CAZy classification

^c Cel, Hem, and Fun indicate a predicted role for the enzyme in cellulose, hemicellulose, and fungal cell wall degradation, respectively. G/T = glycogen/trehalose metabolism. M = cellular metabolism.

^d Sig = Signal peptide; Y and N indicate the presence or absence, respectively, of a predicted signal peptide in the protein sequence. The SignalP 3.0 software was used to predict the occurrence of signal peptides (Bendtsen et al., 2004).

^e Loc = Localization; The pSORTb prediction software (Gardy et al., 2005; <http://www.psort.org/psortb/>) was used to predict the subcellular localization of the protein. C = cytoplasmic, S = secreted/extracellular, and U = unknown localization. SW = proteins are predicted to be secreted as well as cell wall associated; and therefore they could occupy multiple locations.

^f Ref = References. 1 = Baker et al., 1994. 2 = Ding et al., 2003.

Table 3. Genes encoded on the three genomic islands found in the *A. cellulolyticus* 11B genome.

Locus Tag	S	%GC	Size	Product description	Function
<u>Genomic Island 1.</u>					
Acel_0569	+	58.7	446	fumarate reductase/succinate dehydrogenase flavoprotein	Respiration
Acel_0570	+	53.5	333	aryldialkylphosphatase	Organophosphate detoxification
Acel_0571	+	57.6	288	short-chain dehydrogenase/reductase SDR	Metabolism
Acel_0572	+	59.7	236	deoxyribose-phosphate aldolase	Nucleotide metabolism
Acel_0573	+	62.8	342	ROK family protein	Repressor/Kinase/ORF
Acel_0574	-	59.3	254	transcriptional regulator, GntR family	Regulation
Acel_0575	-	61.9	421	ROK family protein	Repressor/Kinase/ORF
Acel_0576	+	58.2	283	SIS (Sugar ISomerase) phosphosugar binding domain protein	Carbohydrate metabolism
Acel_0577	+	58.9	359	periplasmic binding protein/LacI transcriptional regulator	ABC transport
Acel_0578	+	59.0	489	ABC transporter related	ABC transport
Acel_0579	+	58.6	335	inner-membrane translocator	ABC transport
Acel_0580	+	56.6	330	inner-membrane translocator	ABC transport
Acel_0581	+	58.9	391	oxidoreductase domain protein	
Acel_0582	+	53.8	306	Xylose isomerase domain protein TIM barrel	Sugar interconversion
Acel_0583	-	59.1	397	oxidoreductase domain protein	Metabolism
<u>Genomic Island 2.</u>					
Acel_R0021	+	58.7		<i>Xaa tRNA</i>	
Acel_0810	+	59.0	61	DNA binding domain, excisionase family	VrII homolog
Acel_0811	+	59.3	159	conserved hypothetical protein	VrIJ homolog
Acel_0812	+	62.6	1244	conserved hypothetical protein	VrIK homolog
Acel_0813	+	61.0	468	putative transcriptional regulator	Transcriptional regulation
Acel_0814	+	60.5	993	conserved hypothetical protein	
Acel_0815	+	51.1	268	hypothetical protein	
Acel_0816	+	64.7	934	helicase domain protein	VrIO homolog?
Acel_0817	+	57.5	678	conserved hypothetical protein	VrIP homolog
Acel_0818	+	57.0	261	conserved hypothetical protein	VrIQ homolog
Acel_0819	+	67.2	64	hypothetical protein	
Acel_0820	+	68.8	446	metallophosphoesterase	DNA repair
Acel_0821	+	67.6	918	SMC domain protein	DNA repair
Acel_0822	+	66.7	502	acyltransferase 3	Metabolic enzyme
Acel_0823	-	66.5	548	diguanylate cyclase/phosphodiesterase	Metabolic enzyme
Acel_0824	-	65.0	122	hypothetical protein	
Acel_0825	-	66.2	206	protein of unknown function DUF421	
Acel_R0022	+	66.2		<i>Met tRNA</i>	
<u>Genomic Island 3.</u>					
Acel_R0044	+	68.5		<i>Arg tRNA</i>	
Acel_1621	+	51.1	92	hypothetical protein	
Acel_1622	+	62.8	162	hypothetical protein	
Acel_1623	+	64.0	89	transcriptional regulator, XRE family	Transcriptional regulation
Acel_1624	+	55.7	176	hypothetical protein	
Acel_1625	+	66.5	180	hypothetical protein	
Acel_1626	+	63.3	230	ABC transporter related	Transport
Acel_1627	+	65.8	426	protein of unknown function DUF214	
Acel_1628	+	63.7	168	methylglyoxal synthase	Enzyme
Acel_1629	-	64.9	483	methyl-accepting chemotaxis sensory transducer	Chemotaxis
Acel_1630	-	65.9	213	conserved hypothetical protein	
Acel_1631	-	65.5	358	protein of unknown function DUF182	
Acel_1632	-	54.8	208	conserved hypothetical protein	
Acel_1633	-	58.9	602	purine catabolism PucR domain protein	Purine degradation regulator
Acel_1634	-	59.6	327	conserved hypothetical protein	
Acel_1635	-	61.7	403	Pyridoxal-5'-phosphate-dependent enzyme, beta subunit	Metabolic enzyme
Acel_1636	-	62.2	238	carbon monoxide dehydrogenase subunit G, CoxG	CO fixation?
Acel_1637	-	59.7	162	carbon monoxide dehydrogenase small subunit, CoxS	CO fixation?
Acel_1638	-	61.4	296	carbon-monoxide dehydrogenase (acceptor), CoxM	CO fixation?
Acel_1639	-	59.2	231	Asp/Glu racemase	Amino acid metabolism
Acel_1640	-	58.9	560	polar amino acid ABC transporter, inner membrane subunit	Amino acid transport
Acel_1641	-	57.2	303	extracellular solute-binding protein, family 3	Solute uptake
Acel_1642	-	61.0	783	aldehyde oxidase and xanthine dehydrogenase	Metabolic enzyme
Acel_1643	-	60.4	262	coenzyme A transferase	Metabolic enzyme
Acel_1644	-	59.6	318	Glutaconate CoA-transferase	Metabolic enzyme
Acel_1645	-	55.6	316	luciferase family protein	Metabolic enzyme
Acel_1646	+	62.8	230	NADPH-dependent F420 reductase	Metabolic enzyme
Acel_1647	+	67.7	505	Malate dehydrogenase (oxaloacetate-decarboxylating)	Metabolic enzyme
Acel_1648	+	66.1	363	molybdenum cofactor biosynthesis protein A	Metabolic enzyme
Acel_1649	+	69.5	270	Exonuclease, RNase T and DNA polymerase III	Metabolic enzyme
Acel_R0045	+	59.2		<i>His tRNA</i>	

+/- indicates the DNA strand (S) that encodes the gene. The boxes indicate blocks of genes on the same strand with intergenic distance less than 50 bp. Size indicates the length of the predicted protein in amino acids. Product descriptions are based on automatic annotation of the gene. The last column provides a broad function of the protein.

Table 4. Relative proportions of each nucleotide at each of the three codon positions in six actinobacteria.

Organism	OGT (°C)	Nucleotide and Codon base position											
		Position 1 (5' end)				Position 2 (middle)				Position 3 (3' end)			
		A	C	G	T	A	C	G	T	A	C	G	T
<i>A. cellulolyticus</i> 11B	55	0.362	0.280	0.425	0.235	0.457	0.291	0.213	0.533	0.181	0.429	0.362	0.232
<i>Frankia alni</i> ACN14	28	0.388	0.267	0.413	0.255	0.514	0.278	0.213	0.621	0.098	0.455	0.374	0.124
<i>Frankia</i> sp. CcI3	27	0.382	0.277	0.408	0.247	0.487	0.282	0.216	0.580	0.131	0.441	0.376	0.173
<i>S. avermitilis</i> MA-4680	28	0.384	0.261	0.412	0.274	0.518	0.269	0.206	0.617	0.098	0.469	0.382	0.109
<i>S. coelicolor</i> A3(2)	30	0.381	0.258	0.417	0.275	0.534	0.264	0.208	0.644	0.086	0.478	0.375	0.081
<i>T. fusca</i> YX	50	0.357	0.272	0.424	0.256	0.481	0.265	0.212	0.591	0.163	0.463	0.364	0.153
R-squared value:		0.900	0.331	0.885	0.342	0.631	0.085	0.024	0.521	0.795	0.196	0.854	0.484
p-value less than:		0.004	0.232	0.005	0.223	0.059	0.575	0.772	0.105	0.017	0.380	0.008	0.125

Regression (R-squared) and p values were calculated using the R software. p-value less than 0.05 is considered significant.

SUPPLEMENTARY MATERIAL

SUPPLEMENTARY RESULTS AND DISCUSSION

tRNA and codon usage. Forty-five tRNAs representing 43 different anticodons are encoded in the genome (Supplementary Table S1). The tRNA^{Met} is present in three copies in the genome. In contrast to the number of tRNAs, all 61 sense codons are encoded in the genome sequence. The codon usage correlates well with the tRNA complement and is consistent with the high G+C content of the genome as the GC-rich codons predominate in the organism (Supplementary Table S1). Codons ATA (Ile), CGC (Arg), and CGA (Arg) as well as all codons that have a T at the third position, with the exception of CGT (Arg), do not appear to have a cognate tRNA in *A. cellulolyticus*. As a likely evolutionary adaptation to the available tRNAs for any given amino acid, codons that do not have a cognate tRNA occur with the least frequency in the *A. cellulolyticus* genome when compared to synonymous codons differing just in the 3rd position. However, the exceptions to these are for glycine (GGT (18.8%) > GGA (14.4%)), leucine (CTT (10.6%) > CTA (1.4%)), arginine (CGC (33.3%) > CGT (11.1%)), and valine (GTT (10.4%) > GTA (3.8%)). The relative preference for CGC codon over CGT codon follows the high G+C content of the genome, while the remaining four biases mentioned above may simply reflect evolutionary conservation of codon usage, as a similar trend is seen in *Frankia* (Supplementary Table S2). The functional significance of this bias remains elusive.

The *A. cellulolyticus* genome encodes a parsimonious complement of 46 tRNAs. Except for the three copies of tRNA for the ATG codon, all other tRNAs occur in single copy. In general, fast growing organisms have fewer species of tRNAs than slow growing organisms, although they may encode multiple copies of certain tRNAs (Rocha, 2004). Thus, based purely on the diversity of the tRNAs in the *A. cellulolyticus* genome, it can be predicted that the organism may be a relatively slow grower under natural conditions. The doubling time of this bacterium under optimal growth conditions has been estimated to be 6.7 hours (Mohagheghi et al., 1986), which is about 20 times longer than that of

Escherichia coli. However, several factors may influence growth rates of bacteria. Most fast growing bacteria, such as *E. coli* and *Bacillus subtilis*, have multiple copies of ribosomal RNA gene operons and at least one or more of these operons are usually on the leading strand and located close to origin of replication (Guy and Roten, 2004). The position of the single rRNA operon in *A. cellulolyticus* is far away from the replication site. This pattern is similar to that of rRNA operons in other actinobacteria, although in a relatively few actinobacterial genomes that possess multiple copies of *rrn* operons, at least one copy is closer to the OriC. Whether the distant location of the rRNA genes contributes to relatively slower growth rates of actinobacteria in general is yet to be determined.

SUPPLEMENTARY METHODS.

Genomic Regions (GR) in *Acidothermus cellulolyticus*. The method used is implemented in the microbial genome annotation and comparative analysis platform MaGe (Vallenet *et al.*, 2006) developed at Genoscope. It combines conservation of synteny groups between related bacteria, composition abnormalities and GI flanking features such as tRNA, IS and repeats.

In a first step, we delineated the core gene pool from the flexible gene pool of a query sequence (conserved backbone) by comparing this sequence to a selected set of related genomes. The set of orthologous genes (Bidirectional Best Hits or BBH) between the query, here *Acidothermus cellulolyticus* and the compared organisms, (*Streptomyces avermitilis*, *S. coelicolor*, *S. cattleya*, Frankia sp. EAN1pec, Frankia sp. CcI3, Frankia *alni*, and *Thermobifida fusca*) were searched for. The concept of synteny (*i.e.*, local conserved gene organization between organisms) computed as explained in Vallenet *et al.*, 2006, was introduced. Genes in BBH inside synteny groups between all compared organisms are more likely to be part of the query sequence backbone. Then, to delineate Genomic Regions, we retained regions above 5 kb in length, which were found between two conserved blocks in *Acidothermus cellulolyticus* (they actually fall between two synteny break points).

In a second step, these Genomic Regions were analyzed to find some common Genomic Island characteristics such as tRNA and/or tRNA repeats, integrase, atypical GC content and Codon Adaptation

Index value (Sharp and Li, 1987), short DNA repeats or combinations of these features. Finally, to retrieve regions shorter than 5 kb the IVOMs results, software which is based on compositional biases using variable order motif distributions (Vernikos and Parkhill, 2006), was also combined with the set of predicted Genomic Regions.

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Phylogenetic tree of the actinobacteria. Sequences retrieved from 16S rRNA genes of actinobacterial representative completely sequences were aligned using Clustalx (Thompson et al. 1997). Phylogenetic tree reconstruction was based on the Neighbor joining method (Saitou and Nei 1987). Distances were corrected for multiple substitutions (Kimura 1980), using the No-gap option to exclude indel-containing positions, otherwise default settings were used. Numbers give percent bootstrap support from 1000 samples when above 50%. Distances in the bar are in substitutions/site.

Figure S2. Taxonomic distribution of the top BLAST hits to *A. cellulolyticus* proteins.

Figure S3. Reduced dimensionality plot of PCA of global codon usage (A) and synonymous codon usage (B) of the 61 sense codons in 409 prokaryotes. Red: hyperthermophiles, orange: thermophiles, green: mesophiles, blue: psychrophiles, magenta: *A. cellulolyticus*, maroon: *T. fusca*, yellow: *Frankia* sp. (ACN14a, CcI3), and dark green: *Streptomyces* sp. (*S. avermitilis*, *S. coelicolor*).

