**Title:** Complete genome of the cellyloytic thermophile A*cidothermus cellulolyticus 11B* provides insights into its ecophysiological and evolutionary adaptations

**Author(s):** R. Barabote[1,†], G. Xie[1], D. Leu[2], P. Normand[3], A. Necsulea[4], V. Daubin[4], C. Médigue[5], W. Adney[6], X. Xu[2], A. Lapidus[7], C. Detter[1], P. Pujic[3], D. Bruce[1], C. Lavire[3], J. Challacombe[1], T. Brettin[1] and Alison M. Berry[2].

**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 Xie[1], David H. Leu[2], Philippe Normand[3], Anamaria Necsulea[4], Vincent Daubin[4], Claudine Médigue[5], William S. Adney[6], Xin Clare Xu[2], Alla Lapidus[7], Chris Detter[1], Petar Pujic[3], David Bruce[1], Celine Lavire[3], Jean F. Challacombe[1], Thomas S. Brettin[1], and Alison M. Berry[2].

[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 37º and 70ºC; optimal growth temperature [OGT] is 55ºC). 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 81ºC 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 37ºC (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 25-28ºC), 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, aryldialkylphosphatase, 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 *Thermoanerobacter*.

*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, glutaconate 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 *Kineococcu*s. 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 *Nocardioides* 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 in 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 thermophiliy 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 55ºC, 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., Choisne, 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 Micriobiol. 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 (ΔGS; 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*).

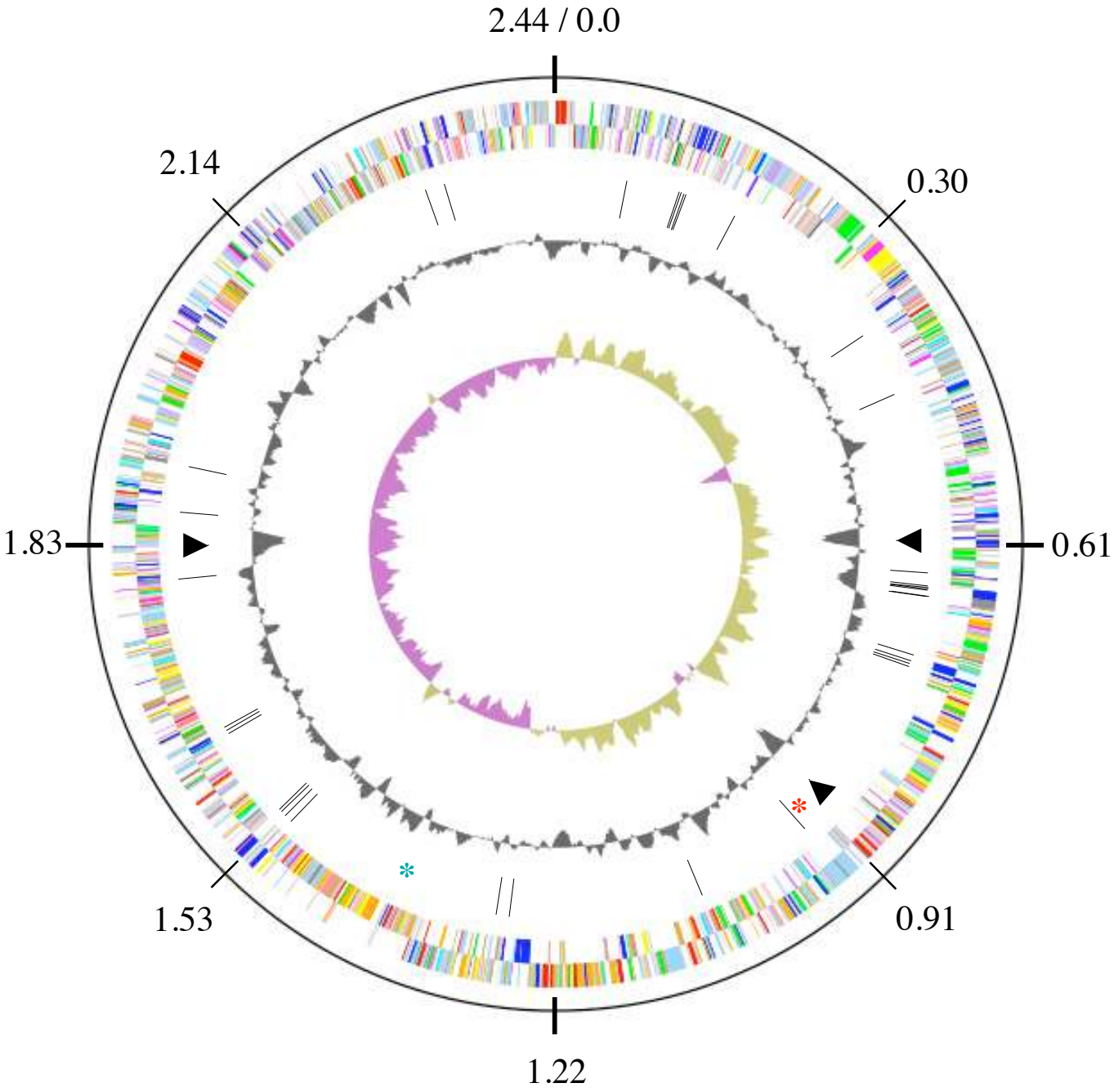


Fig. 1

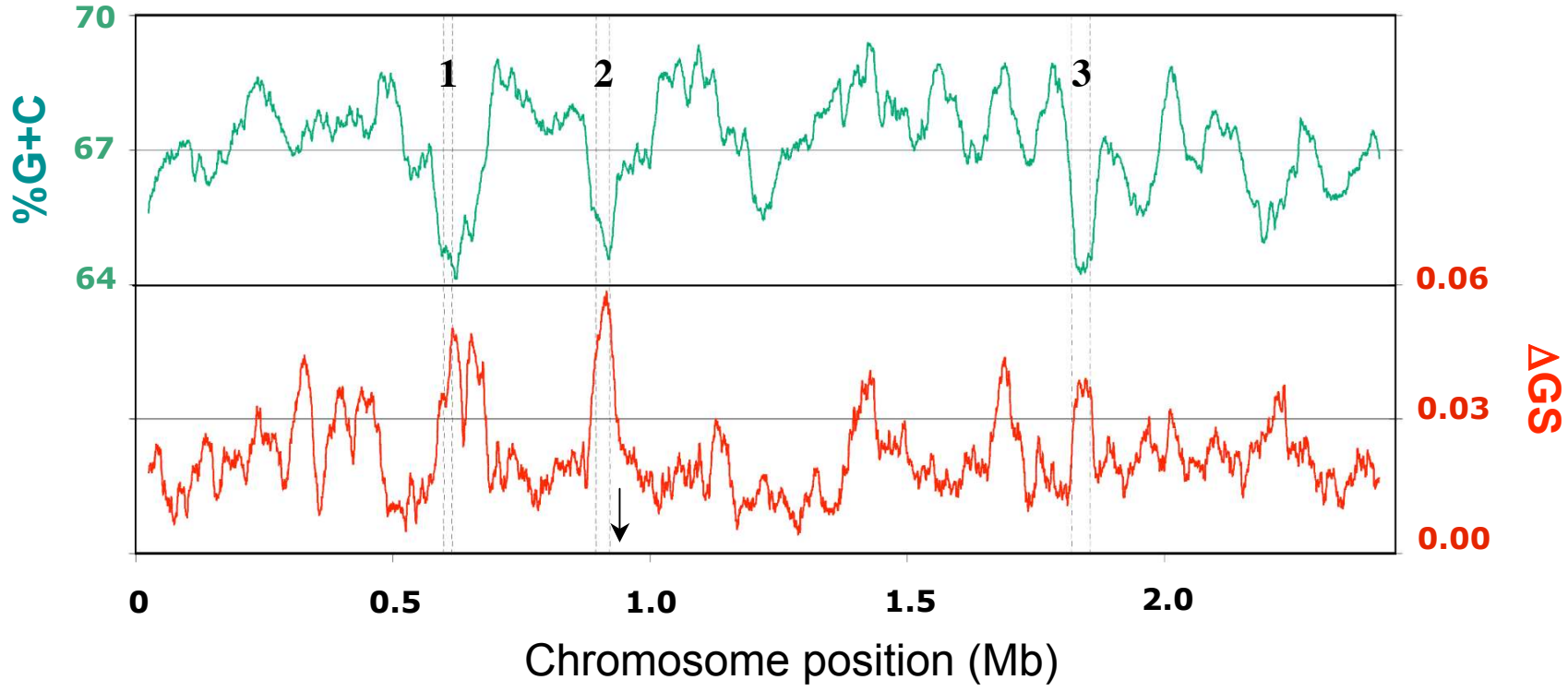
70
67
64
%G+C
1
2
3
0.06
0.03
0.00
ΔGS
0
0.5
1.0
1.5
2.0
Chromosome position (Mb)

Fig. 2

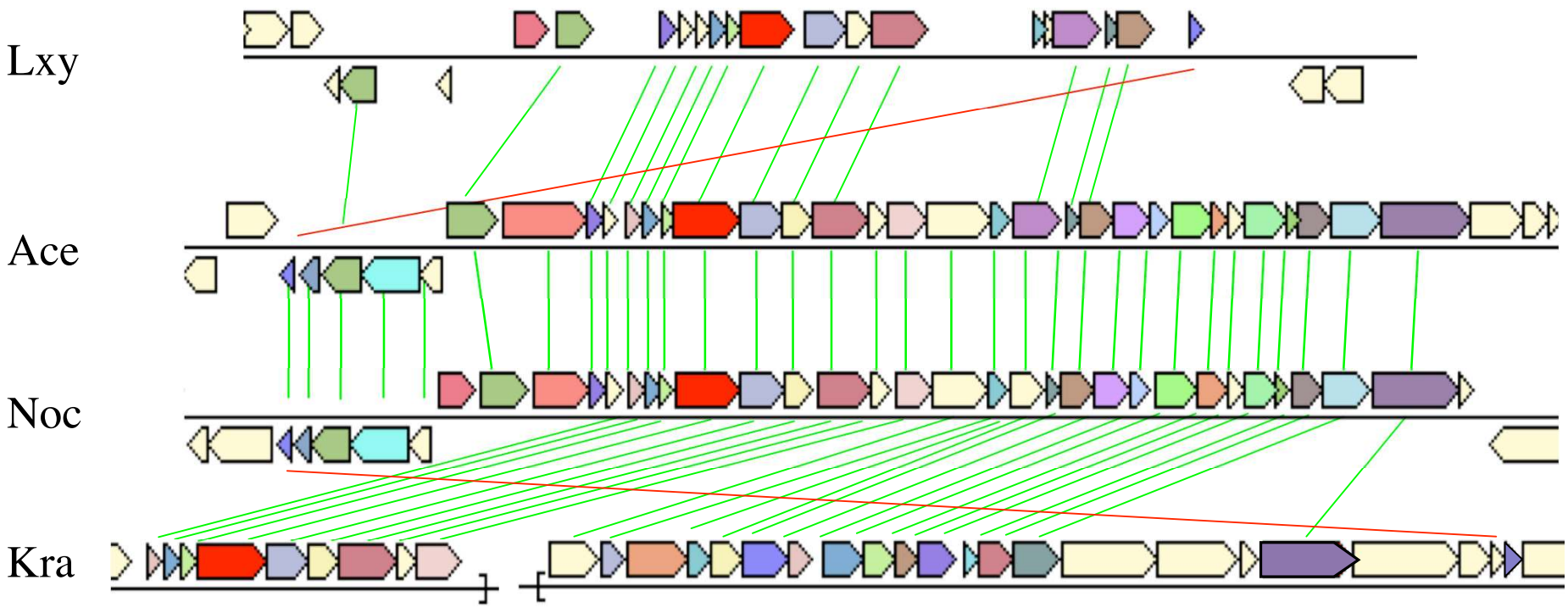


Fig. 3

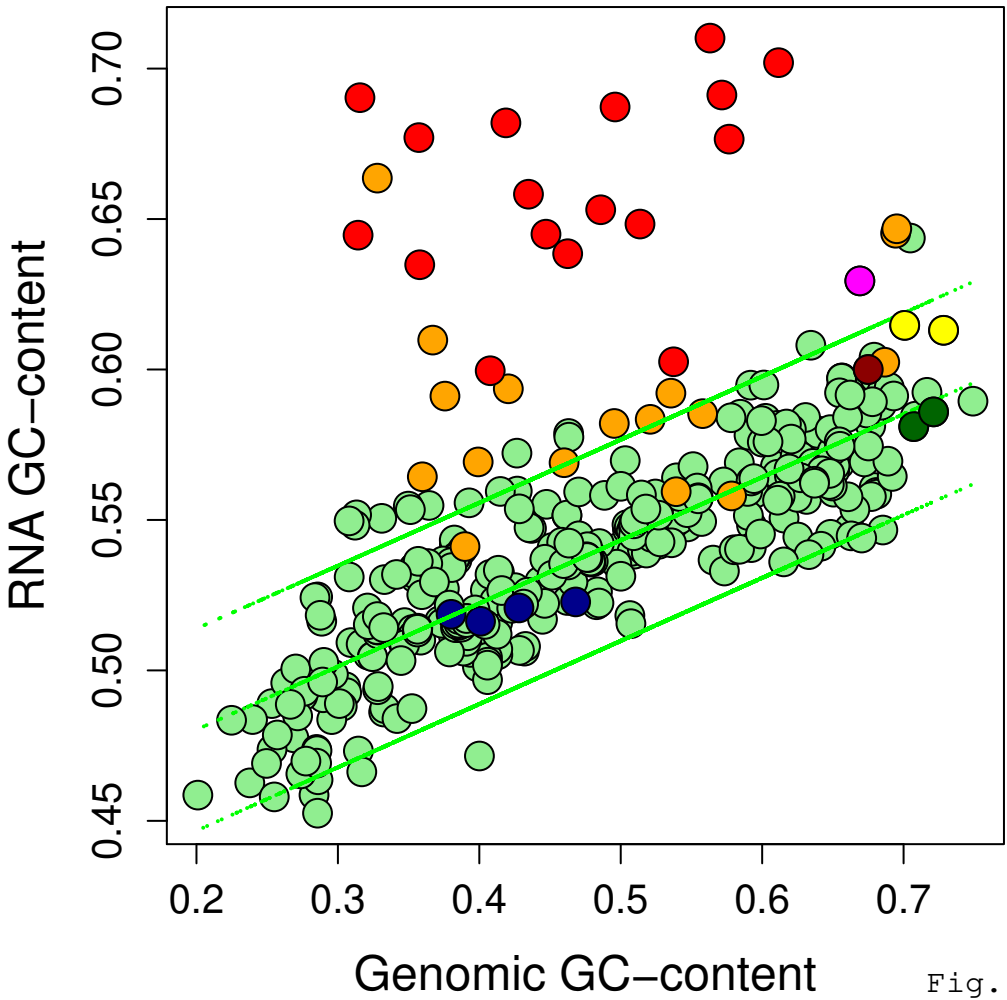
RNA GC−content
0.45
0.50
0.55
0.60
0.65
0.70
0.2
0.3
0.4
0.5
0.6
0.7
Genomic GC−content