Figure S4. Reduced dimensionality plot of PCA of the usage of the twenty amino acids in 409 prokaryotes. Red: hyperthermophiles, orange: thermophiles, green: mesophiles, blue: psychrophiles, magenta: *A. cellulolyticus*, maroon: *T. fusca*, yellow: *Frankia* sp. (ACN14a, CcI3), and dark green: *Streptomyces* sp. (*S. avermitilis*, *S. coelicolor*).

SUPPLEMENTARY TABLES

Table S1. tRNA and codon usage in *A. cellulolyticus* 11B.

Table S2. *A. cellulolyticus* 11B proteins that have best BLAST-hits to Archaea or Eukarya.

Table S3. Salient features of additional genomic regions (GR) identified in the genome of *A. cellulolyticus* 11B.

Table S4. Comparative analysis of global codon usage in six actinobacteria.

Table S5. Total fraction of IVYWREL amino acids in 478 orthologous proteins from each of the six actinobacteria.

Table S6. Total fraction of IVYWREL amino acids in 46 orthologous proteins from forty-five completely sequenced actinobacteria.

Table S7. Features of the 409 prokaryotic organisms used in this study.

Table S8. The 478 *A. cellulolyticus* 11B homologs used in the analysis presented in Table S5.

Table S9. The 46 *A. cellulolyticus* 11B homologs used in the analysis presented in Table S6.

REFERENCES

1. Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16: 111-120.
2. Rocha, E.P. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* 14: 2279-2286.
3. Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406-425.
4. Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25: 4876-4882.

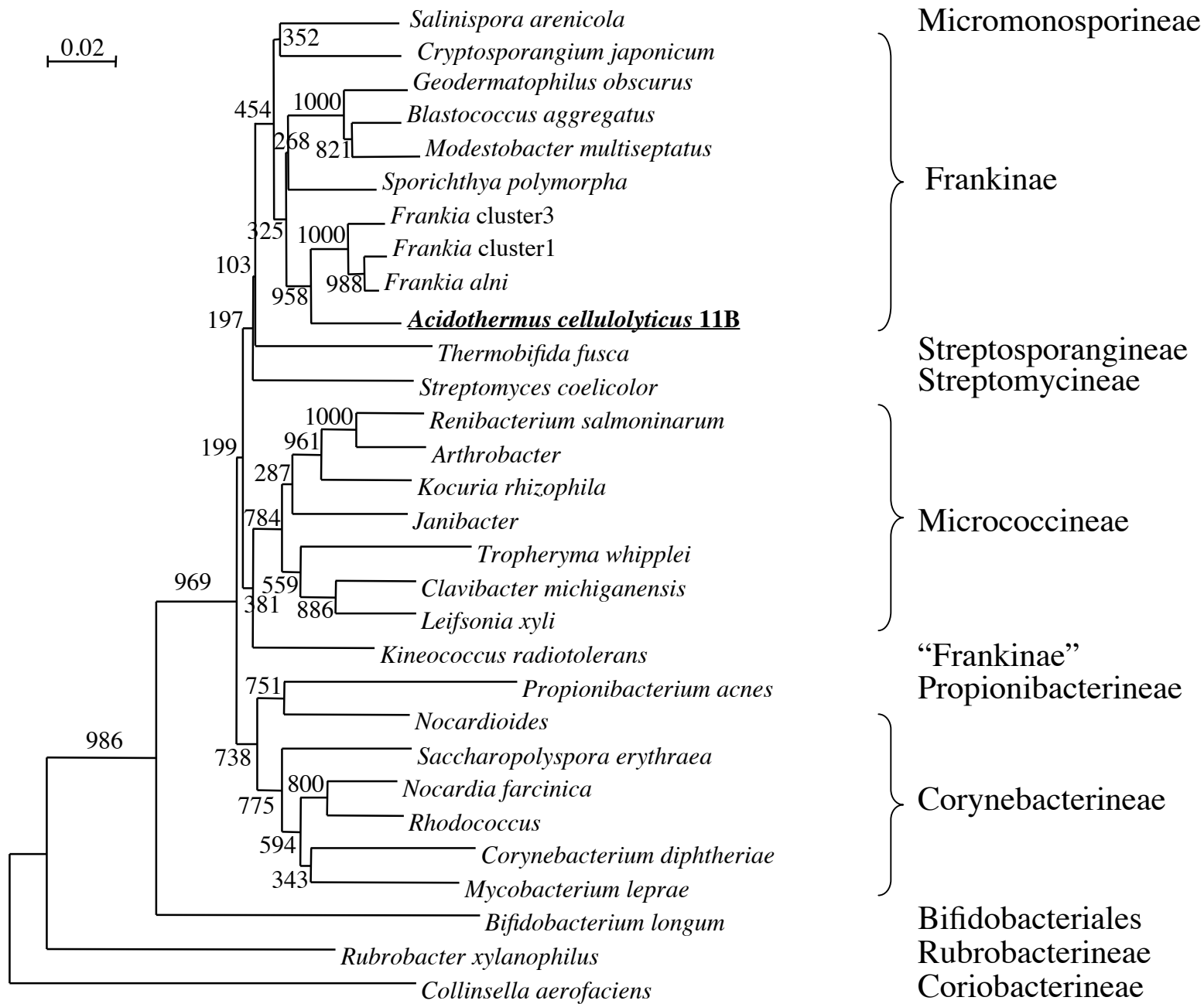


Figure S1

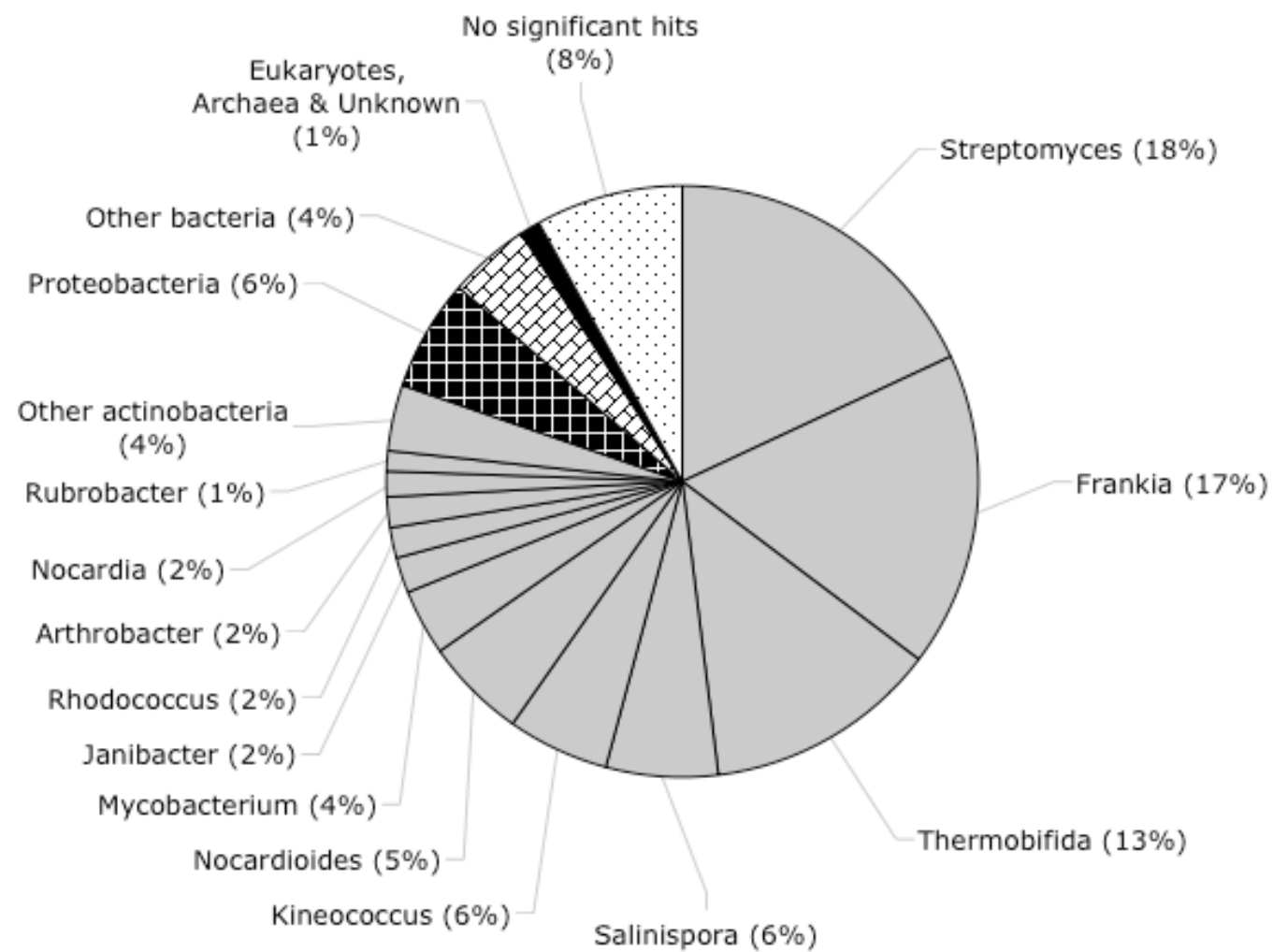


Figure S2

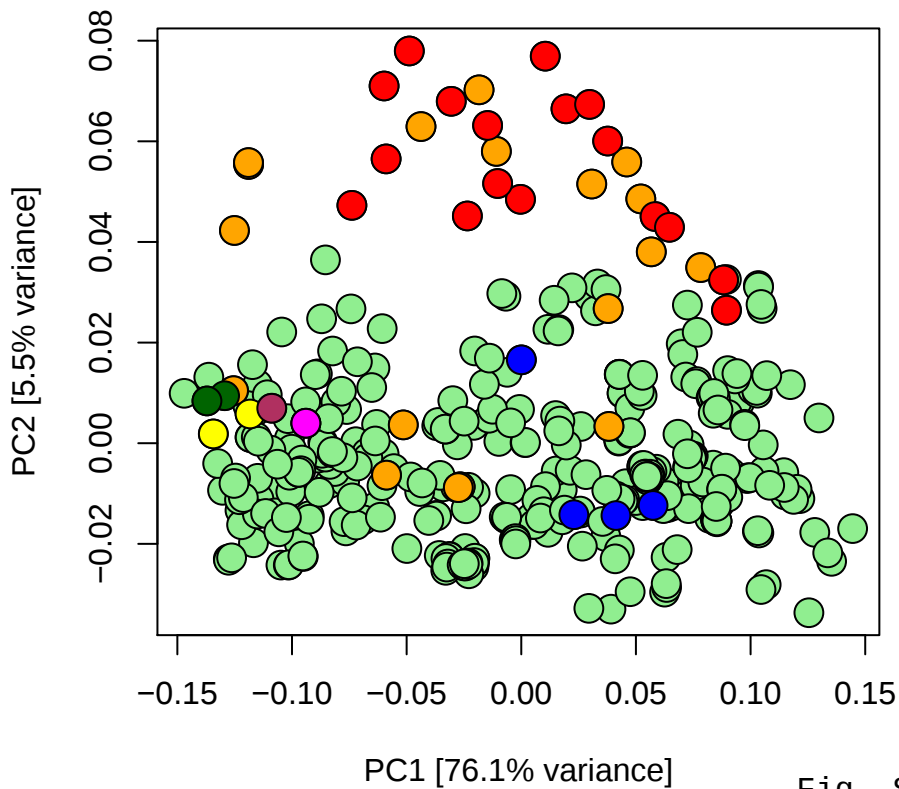


Fig. S3A

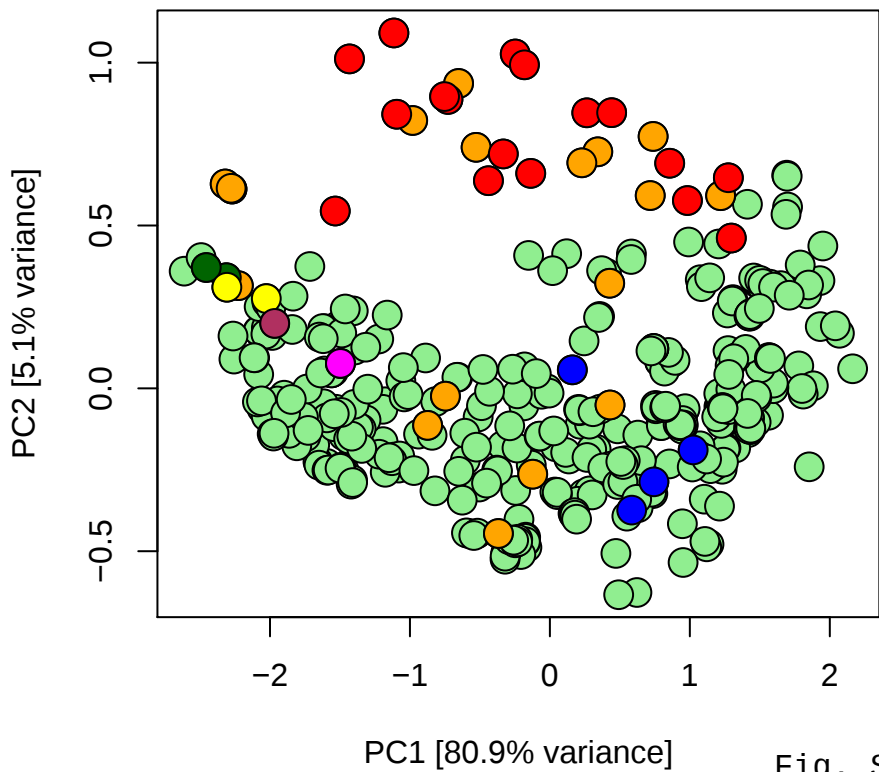


Fig. S3B

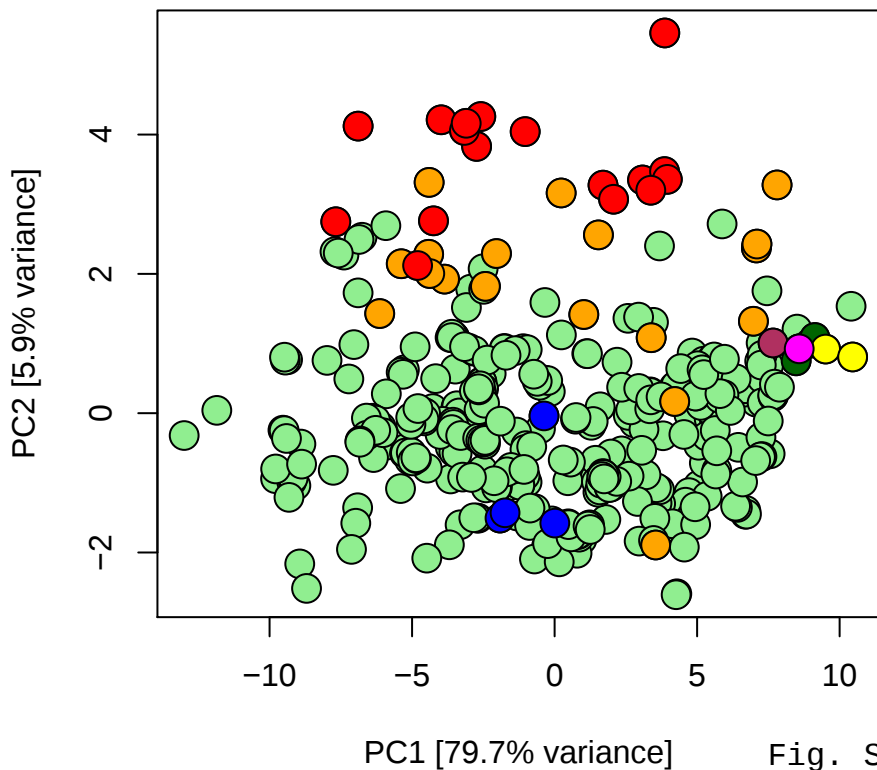


Fig. S4

Table S1. tRNA and codon usage in *A. cellulolyticus* 11B.

Amino acid	Codon	Synonymous codon usage (%)	tRNA location in the genome	Anticodon
Alanine (A)	GCC	40.8	842997..843072	GGC
Alanine (A)	GCG	40.9	179235..179159	CGC
Alanine (A)	GCA	10.1	10372..10447	TGC
Alanine (A)	GCT	8.2	-	-
Arginine (R)	CGG	43.4	677195..677124	CCG
Arginine (R)	CGT	11.1	41078..41153	ACG
Arginine (R)	AGA	1.3	1824619..1824691	TCT
Arginine (R)	AGG	2.3	160793..160721	CCT
Arginine (R)	CGC	33.4	-	-
Arginine (R)	CGA	8.4	-	-
Asparagine (N)	AAC	67.7	893343..893418	GTT
Asparagine (N)	AAT	32.3	-	-
Aspartic acid (D)	GAC	72.9	2339683..2339760	GTC
Aspartic acid (D)	GAT	27.1	-	-
Cysteine (C)	TGC	77.8	1524883..1524954	GCA
Cysteine (C)	TGT	22.2	-	-
Glutamic acid (E)	GAG	63.2	783484..783559	CTC
Glutamic acid (E)	GAA	36.8	2339526..2339598	TTC
Glutamine (Q)	CAG	76.2	783342..783416	CTG
Glutamine (Q)	CAA	23.8	2206822..2206752	TTG
Glycine (G)	GGG	19.3	2389324..2389254	CCC
Glycine (G)	GGA	14.4	1797034..1797104	TCC
Glycine (G)	GGC	47.5	1524762..1524834	GCC
Glycine (G)	GGT	18.8	-	-
Histidine (H)	CAC	72.2	1855719..1855794	GTG
Histidine (H)	CAT	27.8	-	-
Isoleucine (I)	ATC	70.4	10111..10184	GAT
Isoleucine (I)	ATT	26.1	-	-
Isoleucine (I)	ATA	3.4	-	-
Leucine (L)	CTG	34.9	24698..24783	CAG
Leucine (L)	TTA	1.1	2147532..2147460	TAA
Leucine (L)	CTA	1.4	1894564..1894636	TAG
Leucine (L)	CTC	39.5	1294313..1294229	GAG
Leucine (L)	TTG	12.5	1195870..1195798	CAA
Leucine (L)	CTT	10.6	-	-
Lysine (K)	AAA	29.6	2336318..2336246	TTT
Lysine (K)	AAG	70.4	1861376..1861451	CTT
Methionine (M)	ATG	100	921140..921213	CAT
Methionine (M)	ATG	100	509883..509959	CAT
Methionine (M)	ATG	100	308433..308509	CAT
Phenylalanine (F)	TTC	79.3	2339803..2339879	GAA
Phenylalanine (F)	TTT	20.7	-	-
Proline (P)	CCG	63.7	2281657..2281581	CGG
Proline (P)	CCA	7.1	1796952..1796877	TGG
Proline (P)	CCC	23.0	1372838..1372762	GGG
Proline (P)	CCT	6.3	-	-
Serine (S)	AGC	25.8	40835..40924	GCT
Serine (S)	TCA	6.5	2374948..2375032	TGA
Serine (S)	TCG	29.9	2293483..2293394	CGA
Serine (S)	TCC	26.1	2292585..2292671	GGA
Serine (S)	AGT	7.8	-	-
Serine (S)	TCT	4.0	-	-
Threonine (T)	ACA	7.1	89585..89660	TGT
Threonine (T)	ACC	50.3	308349..308424	GGT
Threonine (T)	ACG	37.6	2230630..2230555	CGT
Threonine (T)	ACT	5.0	-	-
Tryptophan (W)	TGG	100	313372..313447	CCA
Tyrosine (Y)	TAC	76.6	307544..307629	GTA
Tyrosine (Y)	TAT	23.4	-	-
Valine (V)	GTA	3.8	974815..974886	TAC
Valine (V)	GTC	50.9	1525000..1525074	GAC
Valine (V)	GTG	34.9	1523720..1523649	CAC
Valine (V)	GTT	10.4	-	-
Unknown	?	-	895201..895275	?
Stop codon	TGA	67.1	-	-
Stop codon	TAG	21.3	-	-
Stop codon	TAA	11.6	-	-

Table S2. *A. cellulolyticus* 11B proteins that have best BLAST-hits to Archaea or Eukarya.

Protein ID	Size	Protein description	GI of Best hit	Best hit organism	Blast2Seq score
<u>(A) Best hits to Archaea</u>					
Acel 0034	245	hypothetical protein Acel 0034	110667166	<i>Haloquadratum walsbyi</i> DSM 16790	224
Acel 0498	111	protein of unknown function DUF59	14590230	<i>Pyrococcus horikoshii</i> OT3	264
Acel 0525	137	hypothetical protein Acel 0525	15897985	<i>Sulfolobus solfataricus</i> P2	291
Acel 0526	204	hypothetical protein Acel 0526	88604270	<i>Methanospirillum hungatei</i> JF-1	111
Acel 0621	366	hypothetical protein Acel 0621	110667166	<i>Haloquadratum walsbyi</i> DSM 16790	225
Acel 0638	381	UDP-N-acetylglucosamine 2-epimerase	15678857	<i>Methanothermobacter thermautotrophicus</i> str. Delta H	519
Acel 0721	283	fructose-bisphosphate aldolase	15922681	<i>Sulfolobus tokodaii</i> str. 7	320
Acel 0888	284	ABC transporter related	73668563	<i>Methanosarcina barkeri</i> str. Fusaro	588
Acel 0910	245	ABC transporter related	110669070	<i>Haloquadratum walsbyi</i> DSM 16790	620
Acel 1270	291	ATP phosphoribosyltransferase (homohexameric)	116753404	<i>Methanosaeta thermophila</i> PT	780
Acel 1310	219	hypothetical protein Acel 1310	110667166	<i>Haloquadratum walsbyi</i> DSM 16790	163
Acel 1639	230	Asp/Glu racemase	14521305	<i>Pyrococcus abyssi</i> GE5	408
Acel 1644	317	Glutaconate CoA-transferase	11498798	<i>Archaeoglobus fulgidus</i> DSM 4304	495
Acel 1897	233	Small-conductance mechanosensitive channel-like	48477585	<i>Picrophilus torridus</i> DSM 9790	200
Acel 1930	282	ABC-2 type transporter	15899372	<i>Sulfolobus solfataricus</i> P2	221
Acel 1931	347	ABC transporter related	119720042	<i>Thermofilum pendens</i> Hrk 5	406
Acel 2066	164	Vitamin K epoxide reductase	70606913	<i>Sulfolobus acidocaldarius</i> DSM 639	193
Acel 2067	303	hypothetical protein Acel 2067	15898666	<i>Sulfolobus solfataricus</i> P2	355
<u>(B) Best hits to Eukarya</u>					
Acel 0064	898	hypothetical protein Acel 0064	118129698	<i>Gallus gallus</i>	294
Acel 0179	656	esterase, PHB depolymerase family	114324587	<i>Volvariella volvacea</i>	753
Acel 0740	439	hypothetical protein Acel 0740	97180301	<i>Sus scrofa</i>	264
Acel 0770	160	hypothetical protein Acel 0770	118085709	<i>Gallus gallus</i>	112
Acel 1067	206	GPR1/FUN34/yaaH family protein	119178442	<i>Coccidioides immitis</i> RS	354
Acel 1163	225	cell wall surface anchor family protein	109658562	<i>Homo sapiens</i>	180
Acel 1712	323	hypothetical protein Acel 1712	46119356	<i>Gibberella zeae</i> PH-1	760
Acel 1727	219	beta-lactamase domain protein	125820913	<i>Danio rerio</i>	438

Table S3. Salient features of additional genomic regions (GR) identified in the genome of *A. cellulolyticus* 11B.

GR number	Start coordinate	End coordinate	Description and Features
GR1	10413	13165	Mainly hypothetical proteins – specific to <i>Acidothermus</i> compared to the 7 selected genomes (see methods).
GR2	24757	34541	Enzymatic activities (ATPase + Kelch repeat possibly in a galactose oxidase).
GR3	40836	53331	Several (conserved) hypothetical protein + enzymatic activities + transporter and 1 regulator.
GR4	121404	141702	Mainly specific (conserved) hypothetical protein in the first part and cellulose transport + metabolism (degradation) shared with <i>Streptomyces</i> species, <i>Frankia</i> EAN1 and <i>T. fusca</i> .
GR5	510185	531977	First part, unknown metabolism with transport (membrane proteins), regulator, and enzymatic activities (transferase, oxidoreductase, phosphoesterase). Second part, highly specific, only hypothetical proteins + one enzyme probably involved in aromatic compound metabolism
GR6	532234	551342	First part, probably nitrate metabolism with transporter and nitrate reductase activity (shared with <i>S. coelicolor</i> only). Second part, specific (conserved) hypothetical proteins.
GR7	650219	677497	Cluster of protein/enzymes involved in cellulose degradation (specific to <i>Acidothermus</i> although partial homologs exists in the compared species).
GR8	688548	694890	Enzymatic activities (glycosyltransferase, carbamoylphosphate, epimerase, hydrolase)
GR9	762052	769864	Pyruvate synthase enzyme (containing iron-sulfur binding domains) specific to <i>Acidothermus</i> – Actually, the genes in the region encoded a pyruvate oxidoreductase or Pyruvic-ferredoxin oxidoreductase. The cluster is also found in <i>Helicobacter pylori</i> strains annotated as: PorC = Pyruvate oxidoreductase gamma chain (ACICE0782) PorD = Pyruvate oxidoreductase delta chain (partial match on ACICE0785 but more than 51% identity in aa). PorA = Pyruvate oxidoreductase alpha chain (ACICE0783) PorB = Pyruvate oxidoreductase beta chain (ACICE0784)
GR10	783644	809485	First part, transport system + regulator + enzymatic activities (kinase, oxidase, glycosyltransferase). Second part, highly specific to <i>Acidothermus</i> , cluster <i>hyf</i> genes coding hydrogenase subunits (NADH dehydrogenase (ubiquinone)/ ATP synthesis coupled electron transport). This cluster is found in a very well conserved synteny in: <i>Anaeromyxobacter</i> species (6 genes, from ACICE0811 to ACICE0816, identities % between 30 and 40) with the annotation ‘NADH dehydrogenase (quinone)’, and in <i>Yersinia</i> species (6 genes from ACICE0811 to ACICE0816, identities % between 25 and 35) with the annotation ‘hydrogenase 4 subunit a, B, C, D, F, G, H, I and J – subunits H and D are missing in <i>Acidothermus</i> .
GR11	975117	987762	Transport system (ABC type, substrate nitrate?), regulator (lacI family) and putative nitrilase. Highly specific to <i>Acidothermus</i> . Nitrilase (ACICE0994) and ACICE0997 in synteny with two genes of the <i>Rhizobium leguminosarum</i> bv. <i>viciae</i> 3841 plasmid pRL80076 (putative aliphatic nitrilase) and pRL80075 (putative endoribonuclease L-PSP family protein).
GR12	1122316	1130148	First part, shared with <i>Frankia</i> Ccl3 only, enzymatic activities (cytochrome c3 hydrogenase), second part more specific to <i>Acidothermus</i> with (conserved) hypothetical proteins.
GR13	1196923	1203395	Transport system (type ABC, substrate amino acid?) shared with <i>Frankia</i> species only.
GR14	1212310	1224927	Metabolism, very probably degradation (monooxygenase, dioxygenase) of glutamine/glutamate? + Regulator (marR family) + transport? (permease). Highly specific to <i>Acidothermus</i> .
GR15	1284136	1295293	Mainly hypothetical proteins with a putative rRNA methylase and exonuclease. Mainly specific to <i>Acidothermus</i> .
GR16	1459534	1477754	Type IV pilus (or type II?) highly specific to <i>Acidothermus</i> . Best synteny group shared with <i>Kineococcus radiotolerans</i> SRS30216 (10 genes, %identity: 30-74), <i>Moorella thermoacetica</i> ATCC 39073 (7 genes, %identity: 40-55), ...
GR17	1513103	1527815	Transport (ABC type, sugar?) + regulator (lacI family) + enzymatic activities (levabase, oxidase)
GR18	1571301	1581786	Mainly specific conserved hypothetical proteins
GR19	1603055	1612988	Cluster of enzymatic functions involved in aromatic compound degradation (paa genes cluster for phenylacetic acid degradation) + regulation (<i>tetR</i> family). Shared with <i>Streptomyces</i> species only.
GR20	1626197	1645147	Mainly conserved hypothetical proteins and 2 copies of a chitinase (involved in Chitin degradation), one is a pseudogene (ACICE1629+1630) and one seems to be functional (ACICE1631).
GR21	2164737	2199723	Many conserved hypothetical proteins + enzymatic activities, probably involved in cell wall biogenesis (glycosyltransferases) = degradation of unknown compound? + Transport system. Highly specific to <i>Acidothermus</i> in the second part.
GR22	2207054	2241063	Mainly conserved hypothetical proteins + transport system + <i>marR</i> family regulator + enzymes (oxidoreductase).
GR23	2299648	2314028	Mainly specific conserved hypothetical proteins + chitinase (chitin degradation)
GR24	2353384	2358909	Only specific conserved hypothetical proteins + regulator (marR family) and probable transporter.
GR25	2383486	2389949	Only conserved hypothetical proteins + regulator (fragment) and probable transporter.