Fig. 4

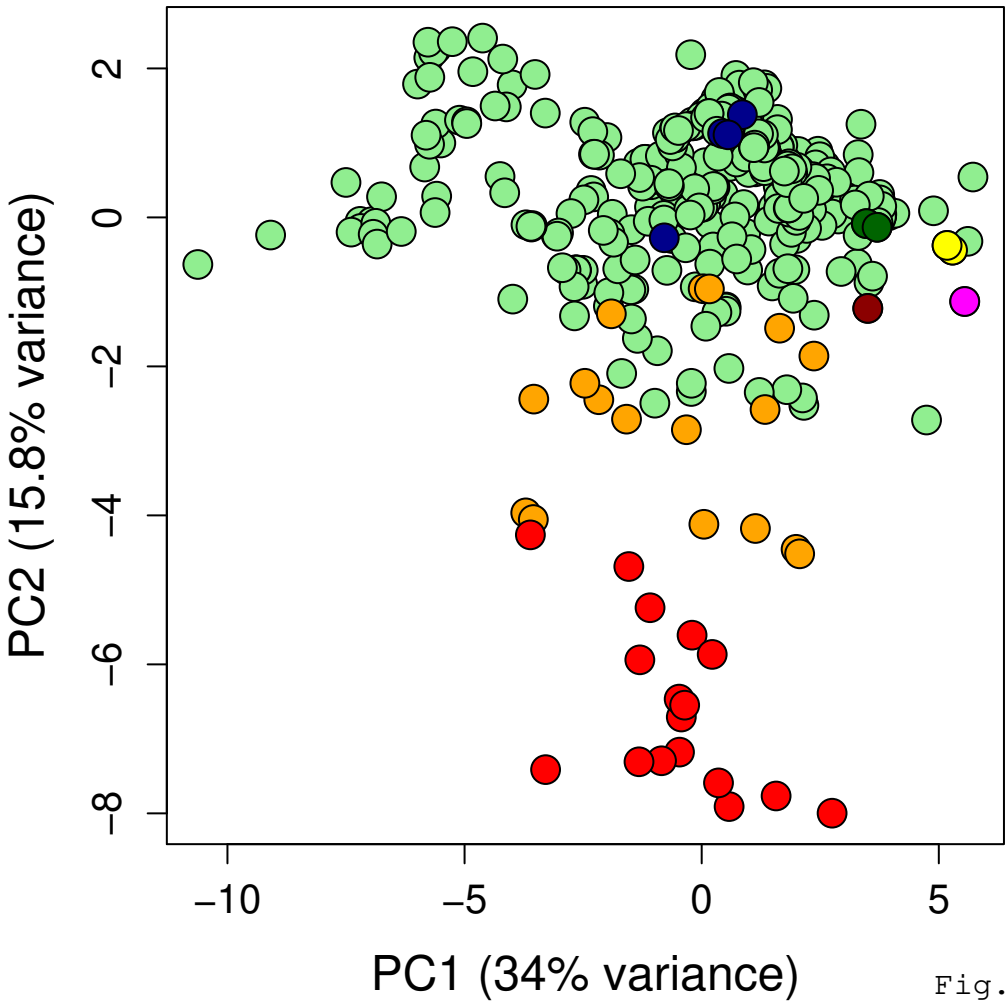


Fig. 5

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

| Features | *Acidothermus cellulolyticus* 11B | *Frankia* sp. CcI3 | *Frankia alni* ACN14a | *Streptomyces avermitilis* MA-4680 | *Streptomyces coelicolor* A3(2) | *Thermobifida fusca* 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 |

[a] The 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/glucodextranase | 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 = cytoplamic, 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 |
|---|---|---|---|---|---|
| Genomic Island 1. | | | | | |
| 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 | Nuclueotide 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 |
| Genomic Island 2. | | | | | |
| Acel_R0021 | + | 58.7 | | *Xaa tRNA* | |
| Acel_0810 | + | 59.0 | 61 | DNA binding domain, excisionase family | VrlI homolog |
| Acel_0811 | + | 59.3 | 159 | conserved hypothetical protein | VrlJ homolog |
| Acel_0812 | + | 62.6 | 1244 | conserved hypothetical protein | VrlK 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 | VrlO homolog? |
| Acel_0817 | + | 57.5 | 678 | conserved hypothetical protein | VrlP homolog |
| Acel_0818 | + | 57.0 | 261 | conserved hypothetical protein | VrlQ 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 | | *Met tRNA* | |
| Genomic Island 3. | | | | | |
| Acel_R0044 | + | 68.5 | | *Arg tRNA* | |
| 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 | | *His tRNA* | |

+/- 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 poistions in six actinobacteria.

| Organism | OGT ($^0$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 |
| *A. cellulolyticus* 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 |
| *Frankia alni* 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 |
| *Frankia* 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 |
| *S. avermitilis* 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 |
| *S. coelicolor* 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 |
| *T. fusca* 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.

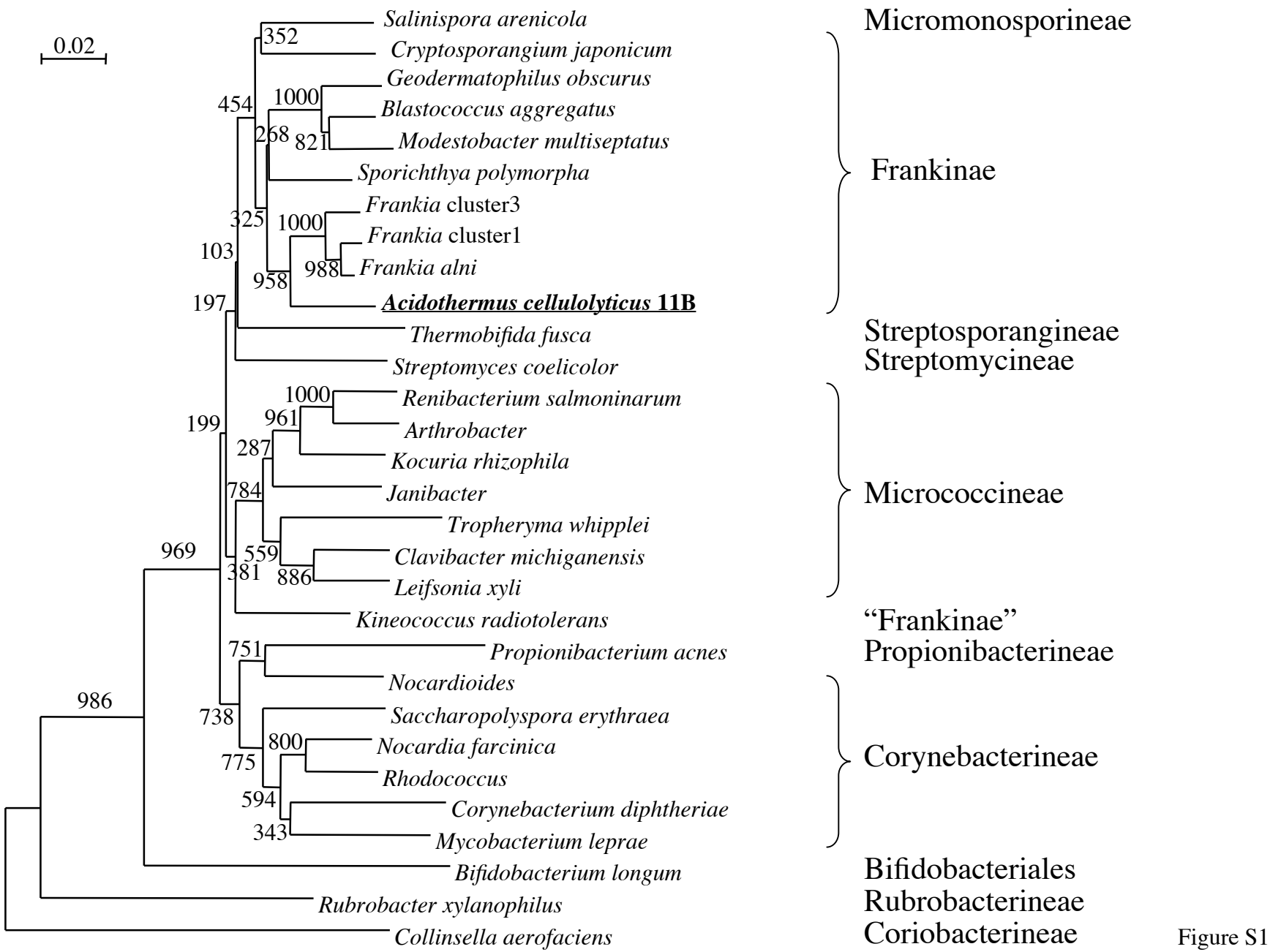
0.02
Salinispora arenicola
Cryptosporangium japonicum
Geodermatophilus obscurus
Blastococcus aggregatus
Modestobacter multiseptatus
Sporichthya polymorpha
Frankia cluster3
Frankia cluster1
Frankia alni
Acidothermus cellulolyticus 11B
Thermobifida fusca
Streptomyces coelicolor
Renibacterium salmoninarum
Arthrobacter
Kocuria rhizophila
Janibacter
Tropheryma whipplei
Clavibacter michiganensis
Leifsonia xyli
Kineococcus radiotolerans
Propionibacterium acnes
Nocardioides
Saccharopolyspora erythraea
Nocardia farcinica
Rhodococcus
Corynebacterium diphtheriae
Mycobacterium leprae
Bifidobacterium longum
Rubrobacter xylanophilus
Collinsella aerofaciens
352
454
1000
268
821
325
1000
103
988
958
197
199
1000
961
287
784
969
559
381
886
751
986
738
800
775
594
343
Micromonosporineae
Frankinae
Streptosporangineae
Streptomycineae
Micrococcineae
"Frankinae"
Propionibacterineae
Corynebacterineae
Bifidobacteriales
Rubrobacterineae
Coriobacterineae

Figure S1

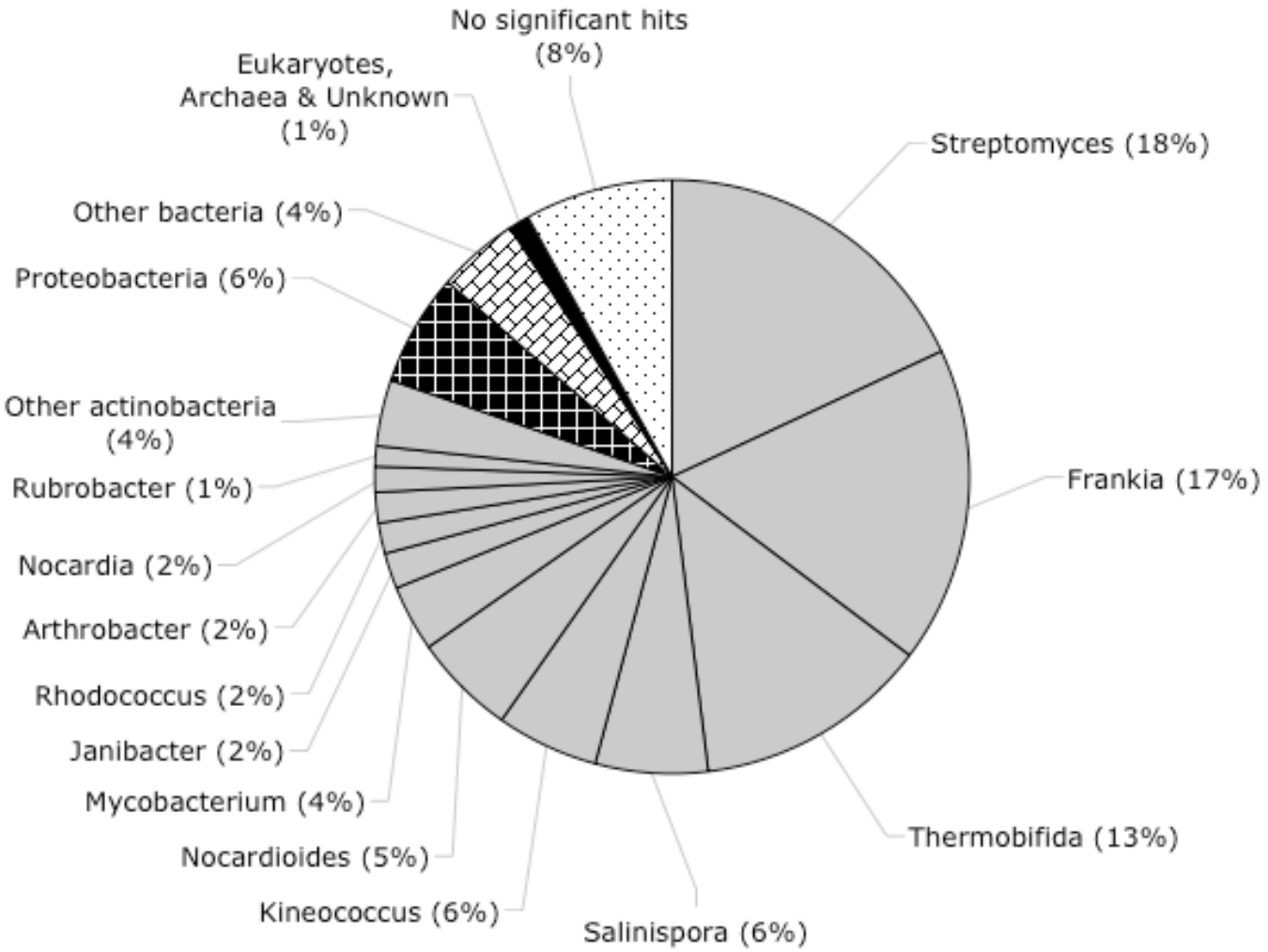
No significant hits
(8%)
Eukaryotes,
Archaea & Unknown
(1%)
Streptomyces (18%)
Other bacteria (4%)
Proteobacteria (6%)
Other actinobacteria
(4%)
Rubrobacter (1%)
Frankia (17%)
Nocardia (2%)
Arthrobacter (2%)
Rhodococcus (2%)
Janibacter (2%)
Mycobacterium (4%)
Nocardioides (5%)
Thermobifida (13%)
Kineococcus (6%)
Salinispora (6%)