Table S4. Comparative analysis of global codon usage in six actinobacteria.

Amino acid	Codon	<i>Acidothermus cellulolyticus</i> 11B	<i>Frankia</i> sp. ACN14	<i>Frankia</i> sp. Ccl3	<i>Streptomyces s. avermitilis</i>	<i>Streptomyces s. coelicolor</i>	<i>Thermobifida fusca</i>
A	GCG	5.62	5.90	5.37	5.40	5.02	4.31
A	GCC	5.58	7.80	6.83	6.99	7.86	5.91
A	GCA	1.37	0.58	0.74	0.62	0.53	1.12
A	GCT	1.12	0.43	0.70	0.40	0.28	1.20
C	TGC	0.66	0.67	0.67	0.69	0.70	0.69
C	TGT	0.19	0.10	0.16	0.11	0.07	0.12
D	GAC	4.20	5.38	4.87	5.43	5.82	5.21
D	GAT	1.56	0.77	1.16	0.48	0.29	0.58
E	GAG	3.23	4.07	3.94	4.62	4.84	3.77
E	GAA	1.89	0.68	0.99	0.97	0.84	2.55
F	TTC	2.32	2.42	2.40	2.66	2.60	2.60
F	TTT	0.60	0.12	0.20	0.07	0.04	0.21
G	GGC	4.17	5.76	4.63	5.81	6.15	4.08
G	GGG	1.67	2.40	2.38	1.85	1.85	2.20
G	GGT	1.64	1.15	1.54	1.04	0.93	0.94
G	GGA	1.25	0.70	0.95	0.77	0.71	1.15
H	CAC	1.57	1.81	1.72	2.03	2.17	2.07
H	CAT	0.60	0.37	0.57	0.31	0.16	0.25
I	ATC	2.95	3.02	3.16	2.95	2.73	3.49
I	ATT	1.09	0.14	0.23	0.10	0.06	0.32
I	ATA	0.14	0.05	0.10	0.08	0.07	0.06
K	AAG	1.21	1.13	1.28	2.16	1.94	1.33
K	AAA	0.50	0.09	0.20	0.14	0.10	0.66
L	CTC	3.97	3.58	3.48	3.91	3.66	3.75
L	CTG	3.52	5.74	5.32	5.67	6.14	5.14
L	TTG	1.25	0.38	0.62	0.34	0.24	0.84
L	CTT	1.06	0.27	0.47	0.23	0.15	0.44
L	CTA	0.12	0.10	0.22	0.05	0.03	0.16
L	TTA	0.10	0.02	0.03	0.01	0.01	0.04
M	ATG	1.49	1.32	1.49	1.60	1.57	1.62
N	AAC	1.29	1.39	1.45	1.69	1.62	1.81
N	AAT	0.61	0.11	0.18	0.12	0.07	0.13
P	CCG	3.94	4.06	3.75	3.29	3.37	2.77
P	CCC	1.41	2.34	2.19	2.41	2.55	2.52
P	CCA	0.43	0.28	0.43	0.17	0.13	0.28
P	CCT	0.38	0.23	0.32	0.20	0.14	0.55
Q	CAG	2.10	2.39	2.31	2.65	2.50	2.52
Q	CAA	0.65	0.13	0.22	0.18	0.13	0.47
R	CGG	3.71	3.77	4.01	2.85	3.22	3.04
R	CGC	2.84	3.64	3.03	3.61	3.90	3.72
R	CGT	0.94	0.64	0.95	0.73	0.54	0.76
R	CGA	0.71	0.45	0.56	0.29	0.24	0.35
R	AGG	0.18	0.31	0.41	0.38	0.36	0.21
R	AGA	0.10	0.08	0.13	0.08	0.08	0.12
S	TCG	1.52	1.69	1.62	1.60	1.39	1.03
S	TCC	1.32	1.57	1.64	1.93	2.03	1.93
S	AGC	1.31	1.42	1.35	1.30	1.23	1.55
S	AGT	0.39	0.19	0.27	0.19	0.15	0.26
S	TCA	0.33	0.13	0.21	0.14	0.10	0.19
S	TCT	0.19	0.09	0.15	0.09	0.06	0.27
T	ACC	2.96	3.68	3.62	3.55	3.97	3.92
T	ACG	2.22	1.95	1.92	2.29	1.91	1.36
T	ACA	0.42	0.17	0.28	0.23	0.15	0.29
T	ACT	0.29	0.14	0.23	0.15	0.11	0.41
V	GTC	4.74	4.81	4.45	4.45	4.72	3.91
V	GTG	3.24	3.54	3.54	3.44	3.54	4.14
V	GTT	0.97	0.27	0.47	0.19	0.14	0.40
V	GTA	0.35	0.14	0.31	0.37	0.26	0.33
W	TGG	1.38	1.41	1.44	1.54	1.51	1.50
Y	TAC	1.63	1.57	1.53	1.92	1.95	1.91
Y	TAT	0.49	0.17	0.31	0.20	0.10	0.27
-	TGA	0.20	0.24	0.22	0.22	0.24	0.19
-	TAG	0.06	0.05	0.05	0.06	0.05	0.07
-	TAA	0.03	0.01	0.02	0.02	0.01	0.05

Table S5. Total fraction of IVYWREL amino acids in 478 orthologous proteins from each of the six actinobacteria.

Organism	OGT (°C)	In the whole proteome	In the cytosolic subproteome	In a set of 478 orthologous proteins
<i>Acidothermus cellulolyticus</i> 11B	55	40.56	41.94	41.76
<i>Frankia alni</i> ACN14a	28	38.83	39.96	39.90
<i>Frankia</i> sp. CcI3	27	39.68	40.68	40.32
<i>Streptomyces avermitilis</i> MA-4680	28	38.97	40.49	40.12
<i>Streptomyces coelicolor</i> A3(2)	30	38.83	40.67	40.00
<i>Thermobifida fusca</i> YX	50	41.19	41.86	41.75
R-squared value		0.77	0.89	0.94
p-value is less than		0.02	0.005	0.001

Total fraction of IVYWREL were computed from the amino acid composition in the whole proteome, cytosolic subproteome, and in a set of 478 highly conserved orthologous proteins in the six actinobacteria. A list of the 478 proteins in *A. cellulolyticus* is provided in supplementary Table S8.

The R-squared and p-values were computed for linear regression between the OGT values and the IVYWREL fractions.

The cytosolic subproteome was predicted using the pSORTb software (Gardy et al., 2005; <http://www.psort.org/psortb/>)

Table S6. Total fraction of IVYWREL amino acids in 46 orthologous proteins from forty-five completely sequenced actinobacteria.

Organism	Genome Size (Mb)	%G+C	OGT (°C)	% IVYWREL		
				In the whole proteome	In the cytosolic subproteome	In the 46 orthologs
<i>Acidothermus cellulolyticus</i> 11B	2.40	66.9	55	40.6	41.9	42.4
<i>Arthrobacter aureus</i> TC1	5.23	62.4	30	38.4	39.2	40.2
<i>Arthrobacter</i> sp. FB24	5.08	65.4	30	38.3	39.2	40.1
<i>Bifidobacterium adolescentis</i> ATCC 15703	2.10	59.2	37	37.6	38.4	39.0
<i>Bifidobacterium longum</i>	2.26	60.1	37	37.2	38.2	39.0
<i>Clavibacter michiganensis</i> NCPPB 382	3.40	72.5	28	39.7	40.6	40.3
<i>Corynebacterium diphtheriae</i>	2.49	53.5	37	38.8	39.4	40.5
<i>Corynebacterium efficiens</i> YS-314	3.15	63.1	37	39.5	39.7	40.9
<i>Corynebacterium glutamicum</i> ATCC 13032 Bielefeld	3.30	53.8	33	39.0	39.6	40.7
<i>Corynebacterium glutamicum</i> ATCC 13032 Kitasato	3.30	53.8	33	39.0	39.5	40.7
<i>Corynebacterium glutamicum</i> R	3.35	54.1	33	39.0	39.5	40.7
<i>Corynebacterium jeikeium</i> K411	2.48	61.4	37	38.4	39.2	39.8
<i>Frankia alni</i> ACN14a	7.50	72.8	28	38.8	40.0	41.4
<i>Frankia</i> sp. CeI3	5.40	70.1	27	39.7	40.7	41.7
<i>Kineococcus radiotolerans</i> SRS30216	4.81	74.2	32	40.2	41.0	40.4
<i>Leifsonia xyli</i> subsp. <i>xyli</i> CTCB0	2.58	67.7	29	40.0	41.0	40.4
<i>Mycobacterium avium</i> 104	5.50	69.0	39	38.8	39.7	40.7
<i>Mycobacterium avium</i> subsp. <i>paratuberculosis</i>	4.80	69.3	39	38.8	39.7	40.7
<i>Mycobacterium bovis</i>	4.35	65.6	37	38.1	39.9	40.7
<i>Mycobacterium bovis</i> BCG Pasteur 1173P2	4.40	65.6	37	38.1	39.9	40.7
<i>Mycobacterium gilvum</i> PYR-GCK	5.96	67.7	30	39.0	39.6	40.3
<i>Mycobacterium leprae</i>	3.27	57.8	37	39.6	40.3	41.0
<i>Mycobacterium smegmatis</i> MC2 155	7.00	67.4	37	39.3	39.8	40.6
<i>Mycobacterium</i> sp. JLS	6.00	68.4	30	39.3	39.8	40.4
<i>Mycobacterium</i> sp. KMS	6.22	68.2	30	39.3	39.8	40.5
<i>Mycobacterium</i> sp. MCS	5.92	68.4	30	39.3	39.8	40.5
<i>Mycobacterium tuberculosis</i> CDC1551	4.40	65.6	37	38.2	39.9	40.8
<i>Mycobacterium tuberculosis</i> F11	4.40	65.6	37	38.0	39.9	40.7
<i>Mycobacterium tuberculosis</i> H37Ra	4.40	65.6	37	38.0	39.9	40.7
<i>Mycobacterium tuberculosis</i> H37Rv	4.40	65.6	37	38.0	40.0	40.7
<i>Mycobacterium ulcerans</i> Agy99	5.60	65.5	32	38.4	39.7	40.8
<i>Mycobacterium vanbaalenii</i> PYR-1	6.50	67.8	30	39.0	39.7	40.4
<i>Nocardia farcinica</i> IFM10152	6.29	70.7	37	39.9	40.6	40.6
<i>Nocardioide</i> sp. JS614	5.31	71.4	30	40.3	41.0	40.5
<i>Propionibacterium acnes</i> KPA171202	2.56	60.0	37	38.9	39.5	40.0
<i>Rhodococcus</i> sp. RHA1	9.67	67.0	30	39.4	39.9	40.4
<i>Rubrobacter xylanophilus</i> DSM 9941	3.23	70.5	60	44.9	45.5	45.0
<i>Saccharopolyspora erythraea</i> NRRL 2338	8.20	71.1	28	40.5	41.1	40.9
<i>Salinispora tropica</i> CNB-440	5.20	69.5	28	40.1	41.0	41.2
<i>Streptomyces avermitilis</i> MA-4680	9.12	70.7	28	39.0	40.5	40.4
<i>Streptomyces coelicolor</i> A3(2)	9.09	72.0	30	39.3	40.7	40.3
<i>Symbiobacterium thermophilum</i> IAM14863	3.60	68.7	60	42.8	43.3	42.8
<i>Thermobifida fusca</i> YX	3.60	67.5	50	41.2	41.9	41.6
<i>Tropheryma whippelii</i> TW08 27	0.93	46.3	37	40.0	40.4	40.0
<i>Tropheryma whippelii</i> Twist	0.93	46.3	37	40.0	40.4	39.8
R-squared value				0.31	0.37	0.40
p-value is less than				8.0E-05	9.8E-06	2.7E-06

Total fraction of IVYWREL were computed from the amino acid composition in the whole proteome, cytosolic subproteome, and in a set of 46 conserved orthologous proteins in the actinobacteria. A list of the 46 proteins in *A. cellulolyticus* is provided in supplementary Table S9. The cytosolic subproteome was predicted using the pSORTb software (Gardy et al., 2005; <http://www.psort.org/psortb/>). The R-squared and p-values were computed for linear regression between the OGT values and the IVYWREL fractions. A p-value of less than 0.05 was considered significant.

Table S7. Features of the 409 prokaryotic organisms used in this study.

Organism	OGT (°C)	%G+C of the genome	%G+C of rRNA+tRNAs
<u>(A) Archaea.</u>			
<i>Aeropyrum pernix</i>	90.0	56.31	71.01
<i>Archaeoglobus fulgidus</i>	85.0	48.58	65.31
<i>Halobacterium sp</i>	37.0	67.91	60.41
<i>Hyperthermus butylicus</i>	99.0	53.74	60.26
<i>Methanobacterium thermoautotrophicum</i>	65.0	49.54	58.21
<i>Methanococcoides burtonii</i> DSM 6242	21.5	40.76	53.19
<i>Methanococcus jannaschii</i>	85.0	31.43	64.47
<i>Methanococcus maripaludis</i> S2	37.0	33.10	55.05
<i>Methanocorpusculum labreanum</i> Z	37.0	50.01	53.13
<i>Methanoculleus marisnigri</i> JR1	27.5	62.06	58.02
<i>Methanopyrus kandleri</i>	98.0	61.16	70.20
<i>Methanosaeta thermophila</i> PT	61.0	53.55	59.21
<i>Methanosarcina acetivorans</i>	38.5	42.68	55.96
<i>Methanosarcina barkeri</i> fusaro	33.5	39.28	55.57
<i>Methanosarcina mazei</i>	37.0	41.48	55.95
<i>Methanospira stadtmanae</i>	37.0	27.63	49.31
<i>Methanospirillum hungatei</i> JF-1	37.0	45.15	54.09
<i>Nanoarchaeum equitans</i>	90.0	31.56	69.03
<i>Natronomonas pharaonis</i>	37.0	63.44	60.81
<i>Picrophilus torridus</i> DSM 9790	55.0	35.97	56.44
<i>Pyrobaculum aerophilum</i>	98.0	51.36	64.83
<i>Pyrobaculum calidifontis</i> JCM 11548	95.0	57.15	69.12
<i>Pyrobaculum islandicum</i> DSM 4184	97.5	49.60	68.73
<i>Pyrococcus abyssi</i>	98.0	44.71	64.50
<i>Pyrococcus furiosus</i>	98.5	40.77	59.96
<i>Pyrococcus horikoshii</i>	95.0	41.88	68.20
<i>Staphylothermus marinus</i> F1	87.5	35.73	67.71
<i>Sulfolobus acidocaldarius</i> DSM 639	70.0	36.71	60.98
<i>Sulfolobus solfataricus</i>	87.0	35.79	63.48
<i>Sulfolobus tokodaii</i>	75.0	32.79	66.36
<i>Thermofilum pendens</i> Hrk 5	88.0	57.67	67.66
<i>Thermoplasma acidophilum</i>	57.5	45.99	56.90
<i>Thermoplasma volcanium</i>	60.0	39.92	56.93
<u>(B) Bacteria.</u>			
<i>Acidobacteria bacterium</i> Ellin345	25.0	58.38	56.90
<i>Acidothermus cellulolyticus</i> 11B	55.0	66.91	62.95
<i>Acidovorax avenae</i> citrulli AAC00-1	25.0	68.53	54.67
<i>Acidovorax</i> JS42	30.0	66.17	54.73
<i>Acinetobacter sp</i> ADP1	30.0	40.43	52.66
<i>Actinobacillus pleuropneumoniae</i> L20	37.0	41.30	51.60
<i>Aeromonas hydrophila</i> ATCC 7966	30.0	61.55	55.06
<i>Alcanivorax borkumensis</i> SK2	28.0	54.73	54.77
<i>Anabaena variabilis</i> ATCC 29413	25.0	41.42	53.26
<i>Anaeromyxobacter dehalogenans</i> 2CP-C	30.0	74.91	58.95
<i>Anaplasma marginale</i> St Maries	37.0	49.76	54.07
<i>Anaplasma phagocytophilum</i> HZ	37.0	41.64	52.50
<i>Aquifex aeolicus</i>	95.0	43.48	65.83
<i>Arthrobacter aureus</i> TC1	30.0	62.34	56.88
<i>Arthrobacter</i> FB24	30.0	65.47	56.84
<i>Aster yellows witches-broom phytoplasma</i> AYWB	25.0	26.89	47.76

<i>Azoarcus</i> BH72	28.0	67.92	56.08
<i>Azoarcus</i> sp EbN1	28.0	65.12	56.62
<i>Bacillus anthracis</i> Ames	30.0	35.38	53.42
<i>Bacillus anthracis</i> Ames 0581	30.0	35.38	53.41
<i>Bacillus anthracis</i> str Sterne	30.0	35.38	53.19
<i>Bacillus cereus</i> ATCC 10987	30.0	35.58	53.42
<i>Bacillus cereus</i> ATCC14579	30.0	35.28	53.23
<i>Bacillus cereus</i> ZK	30.0	35.35	53.15
<i>Bacillus clausii</i> KSM-K16	30.0	44.75	55.52
<i>Bacillus halodurans</i>	30.0	43.69	54.72
<i>Bacillus licheniformis</i> ATCC 14580	37.0	46.19	55.14
<i>Bacillus licheniformis</i> DSM 13	37.0	46.19	55.16
<i>Bacillus subtilis</i>	30.0	43.52	54.74
<i>Bacillus thuringiensis</i> Al Hakam	30.0	35.43	53.08
<i>Bacillus thuringiensis</i> konkukian	30.0	35.41	53.21
<i>Bacteroides fragilis</i> NCTC 9434	37.0	43.19	50.64
<i>Bacteroides fragilis</i> YCH46	37.0	43.27	50.80
<i>Bacteroides thetaiotaomicron</i> VPI-5482	37.0	42.84	50.67
<i>Bartonella bacilliformis</i> KC583	37.0	38.24	54.20
<i>Bartonella henselae</i> Houston-1	37.0	38.23	54.33
<i>Bartonella quintana</i> Toulouse	37.0	38.80	54.00
<i>Baumannia cicadellinicola</i> Homalodisca coagulata	25.0	33.24	48.62
<i>Bdellovibrio bacteriovorus</i>	30.0	50.65	51.83
<i>Bifidobacterium adolescentis</i> ATCC 15703	37.0	59.18	59.48
<i>Bifidobacterium longum</i>	37.0	60.12	59.49
<i>Bordetella bronchiseptica</i>	37.0	68.08	55.92
<i>Bordetella parapertussis</i>	37.0	68.10	55.86
<i>Bordetella pertussis</i>	37.0	67.72	55.78
<i>Borrelia afzelii</i> PKo	37.0	28.31	45.86
<i>Borrelia burgdorferi</i>	35.0	28.59	46.37
<i>Borrelia garinii</i> PBi	37.0	28.30	47.05
<i>Bradyrhizobium japonicum</i>	26.0	64.06	57.95
<i>Buchnera aphidicola</i>	25.0	25.34	47.37
<i>Buchnera aphidicola</i> Cc <i>Cinara cedri</i>	25.0	20.10	45.86
<i>Buchnera aphidicola</i> Sg	25.0	25.33	48.90
<i>Buchnera</i> sp	25.0	26.31	49.58
<i>Campylobacter fetus</i> 82-40	37.0	33.31	48.73
<i>Campylobacter jejuni</i>	37.0	30.55	49.46
<i>Campylobacter jejuni</i> 81-176	37.0	30.62	49.26
<i>Campylobacter jejuni</i> RM1221	37.0	30.31	48.81
<i>Candidatus</i> Blochmannia floridanus	25.0	27.38	46.56
<i>Candidatus</i> Blochmannia pennsylvanicus BPEN	25.0	29.56	48.37
<i>Candidatus</i> Pelagibacter ubique HTCC1062	30.0	29.68	49.87
<i>Carboxydotherrmus hydrogenoformans</i> Z-2901	67.5	42.05	59.36
<i>Caulobacter crescentus</i>	30.0	67.21	56.93
<i>Chlamydia muridarum</i>	37.0	40.34	50.69
<i>Chlamydia trachomatis</i>	37.0	41.31	50.52
<i>Chlamydia trachomatis</i> A HAR-13	37.0	41.30	50.71
<i>Chlamydomphila abortus</i> S26 3	37.0	39.87	50.12
<i>Chlamydomphila caviae</i>	37.0	39.22	50.67
<i>Chlamydomphila felis</i> Fe C-56	37.0	39.38	50.60
<i>Chlamydomphila pneumoniae</i> AR39	37.0	40.57	50.44
<i>Chlamydomphila pneumoniae</i> CWL029	37.0	40.58	49.69
<i>Chlamydomphila pneumoniae</i> J138	37.0	40.58	50.27
<i>Chlamydomphila pneumoniae</i> TW 183	37.0	40.58	50.18

<i>Chlorobium chlorochromatii</i> CaD3	37.0	44.28	52.96
<i>Chlorobium phaeobacteroides</i> DSM 266	25.0	48.35	52.61
<i>Chlorobium tepidum</i> TLS	47.0	56.53	53.67
<i>Chromobacterium violaceum</i>	26.0	64.83	54.52
<i>Chromohalobacter salexigens</i> DSM 3043	30.0	63.91	57.88
<i>Clostridium acetobutylicum</i>	37.0	30.93	50.93
<i>Clostridium novyi</i> NT	37.0	28.86	51.63
<i>Clostridium perfringens</i>	37.0	28.57	52.44
<i>Clostridium perfringens</i> ATCC 13124	37.0	28.38	52.42
<i>Clostridium tetani</i> E88	37.0	28.75	51.83
<i>Clostridium thermocellum</i> ATCC 27405	57.5	38.99	54.12
<i>Colwellia psychrerythraea</i> 34H	10.0	38.01	51.87
<i>Corynebacterium diphtheriae</i>	37.0	53.48	55.01
<i>Corynebacterium efficiens</i> YS-314	37.0	63.14	55.68
<i>Corynebacterium glutamicum</i> ATCC 13032 Bielefeld	33.0	53.84	54.90
<i>Corynebacterium glutamicum</i> ATCC 13032 Kitasato	33.0	53.81	55.07
<i>Corynebacterium jeikeium</i> K411	37.0	61.40	56.50
<i>Coxiella burnetii</i>	37.0	42.66	57.22
<i>Cyanobacteria bacterium</i> Yellowstone A-Prime	27.0	60.24	58.38
<i>Cyanobacteria bacterium</i> Yellowstone B-Prime	27.0	58.45	58.51
<i>Cytophaga hutchinsonii</i> ATCC 33406	30.0	38.85	50.80
<i>Dechloromonas aromatica</i> RCB	30.0	59.25	54.83
<i>Dehalococcoides</i> CBDB1	35.0	47.03	55.93
<i>Dehalococcoides ethenogenes</i> 195	35.0	48.85	55.82
<i>Deinococcus geothermalis</i> DSM 11300	47.5	66.64	59.42
<i>Desulfotobacterium hafniense</i> Y51	37.0	47.36	54.22
<i>Desulfotalea psychrophila</i> LSV54	7.0	46.81	52.28
<i>Desulfovibrio desulfuricans</i> G20	55.0	57.84	55.80
<i>Desulfovibrio vulgaris</i> DP4	30.0	63.01	56.37
<i>Desulfovibrio vulgaris</i> Hildenborough	30.0	63.14	56.54
<i>Ehrlichia canis</i> Jake	37.0	28.96	49.09
<i>Ehrlichia chaffeensis</i> Arkansas	37.0	30.10	48.90
<i>Ehrlichia ruminantium</i> Gardel	37.0	27.51	49.14
<i>Enterococcus faecalis</i> V583	37.0	37.53	52.72
<i>Erwinia carotovora</i> atroseptica SCRI1043	28.0	50.97	53.86
<i>Erythrobacter litoralis</i> HTCC2594	28.0	63.07	58.41
<i>Escherichia coli</i> 536	37.0	50.52	54.82
<i>Escherichia coli</i> APEC O1	37.0	50.55	54.67
<i>Escherichia coli</i> CFT073	37.0	50.48	54.68
<i>Escherichia coli</i> O157H7	37.0	50.54	54.83
<i>Escherichia coli</i> O157H7 EDL933	37.0	50.44	54.85
<i>Escherichia coli</i> UTI89	37.0	50.60	54.77
<i>Escherichia coli</i> W3110	37.0	50.80	53.96
<i>Francisella tularensis</i> FSC 198	37.0	32.26	51.03
<i>Francisella tularensis</i> holarctica	37.0	32.15	50.51
<i>Francisella tularensis</i> holarctica OSU18	37.0	32.16	50.71
<i>Francisella tularensis</i> novicida UI12	37.0	32.48	51.15
<i>Francisella tularensis</i> tularensis	37.0	32.26	51.04
<i>Frankia alni</i> ACN14a	28.0	72.83	61.30
<i>Frankia</i> sp. CcI3	27.0	70.08	61.47
<i>Fusobacterium nucleatum</i>	37.0	27.15	48.49
<i>Geobacillus kaustophilus</i> HTA426	55.0	52.09	58.35
<i>Geobacter metallireducens</i> GS-15	30.0	59.51	55.89
<i>Geobacter sulfurreducens</i>	35.0	60.94	56.79
<i>Gloeobacter violaceus</i>	27.5	62.00	57.72