Figure S2

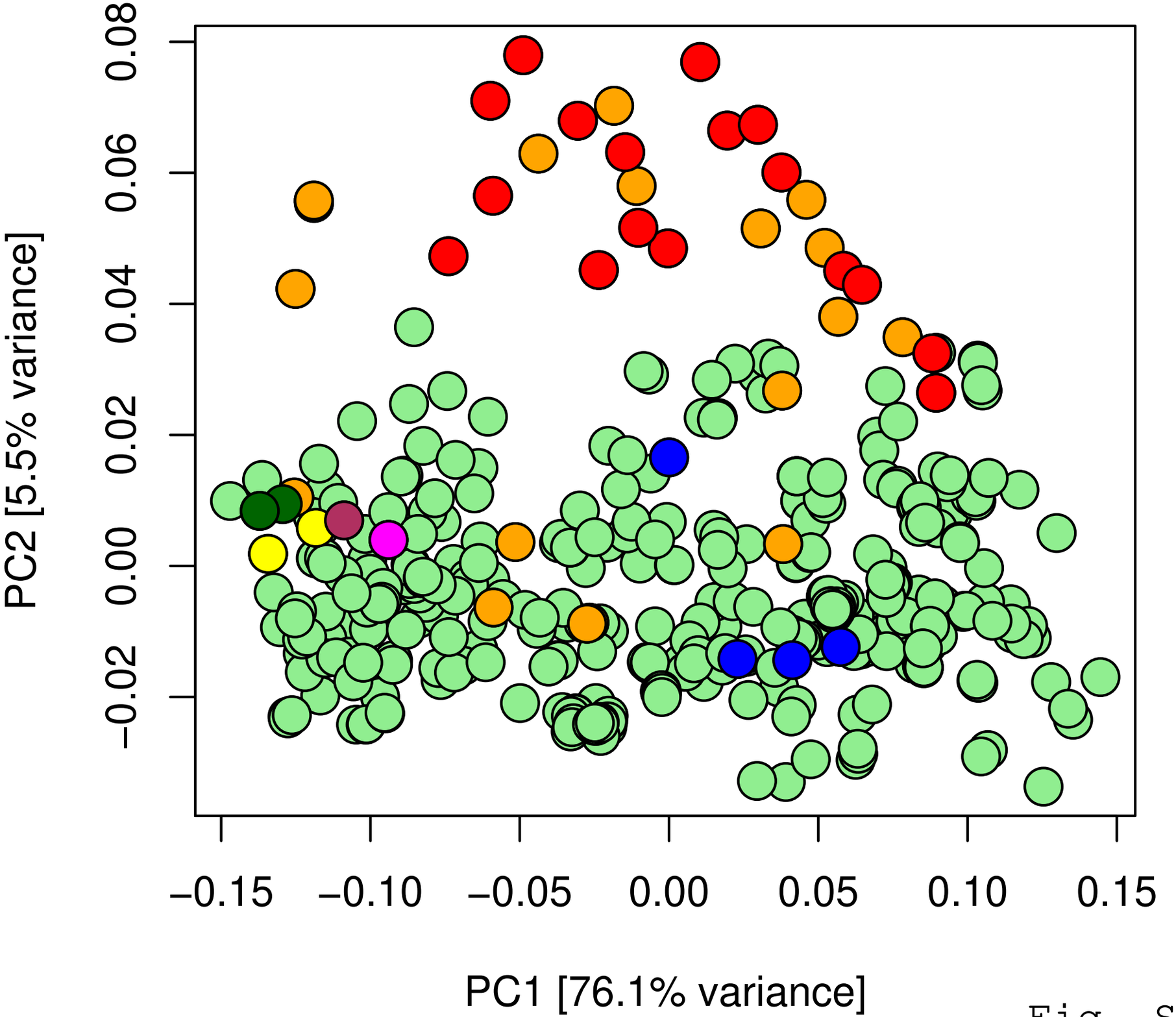
PC2 [5.5% variance]
0.08
0.06
0.04
0.02
0.00
−0.02
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
PC1 [76.1% variance]