<i>Gluconobacter oxydans</i> 621H	26.0	61.07	56.52
<i>Granulobacter bethesdensis</i> CGDNIH1	37.0	59.07	57.82
<i>Haemophilus ducreyi</i> 35000HP	28.0	38.22	51.40
<i>Haemophilus influenzae</i>	37.0	38.15	51.58
<i>Haemophilus influenzae</i> 86 028NP	37.0	38.16	51.57
<i>Haemophilus somnus</i> 129PT	37.0	37.20	51.65
<i>Hahella chejuensis</i> KCTC 2396	28.0	53.87	55.00
<i>Halorhodospira halophila</i> SL1	25.0	67.98	59.71
<i>Helicobacter acinonychis</i> Sheeba	37.0	38.18	51.12
<i>Helicobacter hepaticus</i>	37.0	35.93	52.52
<i>Helicobacter pylori</i> 26695	37.0	38.87	50.64
<i>Helicobacter pylori</i> HPAG1	37.0	39.08	50.97
<i>Helicobacter pylori</i> J99	37.0	39.19	51.15
<i>Hyphomonas neptunium</i> ATCC 15444	26.0	61.93	56.46
<i>Idiomarina loihiensis</i> L2TR	30.0	47.04	54.48
<i>Jannaschia</i> CCS1	25.0	62.33	57.11
<i>Lactobacillus acidophilus</i> NCFM	37.0	34.71	51.59
<i>Lactobacillus brevis</i> ATCC 367	30.0	46.22	52.33
<i>Lactobacillus casei</i> ATCC 334	30.0	46.62	52.94
<i>Lactobacillus delbrueckii</i> bulgaricus	30.0	49.72	53.53
<i>Lactobacillus gasseri</i> ATCC 33323	37.0	35.26	51.36
<i>Lactobacillus johnsonii</i> NCC 533	37.0	34.61	50.93
<i>Lactobacillus plantarum</i>	30.0	44.47	51.70
<i>Lactobacillus sakei</i> 23K	30.0	41.26	51.20
<i>Lactobacillus salivarius</i> UCC118	37.0	32.94	51.21
<i>Lactococcus lactis</i>	30.0	35.33	51.18
<i>Lactococcus lactis</i> cremoris MG1363	30.0	35.75	51.10
<i>Lactococcus lactis</i> cremoris SK11	30.0	35.86	51.01
<i>Lawsonia intracellularis</i> PHE MN1-00	37.0	33.28	53.02
<i>Legionella pneumophila</i> Lens	37.0	38.42	53.67
<i>Legionella pneumophila</i> Paris	37.0	38.37	53.78
<i>Legionella pneumophila</i> Philadelphia 1	37.0	38.27	53.72
<i>Leifsonia xyli</i> xyli CTCB0	29.0	67.68	58.12
<i>Listeria innocua</i>	37.0	37.44	53.64
<i>Listeria monocytogenes</i>	37.0	37.98	53.58
<i>Listeria monocytogenes</i> 4b F2365	37.0	38.04	53.42
<i>Listeria welshimeri</i> serovar 6b SLCC5334	37.0	36.35	53.55
<i>Magnetococcus</i> MC-1	25.0	54.17	55.01
<i>Magnetospirillum magneticum</i> AMB-1	30.0	65.09	56.94
<i>Mannheimia succiniciproducens</i> MBEL55E	37.0	42.54	52.75
<i>Maricaulis maris</i> MCS10	26.0	62.73	57.11
<i>Marinobacter aquaeolei</i> VT8	30.0	57.27	56.26
<i>Mesoplasma florum</i> L1	30.0	27.02	50.05
<i>Mesorhizobium</i> BNC1	26.0	61.07	58.14
<i>Mesorhizobium loti</i>	26.0	62.75	56.56
<i>Methylibium petroleiphilum</i> PMI	30.0	69.20	56.45
<i>Methylobacillus flagellatus</i> KT	30.0	55.72	54.95
<i>Methylococcus capsulatus</i> Bath	37.0	63.58	56.88
<i>Moorella thermoacetica</i> ATCC 39073	55.0	55.79	58.53
<i>Mycobacterium avium</i> 104	39.0	68.99	59.41
<i>Mycobacterium avium</i> paratuberculosis	39.0	69.30	59.12
<i>Mycobacterium bovis</i>	37.0	65.63	59.71
<i>Mycobacterium bovis</i> BCG Pasteur 1173P2	37.0	65.64	59.72
<i>Mycobacterium</i> JLS	30.0	68.36	58.83
<i>Mycobacterium</i> KMS	30.0	68.44	58.82

<i>Mycobacterium leprae</i>	37.0	57.80	58.40
<i>Mycobacterium MCS</i>	30.0	68.45	59.25
<i>Mycobacterium smegmatis MC2 155</i>	37.0	67.40	58.55
<i>Mycobacterium tuberculosis CDC1551</i>	37.0	65.61	59.43
<i>Mycobacterium tuberculosis H37Rv</i>	37.0	65.61	59.71
<i>Mycobacterium ulcerans Agy99</i>	32.0	65.47	59.23
<i>Mycobacterium vanbaalenii PYR-1</i>	30.0	67.79	58.93
<i>Mycoplasma capricolum ATCC 27343</i>	37.0	23.77	46.26
<i>Mycoplasma gallisepticum</i>	37.0	31.45	47.32
<i>Mycoplasma genitalium</i>	37.0	31.69	46.63
<i>Mycoplasma hyopneumoniae 232</i>	37.0	28.56	45.26
<i>Mycoplasma hyopneumoniae 7448</i>	37.0	28.49	47.38
<i>Mycoplasma hyopneumoniae J</i>	37.0	28.52	47.33
<i>Mycoplasma mobile 163K</i>	37.0	24.95	46.91
<i>Mycoplasma mycoides</i>	37.0	23.97	48.37
<i>Mycoplasma penetrans</i>	37.0	25.72	47.86
<i>Mycoplasma pneumoniae</i>	37.0	40.01	47.16
<i>Mycoplasma pulmonis</i>	37.0	26.64	48.87
<i>Mycoplasma synoviae 53</i>	37.0	28.50	46.92
<i>Myxococcus xanthus DK 1622</i>	30.0	68.89	56.95
<i>Neisseria gonorrhoeae FA 1090</i>	37.0	52.69	54.19
<i>Neisseria meningitidis FAM18</i>	37.0	51.62	54.03
<i>Neisseria meningitidis MC58</i>	37.0	51.53	54.03
<i>Neisseria meningitidis Z2491</i>	37.0	51.81	53.90
<i>Neorickettsia sennetsu Miyayama</i>	37.0	41.08	52.31
<i>Nitrobacter hamburgensis X14</i>	28.0	61.71	58.19
<i>Nitrobacter winogradskyi Nb-255</i>	28.0	62.05	57.69
<i>Nitrosococcus oceani ATCC 19707</i>	27.5	50.32	56.97
<i>Nitrosomonas europaea</i>	26.0	50.72	51.55
<i>Nitrosomonas eutropha C71</i>	26.0	48.49	54.22
<i>Nitrospira multiformis ATCC 25196</i>	26.0	53.94	56.40
<i>Nocardia farcinica IFM10152</i>	37.0	70.83	58.45
<i>Nocardioides JS614</i>	30.0	71.65	59.24
<i>Nostoc sp</i>	30.0	41.35	53.33
<i>Novosphingobium aromaticivorans DSM 12444</i>	30.0	65.15	56.80
<i>Oceanobacillus iheyensis</i>	28.0	35.68	53.63
<i>Oenococcus oeni PSU-1</i>	30.0	37.89	50.61
<i>Onion yellows phytoplasma</i>	25.0	27.74	46.99
<i>Parachlamydia sp UWE25</i>	34.5	34.72	50.85
<i>Pasteurella multocida</i>	37.0	40.40	51.53
<i>Pediococcus pentosaceus ATCC 25745</i>	30.0	37.36	51.57
<i>Pelobacter carbinolicus</i>	30.0	55.11	56.77
<i>Pelobacter propionicus DSM 2379</i>	30.0	59.02	55.15
<i>Pelodictyon luteolum DSM 273</i>	25.0	57.33	53.30
<i>Photorhabdus luminescens</i>	28.0	42.83	55.37
<i>Pirellula sp</i>	30.0	55.40	56.06
<i>Polaromonas JS666</i>	25.0	62.47	55.75
<i>Polaromonas naphthalenivorans CJ2</i>	20.0	62.53	54.43
<i>Porphyromonas gingivalis W83</i>	37.0	48.29	52.18
<i>Prochlorococcus marinus AS9601</i>	30.0	31.32	55.17
<i>Prochlorococcus marinus CCMP1375</i>	30.0	36.44	55.48
<i>Prochlorococcus marinus MED4</i>	30.0	30.80	53.11
<i>Prochlorococcus marinus MIT 9303</i>	30.0	50.01	56.12
<i>Prochlorococcus marinus MIT 9312</i>	30.0	31.21	54.83
<i>Prochlorococcus marinus MIT 9515</i>	30.0	30.79	54.97

<i>Prochlorococcus marinus</i> MIT9313	30.0	50.74	54.64
<i>Prochlorococcus marinus</i> NATL1A	30.0	34.98	55.46
<i>Prochlorococcus marinus</i> NATL2A	30.0	35.12	55.34
<i>Propionibacterium acnes</i> KPA171202	37.0	60.01	57.59
<i>Pseudoalteromonas atlantica</i> T6c	20.0	44.62	52.23
<i>Pseudomonas aeruginosa</i>	30.0	66.56	55.35
<i>Pseudomonas aeruginosa</i> UCBPP-PA14	30.0	66.29	54.47
<i>Pseudomonas entomophila</i> L48	30.0	64.16	54.13
<i>Pseudomonas fluorescens</i> Pf-5	30.0	63.30	53.85
<i>Pseudomonas putida</i> KT2440	26.0	61.52	53.61
<i>Pseudomonas syringae</i> phaseolicola 1448A	26.0	58.02	54.01
<i>Pseudomonas syringae</i> pv B728a	26.0	59.23	54.08
<i>Pseudomonas syringae</i> tomato DC3000	26.0	58.40	54.03
<i>Psychrobacter arcticum</i> 273-4	10.0	42.80	52.08
<i>Psychrobacter cryohalolentis</i> K5	23.0	42.29	52.06
<i>Psychromonas ingrahamii</i> 37	10.0	40.09	51.63
<i>Ralstonia solanacearum</i>	28.0	67.04	54.40
<i>Rhizobium etli</i> CFN 42	30.0	61.27	55.79
<i>Rhizobium leguminosarum</i> bv viciae 3841	26.0	61.09	55.66
<i>Rhodococcus</i> RHA1	30.0	67.52	57.46
<i>Rhodoferax ferrireducens</i> T118	25.0	59.88	54.53
<i>Rhodopseudomonas palustris</i> BisA53	25.0	64.44	56.73
<i>Rhodopseudomonas palustris</i> BisB18	25.0	64.96	57.29
<i>Rhodopseudomonas palustris</i> BisB5	25.0	64.81	58.00
<i>Rhodopseudomonas palustris</i> CGA009	25.0	65.04	55.92
<i>Rhodopseudomonas palustris</i> HaA2	25.0	66.04	58.41
<i>Rhodospirillum rubrum</i> ATCC 11170	25.0	65.45	57.60
<i>Rickettsia bellii</i> RML369-C	37.0	31.65	50.87
<i>Rickettsia conorii</i>	37.0	32.44	50.47
<i>Rickettsia felis</i> URRWXC12	37.0	32.45	50.46
<i>Rickettsia prowazekii</i>	37.0	29.00	50.27
<i>Rickettsia typhi</i> wilmsington	37.0	28.92	49.62
<i>Roseobacter denitrificans</i> OCh 114	20.0	58.97	55.95
<i>Rubrobacter xylanophilus</i> DSM 9941	60.0	70.48	64.37
<i>Saccharophagus degradans</i> 2-40	28.0	45.83	54.26
<i>Salinibacter ruber</i> DSM 13855	37.0	66.22	59.15
<i>Salmonella enterica</i> Choleraesuis	37.0	52.16	55.10
<i>Salmonella enterica</i> Paratyphi ATCC 9150	37.0	52.16	54.42
<i>Salmonella typhi</i>	37.0	52.09	54.75
<i>Salmonella typhimurium</i> LT2	37.0	52.22	54.73
<i>Shewanella amazonensis</i> SB2B	37.0	53.59	54.24
<i>Shewanella</i> ANA-3	30.0	48.05	53.62
<i>Shewanella baltica</i> OS155	25.0	46.28	53.51
<i>Shewanella denitrificans</i> OS217	28.0	45.15	53.16
<i>Shewanella frigidimarina</i> NCIMB 400	20.0	41.58	52.66
<i>Shewanella</i> MR-4	30.0	47.89	53.60
<i>Shewanella</i> MR-7	30.0	47.87	53.63
<i>Shewanella oneidensis</i>	30.0	45.96	53.55
<i>Shewanella</i> W3-18-1	30.0	44.63	53.70
<i>Shigella boydii</i> Sb227	37.0	51.21	54.89
<i>Shigella dysenteriae</i>	37.0	51.25	54.71
<i>Shigella flexneri</i> 2a	37.0	50.89	54.79
<i>Shigella flexneri</i> 2a 2457T	37.0	50.91	54.48
<i>Shigella flexneri</i> 5 8401	37.0	50.92	54.87
<i>Shigella sonnei</i> Ss046	37.0	51.01	54.98

<i>Silicibacter pomeroyi</i> DSS-3	30.0	64.22	56.63
<i>Silicibacter</i> TM1040	30.0	60.41	57.60
<i>Sinorhizobium meliloti</i>	26.0	62.73	56.40
<i>Sodalis glossinidius morsitans</i>	28.0	54.70	55.24
<i>Solibacter usitatus</i> Ellin6076	30.0	61.90	56.70
<i>Sphingopyxis alaskensis</i> RB2256	25.0	65.50	57.45
<i>Staphylococcus aureus aureus</i> MRSA252	37.0	32.81	49.36
<i>Staphylococcus aureus aureus</i> MSSA476	37.0	32.85	49.46
<i>Staphylococcus aureus</i> COL	37.0	32.82	51.22
<i>Staphylococcus aureus</i> Mu50	37.0	32.88	51.43
<i>Staphylococcus aureus</i> MW2	37.0	32.83	51.34
<i>Staphylococcus aureus</i> N315	37.0	32.84	51.54
<i>Staphylococcus aureus</i> NCTC 8325	37.0	32.87	51.53
<i>Staphylococcus aureus</i> RF122	37.0	32.78	51.54
<i>Staphylococcus aureus</i> USA300	37.0	32.75	51.36
<i>Staphylococcus epidermidis</i> ATCC 12228	37.0	32.10	52.08
<i>Staphylococcus epidermidis</i> RP62A	37.0	32.15	51.52
<i>Staphylococcus haemolyticus</i>	37.0	32.79	51.79
<i>Staphylococcus saprophyticus</i>	28.0	33.24	51.42
<i>Streptococcus agalactiae</i> 2603	37.0	35.65	51.40
<i>Streptococcus agalactiae</i> A909	37.0	35.62	51.28
<i>Streptococcus agalactiae</i> NEM316	37.0	35.63	51.37
<i>Streptococcus mutans</i>	37.0	36.83	52.94
<i>Streptococcus pneumoniae</i> D39	37.0	39.71	51.84
<i>Streptococcus pneumoniae</i> R6	37.0	39.72	51.82
<i>Streptococcus pneumoniae</i> TIGR4	37.0	39.70	52.01
<i>Streptococcus pyogenes</i> M1 GAS	37.0	38.51	51.64
<i>Streptococcus pyogenes</i> MGAS10270	37.0	38.43	51.42
<i>Streptococcus pyogenes</i> MGAS10394	37.0	38.69	51.68
<i>Streptococcus pyogenes</i> MGAS10750	37.0	38.32	51.43
<i>Streptococcus pyogenes</i> MGAS2096	37.0	38.73	51.44
<i>Streptococcus pyogenes</i> MGAS315	37.0	38.59	51.51
<i>Streptococcus pyogenes</i> MGAS5005	37.0	38.53	51.68
<i>Streptococcus pyogenes</i> MGAS6180	37.0	38.35	51.54
<i>Streptococcus pyogenes</i> MGAS8232	37.0	38.55	51.69
<i>Streptococcus pyogenes</i> MGAS9429	37.0	38.54	51.57
<i>Streptococcus pyogenes</i> SSI-1	37.0	38.55	51.69
<i>Streptococcus sanguinis</i> SK36	37.0	43.40	52.60
<i>Streptococcus thermophilus</i> CNRZ1066	37.0	39.08	51.73
<i>Streptococcus thermophilus</i> LMD-9	37.0	39.08	51.59
<i>Streptococcus thermophilus</i> LMG 18311	37.0	39.09	51.73
<i>Streptomyces avermitilis</i> MA-4680	28.0	70.72	58.10
<i>Streptomyces coelicolor</i> A3(2)	30.0	72.12	58.59
<i>Symbiobacterium thermophilum</i> IAM14863	60.0	68.67	60.24
<i>Synechococcus</i> CC9311	27.5	52.45	55.86
<i>Synechococcus</i> CC9605	27.5	59.22	56.30
<i>Synechococcus</i> CC9902	27.5	54.16	55.78
<i>Synechococcus elongatus</i> PCC 6301	27.5	55.48	55.74
<i>Synechococcus elongatus</i> PCC 7942	27.5	55.47	55.70
<i>Synechococcus</i> sp WH8102	27.5	59.41	56.37
<i>Synechocystis</i> PCC6803	29.5	47.72	53.98
<i>Syntrophobacter fumaroxidans</i> MPOB	37.0	59.95	58.30
<i>Syntrophus aciditrophicus</i> SB	35.0	51.46	55.83
<i>Thermoanaerobacter tengcongensis</i>	75.0	37.57	59.12
<i>Thermobifida fusca</i> YX	50.0	67.50	60.00

<i>Thermosynechococcus elongatus</i>	55.0	53.92	55.94
<i>Thermotoga maritima</i>	80.0	46.25	63.85
<i>Thermus thermophilus</i> HB27	75.0	69.44	64.53
<i>Thermus thermophilus</i> HB8	75.0	69.52	64.68
<i>Thiobacillus denitrificans</i> ATCC 25259	30.0	66.07	55.77
<i>Thiomicrospira crunogena</i> XCL-2	25.0	43.13	52.20
<i>Thiomicrospira denitrificans</i> ATCC 33889	22.0	34.46	50.33
<i>Treponema denticola</i> ATCC 35405	37.0	37.87	52.17
<i>Treponema pallidum</i>	35.0	52.78	54.27
<i>Trichodesmium erythraeum</i> IMS101	37.0	34.14	53.19
<i>Tropheryma whipplei</i> TW08 27	37.0	46.31	57.87
<i>Tropheryma whipplei</i> Twist	37.0	46.33	57.75
<i>Ureaplasma urealyticum</i>	37.0	25.50	45.80
<i>Wigglesworthia brevialpilis</i>	27.5	22.48	48.34
<i>Wolbachia endosymbiont of Brugia malayi</i> TRS	27.5	34.18	48.40
<i>Wolbachia endosymbiont of Drosophila melanogaster</i>	27.5	35.23	48.73
<i>Wolinella succinogenes</i>	37.0	48.46	52.24
<i>Xanthomonas campestris</i> 8004	26.0	64.96	56.13
<i>Xanthomonas campestris vesicatoria</i> 85-10	26.0	64.75	56.18
<i>Xanthomonas citri</i>	26.0	64.77	56.55
<i>Xanthomonas oryzae</i> KACC10331	26.0	63.69	56.29
<i>Xanthomonas oryzae</i> MAFF 311018	26.0	63.70	56.22
<i>Xylella fastidiosa</i>	26.0	52.67	55.16
<i>Xylella fastidiosa</i> Temecula1	26.0	51.78	55.37
<i>Yersinia pestis</i> Antiqua	37.0	47.70	53.75
<i>Yersinia pestis</i> CO92	37.0	47.64	53.90
<i>Yersinia pestis</i> KIM	37.0	47.64	54.11
<i>Yersinia pestis</i> Nepal516	37.0	47.58	53.78
<i>Yersinia pseudotuberculosis</i> IP32953	37.0	47.61	53.75
<i>Zymomonas mobilis</i> ZM4	28.0	46.33	54.23

Table S8. The 478 *A. cellulolyticus* 11B homologs used in the analysis presented in Table S5.

GI	Product
117927212	chromosomal replication initiator protein DnaA
117927214	recombination protein F
117927217	DNA gyrase subunit A
117927234	FHA domain-containing protein
117927239	hypothetical protein Accl_0028
117927270	thiosulfate sulfurtransferase
117927273	glycine cleavage T protein (aminomethyl transferase)
117927284	glycosyl transferase, group 1
117927286	phosphoglycerate mutase
117927289	CarD family transcriptional regulator
117927292	2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
117927293	cysteinyI-tRNA synthetase
117927294	RNA methyltransferase
117927301	phosphate uptake regulator, PhoU
117927303	phosphate ABC transporter, inner membrane subunit PstC
117927304	phosphate ABC transporter, inner membrane subunit PstA
117927305	phosphate ABC transporter, ATPase subunit
117927311	threonine synthase
117927315	cold-shock DNA-binding protein family protein
117927320	MarR family transcriptional regulator
117927325	phosphoserine aminotransferase
117927326	citrate synthase 2
117927327	pyridoxamine 5'-phosphate oxidase
117927365	hypothetical protein Accl_0155
117927376	UDP-glucose pyrophosphorylase
117927377	molybdopterin molybdochelataase
117927378	GTP cyclohydrolase subunit MoaC
117927379	molybdopterin adenylyltransferase
117927386	TatD family hydrolase
117927388	dimethyladenosine transferase
117927391	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase
117927408	inorganic diphosphatase
117927410	hypothetical protein Accl_0201
117927412	hypoxanthine phosphoribosyltransferase
117927413	Mername-AA223 peptidase
117927414	GTP cyclohydrolase I
117927416	dihydropteroate synthase
117927420	pantothenate synthetase
117927422	L-aspartate oxidase
117927423	nicotinate-nucleotide pyrophosphorylase
117927424	pantothenate kinase
117927430	HhH-GPD family protein
117927437	DNA repair protein RadA
117927440	glutamate-1-semialdehyde 2,1-aminomutase
117927444	delta-aminolevulinic acid dehydratase
117927445	uroporphyrinogen-III synthase / uroporphyrinogen-III C-methyltransferase
117927448	redox-sensing transcriptional repressor Rex
117927455	histone deacetylase superfamily protein
117927456	pyrroline-5-carboxylate reductase
117927457	L-proline dehydrogenase
117927458	xylose isomerase domain-containing protein
117927459	phosphoglycerate mutase
117927475	geranylgeranyl reductase
117927476	NADH dehydrogenase subunit A
117927478	NADH dehydrogenase subunit C

117927479	NADH dehydrogenase subunit D
117927480	NADH dehydrogenase subunit E
117927481	NADH-quinone oxidoreductase, F subunit
117927482	NADH dehydrogenase subunit G
117927483	NADH dehydrogenase subunit H
117927484	NADH dehydrogenase subunit I
117927485	NADH dehydrogenase subunit J
117927487	NADH dehydrogenase subunit L
117927488	proton-translocating NADH-quinone oxidoreductase, chain M
117927489	NADH dehydrogenase subunit N
117927490	trans-hexaprenyltranstransferase
117927495	nucleotide-binding protein
117927498	dehydratase
117927501	aspartate aminotransferase
117927503	transcription antitermination protein nusG
117927505	50S ribosomal protein L1P
117927506	50S ribosomal protein L10
117927507	50S ribosomal protein L12P
117927508	DNA-directed RNA polymerase subunit beta
117927509	DNA-directed RNA polymerase subunit beta'
117927510	30S ribosomal protein S12
117927511	30S ribosomal protein S7
117927514	30S ribosomal protein S10
117927515	50S ribosomal protein L3P
117927516	50S ribosomal protein L4P
117927517	50S ribosomal protein L23P
117927518	50S ribosomal protein L2
117927519	SSU ribosomal protein S19P
117927521	30S ribosomal protein S3
117927522	50S ribosomal protein L16
117927523	50S ribosomal protein L29P
117927524	SSU ribosomal protein S17P
117927525	50S ribosomal protein L14
117927526	50S ribosomal protein L24P
117927527	50S ribosomal protein L5
117927528	30S ribosomal protein S14
117927529	SSU ribosomal protein S8P
117927530	50S ribosomal protein L6P
117927531	50S ribosomal protein L18P
117927532	SSU ribosomal protein S5P
117927533	50S ribosomal protein L30P
117927534	50S ribosomal protein L15P
117927535	preprotein translocase subunit SecY
117927536	adenylate kinase
117927539	50S ribosomal protein L36
117927540	30S ribosomal protein S13
117927541	30S ribosomal protein S11
117927544	50S ribosomal protein L17P
117927545	tRNA pseudouridine synthase A
117927550	LSU ribosomal protein L13P
117927551	SSU ribosomal protein S9P
117927552	phosphoglucosamine mutase
117927553	D-fructose-6-phosphate amidotransferase
117927554	carbohydrate kinase, YjeF related protein
117927564	alanine racemase
117927566	hypothetical protein Acel_0357
117927569	O-sialoglycoprotein endopeptidase
117927571	chaperonin Cpn10

117927575	inosine-5'-monophosphate dehydrogenase
117927576	inositol-5-monophosphate dehydrogenase
117927580	bifunctional GMP synthase/glutamine amidotransferase protein
117927584	ATP-dependent DNA helicase PcrA
117927592	phosphoribosylglycinamide formyltransferase
117927593	bifunctional phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase
117927594	methenyltetrahydrofolate cyclohydrolase
117927603	flavoprotein disulfide reductase
117927618	acyl-CoA dehydrogenase domain-containing protein
117927664	LPPG:FO 2-phospho-L-lactate transferase
117927672	hypothetical protein Acel_0463
117927674	phosphomannomutase
117927677	S-adenosyl-L-homocysteine hydrolase
117927679	hypothetical protein Acel_0470
117927681	molybdopterin synthase subunit MoaE
117927729	putative phosphoketolase
117927741	histidinol-phosphate phosphatase
117927742	GTPase EngC
117927743	3-phosphoshikimate 1-carboxyvinyltransferase
117927747	anti-sigma factor
117927767	signal transduction histidine kinase
117927792	alcohol dehydrogenase
117927793	hypothetical protein Acel_0585
117927794	alpha-ketoglutarate decarboxylase
117927800	LamB/YcsF family protein
117927838	homoserine dehydrogenase
117927839	threonine synthase
117927841	transcription termination factor Rho
117927848	50S ribosomal protein L31P
117927849	peptide chain release factor 1
117927850	HemK family modification methylase
117927851	translation factor SUA5
117927852	glycosyl transferase family protein
117927859	F0F1 ATP synthase subunit alpha
117927860	ATP synthase F1, gamma subunit
117927861	F0F1 ATP synthase subunit beta
117927864	ATP:cob(I)alamin adenosyltransferase
117927872	hypothetical protein Acel_0664
117927880	acetyl-CoA acetyltransferase
117927888	alpha-glucan phosphorylases
117927890	short chain enoyl-CoA hydratase
117927892	electron transfer flavoprotein beta-subunit
117927893	electron transfer flavoprotein, alpha subunit
117927896	aminotransferase, class V
117927897	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase
117927901	methionine synthase, vitamin-B12 independent
117927902	DNA ligase, NAD-dependent
117927904	aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit C
117927905	aspartyl/glutamyl-tRNA amidotransferase subunit A
117927906	aspartyl/glutamyl-tRNA amidotransferase subunit B
117927914	acetolactate synthase 1 catalytic subunit
117927915	acetolactate synthase 3 regulatory subunit
117927917	D-3-phosphoglycerate dehydrogenase
117927918	3-isopropylmalate dehydrogenase
117927920	alpha-isopropylmalate/homocitrate synthase family transferase
117927921	5-carboxymethyl-2-hydroxymuconate delta-isomerase
117927944	trigger factor
117927947	ATP-dependent protease ATP-binding subunit

117927955	nucleoside diphosphate kinase
117927961	radical SAM domain-containing protein
117927962	hypothetical protein Acel_0754
117927963	ribonuclease G
117927965	50S ribosomal protein L27
117927966	GTPase ObgE
117927967	gamma-glutamyl kinase
117927969	gamma-glutamyl phosphate reductase
117927970	nicotinate (nicotinamide) nucleotide adenylyltransferase
117927972	iojap-like protein
117927977	FO synthase
117927986	GTP-binding protein LepA
117927989	coproporphyrinogen III oxidase, anaerobic
117927991	heat-inducible transcription repressor
117927992	chaperone protein DnaJ
117927997	PhoH family protein
117927998	putative metalloprotease
117927999	CBS domain-containing protein
117928000	hypothetical protein Acel_0792
117928001	GTP-binding protein Era
117928002	2-isopropylmalate synthase
117928003	DNA repair protein RecO
117928014	DNA primase
117928084	acyl carrier protein
117928086	(acyl-carrier-protein) S-malonyltransferase-like
117928090	hypothetical protein Acel_0882
117928109	hypothetical protein Acel_0901
117928120	PBP family phospholipid-binding protein
117928125	3-methyl-2-oxobutanoate hydroxymethyltransferase
117928127	L-glutamine synthetase
117928128	(glutamate--ammonia-ligase) adenylyltransferase
117928131	L-glutamine synthetase
117928134	lipoyl synthase
117928145	leucyl aminopeptidase
117928162	iron-sulfur cluster assembly accessory protein
117928163	ribokinase-like domain-containing protein
117928164	cytochrome c oxidase, subunit II
117928166	putative integral membrane protein
117928172	cytochrome c oxidase, subunit III
117928173	response regulator receiver protein
117928195	thiazole synthase
117928198	thiamine-phosphate pyrophosphorylase
117928199	5,10-methylenetetrahydrofolate reductase
117928210	S-adenosyl-methyltransferase MraW
117928213	UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase
117928214	UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase
117928215	phospho-N-acetylmuramoyl-pentapeptide-transferase
117928216	UDP-N-acetylmuramoylalanine--D-glutamate ligase
117928217	cell division protein FtsW
117928218	N-acetylglucosaminyl transferase
117928220	cell division protein FtsZ
117928222	hypothetical protein Acel_1014
117928232	hypothetical protein Acel_1024
117928234	hypothetical protein Acel_1026
117928237	isoleucyl-tRNA synthetase
117928240	ribosomal large subunit pseudouridine synthase D
117928242	hypothetical protein Acel_1034
117928266	histidinol dehydrogenase