Fig. S3A

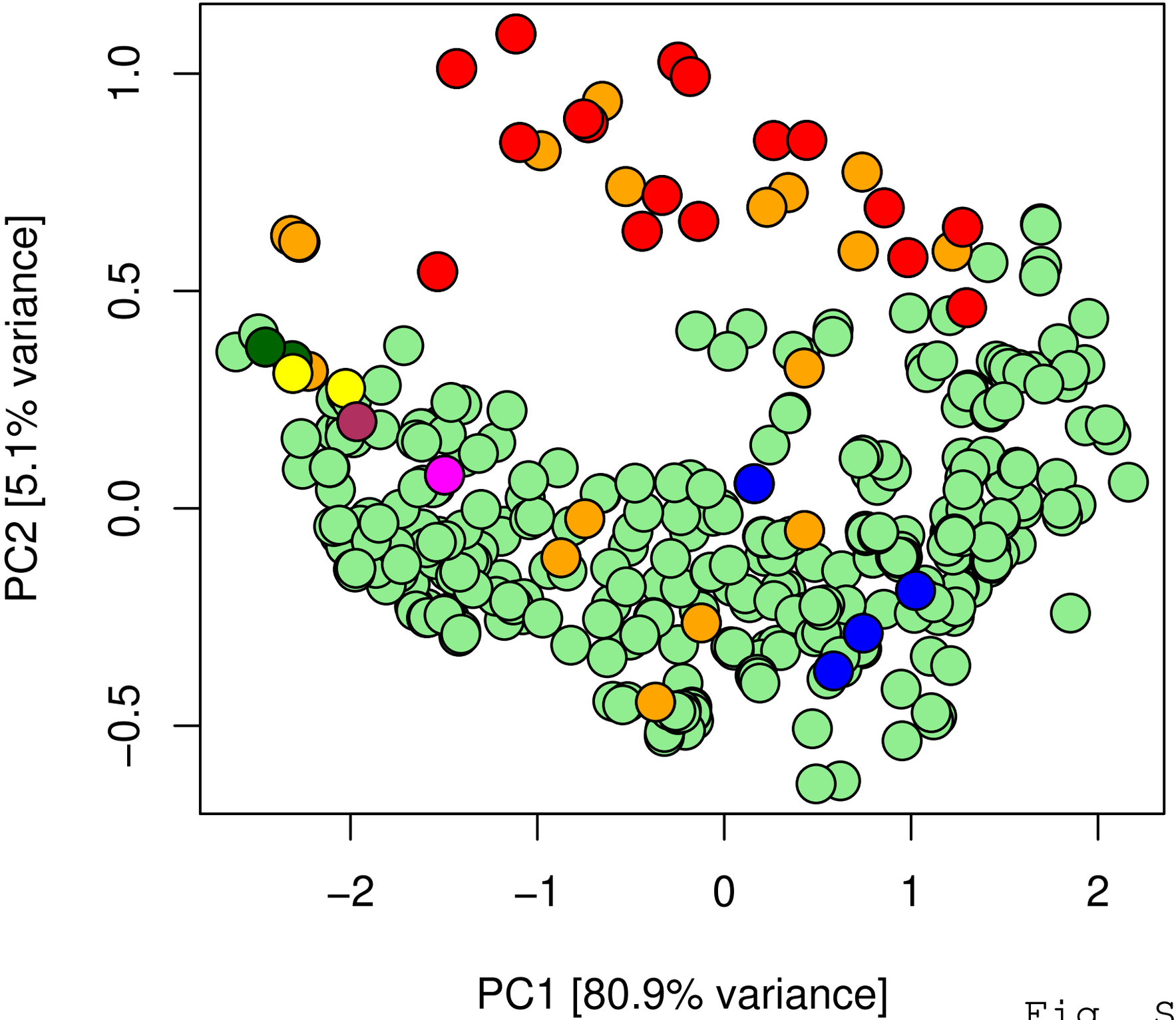


Fig. S3B

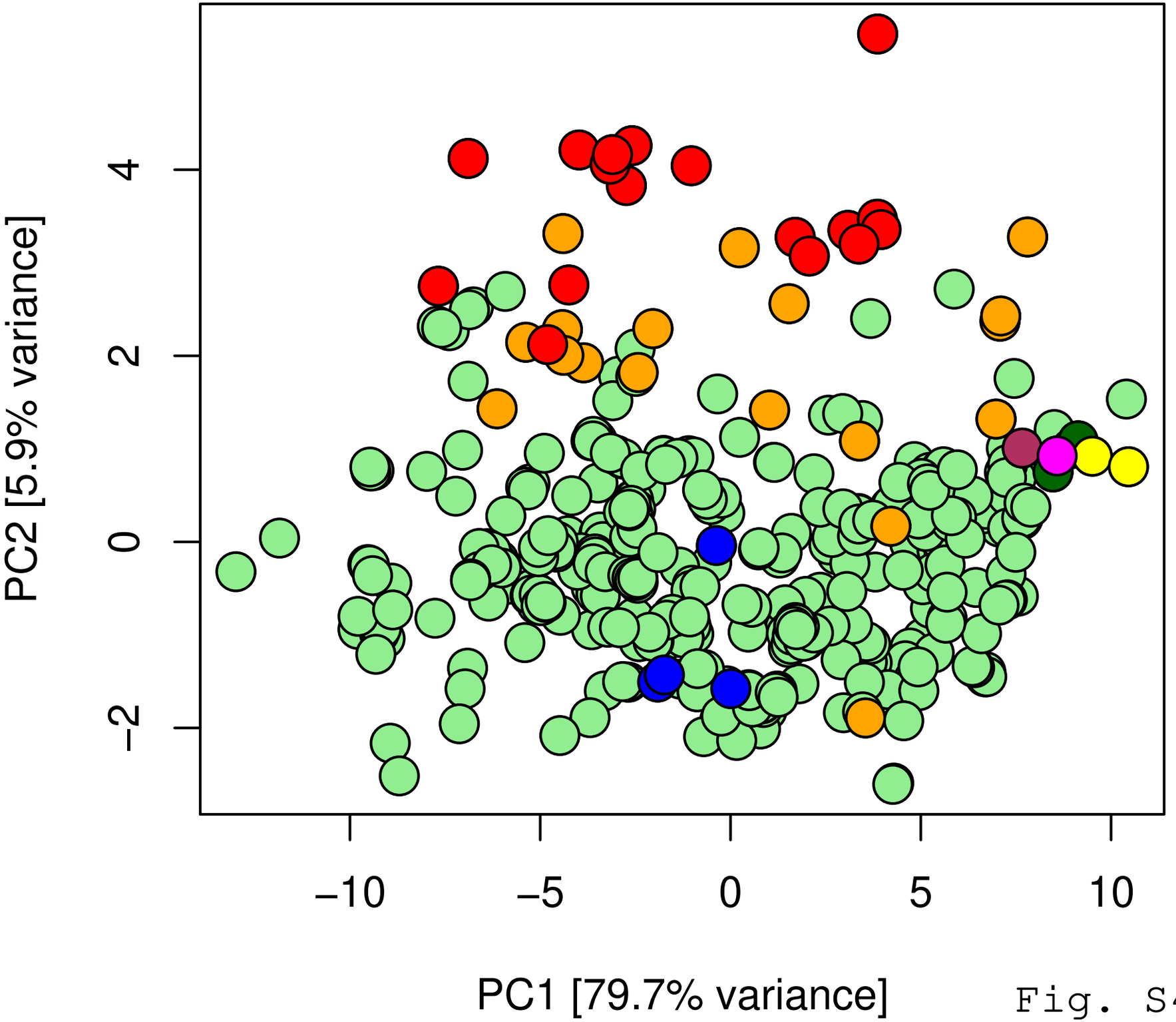


Fig. S4

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 |
|---|---|---|---|---|---|
| (A) Best hits to Archaea | | | | | |
| Acel_0034 | 245 | hypothetical protein Acel_0034 | 110667166 | *Haloquadratum walsbyi DSM 16790* | 224 |
| Acel_0498 | 111 | protein of unknown function DUF59 | 14590230 | *Pyrococcus horikoshii OT3* | 264 |
| Acel_0525 | 137 | hypothetical protein Acel_0525 | 15897985 | *Sulfolobus solfataricus P2* | 291 |
| Acel_0526 | 204 | hypothetical protein Acel_0526 | 88604270 | *Methanospirillum hungatei JF-1* | 111 |
| Acel_0621 | 366 | hypothetical protein Acel_0621 | 110667166 | *Haloquadratum walsbyi DSM 16790* | 225 |
| Acel_0638 | 381 | UDP-N-acetylglucosamine 2-epimerase | 15678857 | *Methanothermobacter thermautotrophicus str. Delta H* | 519 |
| Acel_0721 | 283 | fructose-bisphosphate aldolase | 15922681 | *Sulfolobus tokodaii str. 7* | 320 |
| Acel_0888 | 284 | ABC transporter related | 73668563 | *Methanosarcina barkeri str. Fusaro* | 588 |
| Acel_0910 | 245 | ABC transporter related | 110669070 | *Haloquadratum walsbyi DSM 16790* | 620 |
| Acel_1270 | 291 | ATP phosphoribosyltransferase (homohexameric) | 116753404 | *Methanosaeta thermophila PT* | 780 |
| Acel_1310 | 219 | hypothetical protein Acel_1310 | 110667166 | *Haloquadratum walsbyi DSM 16790* | 163 |
| Acel_1639 | 230 | Asp/Glu racemase | 14521305 | *Pyrococcus abyssi GE5* | 408 |
| Acel_1644 | 317 | Glutaconate CoA-transferase | 11498798 | *Archaeoglobus fulgidus DSM 4304* | 495 |
| Acel_1897 | 233 | Small-conductance mechanosensitive channel-like | 48477585 | *Picrophilus torridus DSM 9790* | 200 |
| Acel_1930 | 282 | ABC-2 type transporter | 15899372 | *Sulfolobus solfataricus P2* | 221 |
| Acel_1931 | 347 | ABC transporter related | 119720042 | *Thermofilum pendens Hrk 5* | 406 |
| Acel_2066 | 164 | Vitamin K epoxide reductase | 70606913 | *Sulfolobus acidocaldarius DSM 639* | 193 |
| Acel_2067 | 303 | hypothetical protein Acel_2067 | 15898666 | *Sulfolobus solfataricus P2* | 355 |
| (B) Best hits to Eukarya | | | | | |
| Acel_0064 | 898 | hypothetical protein Acel_0064 | 118129698 | *Gallus gallus* | 294 |
| Acel_0179 | 656 | esterase, PHB depolymerase family | 114324587 | *Volvariella volvacea* | 753 |
| Acel_0740 | 439 | hypothetical protein Acel_0740 | 97180301 | *Sus scrofa* | 264 |
| Acel_0770 | 160 | hypothetical protein Acel_0770 | 118085709 | *Gallus gallus* | 112 |
| Acel_1067 | 206 | GPR1/FUN34/yaaH family protein | 119178442 | *Coccidioides immitis RS* | 354 |
| Acel_1163 | 225 | cell wall surface anchor family protein | 109658562 | *Homo sapiens* | 180 |
| Acel_1712 | 323 | hypothetical protein Acel_1712 | 46119356 | *Gibberella zeae PH-1* | 760 |
| Acel_1727 | 219 | beta-lactamase domain protein | 125820913 | *Danio rerio* | 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 *Acidothermus* 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 *Streptomyces* species, *Frankia* EAN1 and *T. fusca*. |
| 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 *S. coelicolor* only). Second part, specific (conserved) hypothetical proteins. |
| GR7 | 650219 | 677497 | Cluster of protein/enzymes involved in cellulose degradation (specific to *Acidothermus* 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 *Acidothermus* – Actually, the genes in the region encoded a pyruvate oxidoreductase or Pyruvic-ferredoxin oxidoreductase. The cluster is also find in *Helicobacter pylori* strains annotated as:<br>PorC = Pyruvate oxidoreductase gamma chain (ACICE0782)<br>PorD = Pyruvate oxidoreductase delta chain (partial match on ACICE0785 but more than 51% identity in aa).<br>PorA = Pyruvate oxidoreductase alpha chain (ACICE0783)<br>PorB = Pyruvate oxidoreductase beta chain (ACICE0784) |
| GR10 | 783644 | 809485 | First part, transport system + regulator + enzymatic activities (kinase, oxidase, glycosyltransferase). Second part, highly specific to *Acidothermus*, cluster *hyf* genes coding hydrogenase subunits (NADH dehydrogenase (ubiquinone)/ ATP synthesis coupled electron transport). This cluster is found in a very well conserved synteny in: *Anaeromyxobacter* species (6 genes, from ACICE0811 to ACICE0816, identities % between 30 and 40) with the annotation 'NADH dehydrogenase (quinone)', and in *Yersinia* 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 *Acidothermus*. |
| GR11 | 975117 | 987762 | Transport system (ABC type, susbtrat nitrate?), regulator (lacI family) and putative nitrilase. Highly specific to *Acidothermus*. Nitrilase (ACICE0994) and ACICE0997 in synteny with two genes of the *Rhizobium leguminosarum* bv. *viciae* 3841 plasmid pRL80076 (putative aliphatic nitrilase) and pRL80075 (putative endoribonuclease L-PSP family protein). |
| GR12 | 1122316 | 1130148 | First part, shared with Frankia CcI3 only, enzymatic activities (cytochrome c3 hydrogenase), second part more specific to *Acidothermus* with (conserved) hypothetical proteins. |
| GR13 | 1196923 | 1203395 | Transport system (type ABC, substrate amino acid?) shared with *Frankia* species only. |
| GR14 | 1212310 | 1224927 | Metabolism, very probably degradation (monooxygenase, dioxygenase) of glutamine/glutamate? + Regulator (marR family) + transport? (permease). Highly specific to *Acidothermus*. |
| GR15 | 1284136 | 1295293 | Mainly hypothetical proteins with a putative rRNA methylase and exonuclease. Mainly specific to *Acidothermus*. |
| GR16 | 1459534 | 1477754 | Type IV pilus (or type II?) highly specific to *Acidothermus*. Best synteny group shared with *Kineococcus radiotolerans* SRS30216 (10 genes, %identity: 30-74), *Moorella thermoacetica* 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 phenylacetatic acid degradation) + regulation (*tetR* family). Shared with *Streptomyces* 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 *Acidothermus* in the second part. |
| GR22 | 2207054 | 2241063 | Mainly conserved hypothetical proteins + transport system + *marR* 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 | *Acidothermus cellulolyticus* 11B | *Frankia* sp. ACN14 | *Frankia* sp. CcI3 | *Streptomyces avermitilis* | *Streptomyces coelicolor* | *Thermobifida fusca* |
|---|---|---|---|---|---|---|---|
| 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 orthologus proteins |
|---|---|---|---|---|
| *Acidothermus cellulolyticus* 11B | 55 | 40.56 | 41.94 | 41.76 |
| *Frankia alni* ACN14a | 28 | 38.83 | 39.96 | 39.90 |
| *Frankia* sp. CcI3 | 27 | 39.68 | 40.68 | 40.32 |
| *Streptomyces avermitilis* MA-4680 | 28 | 38.97 | 40.49 | 40.12 |
| *Streptomyces coelicolor* A3(2) | 30 | 38.83 | 40.67 | 40.00 |
| *Thermobifida fusca* 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 |
| *Acidothermus cellulolyticus* 11B | 2.40 | 66.9 | 55 | 40.6 | 41.9 | 42.4 |
| *Arthrobacter aurescens* TC1 | 5.23 | 62.4 | 30 | 38.4 | 39.2 | 40.2 |
| *Arthrobacter* sp. FB24 | 5.08 | 65.4 | 30 | 38.3 | 39.2 | 40.1 |
| *Bifidobacterium adolescentis* ATCC 15703 | 2.10 | 59.2 | 37 | 37.6 | 38.4 | 39.0 |
| *Bifidobacterium longum* | 2.26 | 60.1 | 37 | 37.2 | 38.2 | 39.0 |
| *Clavibacter michiganensis* NCPPB 382 | 3.40 | 72.5 | 28 | 39.7 | 40.6 | 40.3 |
| *Corynebacterium diphtheriae* | 2.49 | 53.5 | 37 | 38.8 | 39.4 | 40.5 |
| *Corynebacterium efficiens* YS-314 | 3.15 | 63.1 | 37 | 39.5 | 39.7 | 40.9 |
| *Corynebacterium glutamicum* ATCC 13032 Bielefeld | 3.30 | 53.8 | 33 | 39.0 | 39.6 | 40.7 |
| *Corynebacterium glutamicum* ATCC 13032 Kitasato | 3.30 | 53.8 | 33 | 39.0 | 39.5 | 40.7 |
| *Corynebacterium glutamicum* R | 3.35 | 54.1 | 33 | 39.0 | 39.5 | 40.7 |
| *Corynebacterium jeikeium* K411 | 2.48 | 61.4 | 37 | 38.4 | 39.2 | 39.8 |
| *Frankia alni* ACN14a | 7.50 | 72.8 | 28 | 38.8 | 40.0 | 41.4 |
| *Frankia* sp. CcI3 | 5.40 | 70.1 | 27 | 39.7 | 40.7 | 41.7 |
| *Kineococcus radiotolerans* SRS30216 | 4.81 | 74.2 | 32 | 40.2 | 41.0 | 40.4 |
| *Leifsonia xyli* subsp. *xyli* CTCB0 | 2.58 | 67.7 | 29 | 40.0 | 41.0 | 40.4 |
| *Mycobacterium avium* 104 | 5.50 | 69.0 | 39 | 38.8 | 39.7 | 40.7 |
| *Mycobacterium avium* subsp. *paratuberculosis* | 4.80 | 69.3 | 39 | 38.8 | 39.7 | 40.7 |
| *Mycobacterium bovis* | 4.35 | 65.6 | 37 | 38.1 | 39.9 | 40.7 |
| *Mycobacterium bovis* BCG Pasteur 1173P2 | 4.40 | 65.6 | 37 | 38.1 | 39.9 | 40.7 |
| *Mycobacterium gilvum* PYR-GCK | 5.96 | 67.7 | 30 | 39.0 | 39.6 | 40.3 |
| *Mycobacterium leprae* | 3.27 | 57.8 | 37 | 39.6 | 40.3 | 41.0 |
| *Mycobacterium smegmatis* MC2 155 | 7.00 | 67.4 | 37 | 39.3 | 39.8 | 40.6 |
| *Mycobacterium* sp. JLS | 6.00 | 68.4 | 30 | 39.3 | 39.8 | 40.4 |
| *Mycobacterium* sp. KMS | 6.22 | 68.2 | 30 | 39.3 | 39.8 | 40.5 |
| *Mycobacterium* sp. MCS | 5.92 | 68.4 | 30 | 39.3 | 39.8 | 40.5 |
| *Mycobacterium tuberculosis* CDC1551 | 4.40 | 65.6 | 37 | 38.2 | 39.9 | 40.8 |
| *Mycobacterium tuberculosis* F11 | 4.40 | 65.6 | 37 | 38.0 | 39.9 | 40.7 |
| *Mycobacterium tuberculosis* H37Ra | 4.40 | 65.6 | 37 | 38.0 | 39.9 | 40.7 |
| *Mycobacterium tuberculosis* H37Rv | 4.40 | 65.6 | 37 | 38.0 | 40.0 | 40.7 |
| *Mycobacterium ulcerans* Agy99 | 5.60 | 65.5 | 32 | 38.4 | 39.7 | 40.8 |
| *Mycobacterium vanbaalenii* PYR-1 | 6.50 | 67.8 | 30 | 39.0 | 39.7 | 40.4 |
| *Nocardia farcinica* IFM10152 | 6.29 | 70.7 | 37 | 39.9 | 40.6 | 40.6 |
| *Nocardioides* sp. JS614 | 5.31 | 71.4 | 30 | 40.3 | 41.0 | 40.5 |
| *Propionibacterium acnes* KPA171202 | 2.56 | 60.0 | 37 | 38.9 | 39.5 | 40.0 |
| *Rhodococcus* sp. RHA1 | 9.67 | 67.0 | 30 | 39.4 | 39.9 | 40.4 |
| *Rubrobacter xylanophilus* DSM 9941 | 3.23 | 70.5 | 60 | 44.9 | 45.5 | 45.0 |
| *Saccharopolyspora erythraea* NRRL 2338 | 8.20 | 71.1 | 28 | 40.5 | 41.1 | 40.9 |
| *Salinispora tropica* CNB-440 | 5.20 | 69.5 | 28 | 40.1 | 41.0 | 41.2 |
| *Streptomyces avermitilis* MA-4680 | 9.12 | 70.7 | 28 | 39.0 | 40.5 | 40.4 |
| *Streptomyces coelicolor* A3(2) | 9.09 | 72.0 | 30 | 39.3 | 40.7 | 40.3 |
| *Symbiobacterium thermophilum* IAM14863 | 3.60 | 68.7 | 60 | 42.8 | 43.3 | 42.8 |
| *Thermobifida fusca* YX | 3.60 | 67.5 | 50 | 41.2 | 41.9 | 41.6 |
| *Tropheryma whipplei* TW08 27 | 0.93 | 46.3 | 37 | 40.0 | 40.4 | 40.0 |
| *Tropheryma whipplei* Twist | 0.93 | 46.3 | 37 | 40.0 | 40.4 | 39.8 |
| | R-sqaured 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 |
|---|---|---|---|
| (A) Archaea. | | | |
| *Aeropyrum pernix* | 90.0 | 56.31 | 71.01 |
| *Archaeoglobus fulgidus* | 85.0 | 48.58 | 65.31 |
| *Halobacterium sp* | 37.0 | 67.91 | 60.41 |
| *Hyperthermus butylicus* | 99.0 | 53.74 | 60.26 |
| *Methanobacterium thermoautotrophicum* | 65.0 | 49.54 | 58.21 |
| *Methanococcoides burtonii DSM 6242* | 21.5 | 40.76 | 53.19 |
| *Methanococcus jannaschii* | 85.0 | 31.43 | 64.47 |
| *Methanococcus maripaludis S2* | 37.0 | 33.10 | 55.05 |
| *Methanocorpusculum labreanum Z* | 37.0 | 50.01 | 53.13 |
| *Methanoculleus marisnigri JR1* | 27.5 | 62.06 | 58.02 |
| *Methanopyrus kandleri* | 98.0 | 61.16 | 70.20 |
| *Methanosaeta thermophila PT* | 61.0 | 53.55 | 59.21 |
| *Methanosarcina acetivorans* | 38.5 | 42.68 | 55.96 |
| *Methanosarcina barkeri fusaro* | 33.5 | 39.28 | 55.57 |
| *Methanosarcina mazei* | 37.0 | 41.48 | 55.95 |
| *Methanosphaera stadtmanae* | 37.0 | 27.63 | 49.31 |
| *Methanospirillum hungatei JF-1* | 37.0 | 45.15 | 54.09 |
| *Nanoarchaeum equitans* | 90.0 | 31.56 | 69.03 |
| *Natronomonas pharaonis* | 37.0 | 63.44 | 60.81 |
| *Picrophilus torridus DSM 9790* | 55.0 | 35.97 | 56.44 |
| *Pyrobaculum aerophilum* | 98.0 | 51.36 | 64.83 |
| *Pyrobaculum calidifontis JCM 11548* | 95.0 | 57.15 | 69.12 |
| *Pyrobaculum islandicum DSM 4184* | 97.5 | 49.60 | 68.73 |
| *Pyrococcus abyssi* | 98.0 | 44.71 | 64.50 |
| *Pyrococcus furiosus* | 98.5 | 40.77 | 59.96 |
| *Pyrococcus horikoshii* | 95.0 | 41.88 | 68.20 |
| *Staphylothermus marinus F1* | 87.5 | 35.73 | 67.71 |
| *Sulfolobus acidocaldarius DSM 639* | 70.0 | 36.71 | 60.98 |
| *Sulfolobus solfataricus* | 87.0 | 35.79 | 63.48 |
| *Sulfolobus tokodaii* | 75.0 | 32.79 | 66.36 |
| *Thermofilum pendens Hrk 5* | 88.0 | 57.67 | 67.66 |
| *Thermoplasma acidophilum* | 57.5 | 45.99 | 56.90 |
| *Thermoplasma volcanium* | 60.0 | 39.92 | 56.93 |
| (B) Bacteria. | | | |
| *Acidobacteria bacterium Ellin345* | 25.0 | 58.38 | 56.90 |
| *Acidothermus cellulolyticus 11B* | 55.0 | 66.91 | 62.95 |
| *Acidovorax avenae citrulli AAC00-1* | 25.0 | 68.53 | 54.67 |
| *Acidovorax JS42* | 30.0 | 66.17 | 54.73 |
| *Acinetobacter sp ADP1* | 30.0 | 40.43 | 52.66 |
| *Actinobacillus pleuropneumoniae L20* | 37.0 | 41.30 | 51.60 |
| *Aeromonas hydrophila ATCC 7966* | 30.0 | 61.55 | 55.06 |
| *Alcanivorax borkumensis SK2* | 28.0 | 54.73 | 54.77 |
| *Anabaena variabilis ATCC 29413* | 25.0 | 41.42 | 53.26 |
| *Anaeromyxobacter dehalogenans 2CP-C* | 30.0 | 74.91 | 58.95 |
| *Anaplasma marginale St Maries* | 37.0 | 49.76 | 54.07 |
| *Anaplasma phagocytophilum HZ* | 37.0 | 41.64 | 52.50 |
| *Aquifex aeolicus* | 95.0 | 43.48 | 65.83 |
| *Arthrobacter aurescens TC1* | 30.0 | 62.34 | 56.88 |
| *Arthrobacter FB24* | 30.0 | 65.47 | 56.84 |
| *Aster yellows witches-broom phytoplasma AYWB* | 25.0 | 26.89 | 47.76 |