117928267	histidinol-phosphate aminotransferase
117928269	imidazole glycerol phosphate synthase subunit HisH
117928270	1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino] imidazole-4-carboxamide isomerase
117928276	imidazoleglycerol phosphate synthase, cyclase subunit
117928277	phosphoribosyl-AMP cyclohydrolase
117928280	indole-3-glycerol phosphate synthase
117928282	tryptophan synthase, alpha chain
117928285	glutamate synthase (NADH) large subunit
117928288	response regulator receiver/ANTAR domain-containing protein
117928295	DNA polymerase I
117928296	30S ribosomal protein S1
117928300	4-aminobutyrate aminotransferase
117928313	excinuclease ABC subunit B
117928315	beta-lactamase domain-containing protein
117928316	excinuclease ABC subunit A
117928317	excinuclease ABC subunit C
117928318	hypothetical protein Acel_1111
117928319	hypothetical protein Acel_1112
117928320	hypothetical protein Acel_1113
117928328	preprotein translocase, SecE subunit
117928329	6-phosphogluconolactonase
117928335	protoheme IX farnesyltransferase
117928341	putative transcriptional regulator
117928342	FeS assembly protein SufB
117928343	FeS assembly protein SufD
117928344	FeS assembly ATPase SufC
117928345	SufS subfamily cysteine desulfurase
117928346	NifU family SUF system FeS assembly protein
117928347	hypothetical protein Acel_1140
117928348	transcriptional regulator
117928355	hypothetical protein Acel_1148
117928356	peptidase U62, modulator of DNA gyrase
117928361	ABC transporter related
117928367	tRNA/rRNA methyltransferase (SpoU)
117928376	phosphoglycerate mutase
117928377	hypothetical protein Acel_1170
117928378	hypothetical protein Acel_1171
117928379	cysteinyl-tRNA synthetase
117928380	hypothetical protein Acel_1173
117928387	RecB family exonuclease
117928389	tRNA (adenine-N(1)-)-methyltransferase
117928391	vesicle-fusing ATPase
117928393	hypothetical protein Acel_1186
117928394	hypothetical protein Acel_1187
117928396	20S proteasome, A and B subunits
117928398	hypothetical protein Acel_1191
117928419	hypothetical protein Acel_1212
117928429	glycine dehydrogenase
117928430	MerR family transcriptional regulator
117928431	hypothetical protein Acel_1224
117928433	FHA domain-containing protein
117928434	glycine cleavage system H protein
117928438	nucleotidyl transferase
117928439	CDP-alcohol phosphatidyltransferase
117928440	small GTP-binding protein
117928443	ribosomal large subunit pseudouridine synthase B
117928444	condensin subunit ScpB
117928445	condensin subunit ScpA

117928447	cobyrinic acid a,c-diamide synthase
117928450	CTP synthetase
117928452	DNA repair protein RecN
117928453	NAD(+) kinase
117928454	hemolysin A
117928457	HAD family hydrolase
117928463	argininosuccinate lyase
117928465	ArgR family transcriptional regulator
117928467	acetylornithine aminotransferase
117928468	acetylglutamate kinase
117928469	bifunctional ornithine acetyltransferase/N-acetylglutamate synthase protein
117928470	N-acetyl-gamma-glutamyl-phosphate reductase
117928471	phenylalanyl-tRNA synthetase subunit beta
117928472	phenylalanyl-tRNA synthetase subunit alpha
117928473	PAS/PAC sensor signal transduction histidine kinase
117928474	50S ribosomal protein L20
117928475	50S ribosomal protein L35P
117928476	translation initiation factor 3
117928478	6,7-dimethyl-8-ribityllumazine synthase
117928479	bifunctional 3,4-dihydroxy-2-butanone 4-phosphate synthase/GTP cyclohydrolase II protein
117928483	ribulose-5-phosphate 3-epimerase
117928484	Fmu (Sun) domain-containing protein
117928485	methionyl-tRNA formyltransferase
117928493	AsnC family transcriptional regulator
117928496	primosomal protein N'
117928498	phosphopantothenate-cysteine ligase / phosphopantothenoylcysteine decarboxylase
117928499	DNA-directed RNA polymerase subunit omega
117928500	guanylate kinase
117928501	putative integration host factor MihF
117928507	dihydroorotase
117928508	aspartate carbamoyltransferase catalytic subunit
117928512	elongation factor P
117928514	3-dehydroquinate synthase
117928515	shikimate kinase
117928516	chorismate synthase
117928536	Holliday junction resolvase YqgF
117928537	alanyl-tRNA synthetase
117928540	recombination factor protein RarA
117928546	adenine phosphoribosyltransferase
117928550	Holliday junction DNA helicase B
117928551	Holliday junction DNA helicase subunit RuvA
117928552	Holliday junction endonuclease RuvC
117928553	hypothetical protein Acel_1346
117928554	glutamine amidotransferase subunit PdxT
117928555	pyridoxine biosynthesis protein
117928560	phosphatidylinositol alpha-mannosyltransferase
117928564	histidine triad (HIT) protein
117928586	methionine-R-sulfoxide reductase
117928587	chlorite dismutase
117928589	uroporphyrinogen decarboxylase
117928598	3'-5' exonuclease
117928601	aconitase
117928605	TrkA domain-containing protein
117928606	TrkA domain-containing protein
117928609	deoxyuridine 5'-triphosphate nucleotidohydrolase Dut
117928612	inositol-phosphate phosphatase
117928614	ferrochelataase
117928676	exodeoxyribonuclease III

117928683	vitamin B12-dependent ribonucleotide reductase
117928684	ribonucleotide reductase regulator NrdR-like
117928686	LexA repressor
117928689	small GTP-binding protein
117928691	tRNA delta(2)-isopentenylpyrophosphate transferase
117928695	RNA modification protein
117928700	hypothetical protein Acel_1493
117928705	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase
117928706	MiaB-like tRNA modifying enzyme YliG
117928708	cell divisionFtsK/SpoIIIE
117928709	beta-lactamase domain-containing protein
117928713	dihydrodipicolinate reductase
117928714	peptidase M16 domain-containing protein
117928715	polynucleotide phosphorylase/polyadenylase
117928716	30S ribosomal protein S15
117928719	tRNA pseudouridine synthase B
117928720	ribosome-binding factor A
117928721	hypothetical protein Acel_1514
117928722	translation initiation factor IF-2
117928724	transcription elongation factor NusA
117928728	GCN5-related N-acetyltransferase
117928731	1-deoxy-D-xylulose 5-phosphate reductoisomerase
117928743	radical SAM protein
117928746	ribosome recycling factor
117928749	SSU ribosomal protein S2P
117928756	Mg chelatase, subunit ChII
117928760	hypothetical protein Acel_1553
117928764	50S ribosomal protein L19
117928765	tRNA (Guanine37-N1) methyltransferase
117928766	16S rRNA processing protein RimM
117928768	SSU ribosomal protein S16P
117928770	signal recognition particle subunit FFH/SRP54 (srp54)
117928774	signal recognition particle-docking protein FtsY
117928777	condensin subunit Smc
117928780	formamidopyrimidine-DNA glycosylase
117928781	ribonuclease III
117928785	phosphopantetheine adenylyltransferase
117928789	50S ribosomal protein L28
117928791	thiamine monophosphate kinase
117928792	AsnC family transcriptional regulator
117928793	D-alanine--D-alanine ligase
117928794	glycerol-3-phosphate dehydrogenase (NAD(P)(+))
117928797	isopropylmalate isomerase small subunit
117928798	isopropylmalate isomerase large subunit
117928799	IclR family transcriptional regulator
117928805	ribose-5-phosphate isomerase B
117928810	HNH endonuclease
117928816	globin
117928821	putative ABC transporter ATP-binding protein
117928856	oligoribonuclease
117928857	hypothetical protein Acel_1650
117928889	redoxin domain-containing protein
117928890	RdgB/HAM1 family non-canonical purine NTP pyrophosphatase
117928891	ribonuclease PH
117928894	cysteine synthase
117928896	Mov34/MPN/PAD-1 family protein
117928898	ATP-dependent Clp protease adaptor protein ClpS
117928899	nicotinate phosphoribosyltransferase

117928935	SsrA-binding protein
117928938	cell division ATP-binding protein FtsE
117928939	peptide chain release factor 2
117928957	delta-1-pyrroline-5-carboxylate dehydrogenase
117928977	transcription factor WhiB
117928979	glutaredoxin-like protein
117928984	molybdopterin biosynthesis-like protein MoeZ
117928991	hypothetical protein Acel_1784
117929031	hypothetical protein Acel_1824
117929033	ECF subfamily RNA polymerase sigma-24 factor
117929042	hypothetical protein Acel_1835
117929044	DNA-3-methyladenine glycosylase I
117929046	dihydropteroate synthase
117929049	hypothetical protein Acel_1842
117929050	dipeptidase
117929062	N-succinyldiaminopimelate aminotransferase
117929068	translation-associated GTPase
117929082	exodeoxyribonuclease VII large subunit
117929084	fructose 1,6-bisphosphatase II
117929088	undecaprenyl pyrophosphate synthetase
117929090	hypothetical protein Acel_1883
117929093	LmbE family protein
117929094	transcription elongation factor GreA
117929114	hypothetical protein Acel_1907
117929117	nucleoside triphosphate pyrophosphohydrolase
117929120	transcription-repair coupling factor
117929151	peptidyl-tRNA hydrolase
117929152	50S ribosomal protein L25/general stress protein Ctc
117929153	ribose-phosphate pyrophosphokinase
117929154	UDP-N-acetylglucosamine pyrophosphorylase / glucosamine-1-phosphate N-acetyltransferase
117929169	F420-0--gamma-glutamyl ligase
117929179	DNA topoisomerase I
117929180	membrane-bound proton-translocating pyrophosphatase
117929183	DEAD/DEAH box helicase domain-containing protein
117929189	type II secretion system protein E
117929191	acetyl-coenzyme A synthetase
117929195	peptidase S1 and S6, chymotrypsin/Hap
117929197	alpha/beta hydrolase fold
117929200	endonuclease III / DNA-(apurinic or apyrimidinic site) lyase
117929202	transcriptional regulator
117929206	endoribonuclease L-PSP
117929207	hypothetical protein Acel_2000
117929208	putative ion-transporting ATPase
117929209	anion-transporting ATPase
117929212	GatB/Yqey domain-containing protein
117929213	metallophosphoesterase
117929219	aspartate kinase
117929222	recombination protein RecR
117929223	DNA polymerase III, subunits gamma and tau
117929231	thiamineS protein
117929232	response regulator receiver protein
117929246	NifC-like ABC-type porter
117929278	phosphoribosylformylglycinamide cyclo-ligase
117929279	amidophosphoribosyltransferase
117929281	phosphoribosylformylglycinamide synthase II
117929282	phosphoribosylformylglycinamide synthase I
117929283	phosphoribosylformylglycinamide synthase subunit PurS
117929292	zinc uptake regulator

117929308	alcohol dehydrogenase
117929309	hypothetical protein Acel_2102
117929312	prephenate dehydratase
117929319	deoxycytidine triphosphate deaminase
117929324	chaperone protein DnaJ
117929332	fructose-bisphosphate aldolase
117929333	adenylosuccinate synthetase
117929334	phosphoribosylamine--glycine ligase
117929336	phosphoribosylaminoimidazole-succinocarboxamide synthase
117929337	50S ribosomal protein L9P
117929340	SSU ribosomal protein S6P
117929344	PadR family transcriptional regulator
117929345	myo-inositol-1-phosphate synthase
117929346	metal dependent phosphohydrolase
117929361	chromosome segregation DNA-binding protein
117929362	cobyrinic acid a,c-diamide synthase
117929364	single-stranded nucleic acid binding R3H domain-containing protein

Table S9. The 46 *A. cellulolyticus* 11B homologs used in the analysis presented in Table S6.

GI	Product
117927505	50S ribosomal protein L1P
117927507	50S ribosomal protein L12P
117927509	DNA-directed RNA polymerase subunit beta'
117927510	30S ribosomal protein S12
117927511	30S ribosomal protein S7
117927514	30S ribosomal protein S10
117927517	50S ribosomal protein L23P
117927518	50S ribosomal protein L2
117927522	50S ribosomal protein L16
117927524	SSU ribosomal protein S17P
117927525	50S ribosomal protein L14
117927527	50S ribosomal protein L5
117927529	SSU ribosomal protein S8P
117927530	50S ribosomal protein L6P
117927531	50S ribosomal protein L18P
117927532	SSU ribosomal protein S5P
117927535	preprotein translocase subunit SecY
117927540	30S ribosomal protein S13
117927541	30S ribosomal protein S11
117927550	LSU ribosomal protein L13P
117927551	SSU ribosomal protein S9P
117927552	phosphoglucosamine mutase
117927571	chaperonin Cpn10
117927580	bifunctional GMP synthase/glutamine amidotransferase protein
117927593	bifunctional phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase
117927849	peptide chain release factor 1
117927859	F0F1 ATP synthase subunit alpha
117927861	F0F1 ATP synthase subunit beta
117927905	aspartyl/glutamyl-tRNA amidotransferase subunit A
117927906	aspartyl/glutamyl-tRNA amidotransferase subunit B
117927965	50S ribosomal protein L27
117928220	cell division protein FtsZ
117928313	excinuclease ABC subunit B
117928316	excinuclease ABC subunit A
117928344	FeS assembly ATPase SufC
117928447	cobyrinic acid a,c-diamide synthase
117928474	50S ribosomal protein L20
117928516	chorismate synthase
117928550	Holliday junction DNA helicase B
117928553	hypothetical protein Acel_1346
117928716	30S ribosomal protein S15
117928938	cell division ATP-binding protein FtsE
117928939	peptide chain release factor 2
117929068	translation-associated GTPase
117929222	recombination protein RecR
117929333	adenylosuccinate synthetase