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

| | | | |
|---|---|---|---|
| *Chlorobium chlorochromatii CaD3* | 37.0 | 44.28 | 52.96 |
| *Chlorobium phaeobacteroides DSM 266* | 25.0 | 48.35 | 52.61 |
| *Chlorobium tepidum TLS* | 47.0 | 56.53 | 53.67 |
| *Chromobacterium violaceum* | 26.0 | 64.83 | 54.52 |
| *Chromohalobacter salexigens DSM 3043* | 30.0 | 63.91 | 57.88 |
| *Clostridium acetobutylicum* | 37.0 | 30.93 | 50.93 |
| *Clostridium novyi NT* | 37.0 | 28.86 | 51.63 |
| *Clostridium perfringens* | 37.0 | 28.57 | 52.44 |
| *Clostridium perfringens ATCC 13124* | 37.0 | 28.38 | 52.42 |
| *Clostridium tetani E88* | 37.0 | 28.75 | 51.83 |
| *Clostridium thermocellum ATCC 27405* | 57.5 | 38.99 | 54.12 |
| *Colwellia psychrerythraea 34H* | 10.0 | 38.01 | 51.87 |
| *Corynebacterium diphtheriae* | 37.0 | 53.48 | 55.01 |
| *Corynebacterium efficiens YS-314* | 37.0 | 63.14 | 55.68 |
| *Corynebacterium glutamicum ATCC 13032 Bielefeld* | 33.0 | 53.84 | 54.90 |
| *Corynebacterium glutamicum ATCC 13032 Kitasato* | 33.0 | 53.81 | 55.07 |
| *Corynebacterium jeikeium K411* | 37.0 | 61.40 | 56.50 |
| *Coxiella burnetii* | 37.0 | 42.66 | 57.22 |
| *Cyanobacteria bacterium Yellowstone A-Prime* | 27.0 | 60.24 | 58.38 |
| *Cyanobacteria bacterium Yellowstone B-Prime* | 27.0 | 58.45 | 58.51 |
| *Cytophaga hutchinsonii ATCC 33406* | 30.0 | 38.85 | 50.80 |
| *Dechloromonas aromatica RCB* | 30.0 | 59.25 | 54.83 |
| *Dehalococcoides CBDB1* | 35.0 | 47.03 | 55.93 |
| *Dehalococcoides ethenogenes 195* | 35.0 | 48.85 | 55.82 |
| *Deinococcus geothermalis DSM 11300* | 47.5 | 66.64 | 59.42 |
| *Desulfitobacterium hafniense Y51* | 37.0 | 47.36 | 54.22 |
| *Desulfotalea psychrophila LSv54* | 7.0 | 46.81 | 52.28 |
| *Desulfovibrio desulfuricans G20* | 55.0 | 57.84 | 55.80 |
| *Desulfovibrio vulgaris DP4* | 30.0 | 63.01 | 56.37 |
| *Desulfovibrio vulgaris Hildenborough* | 30.0 | 63.14 | 56.54 |
| *Ehrlichia canis Jake* | 37.0 | 28.96 | 49.09 |
| *Ehrlichia chaffeensis Arkansas* | 37.0 | 30.10 | 48.90 |
| *Ehrlichia ruminantium Gardel* | 37.0 | 27.51 | 49.14 |
| *Enterococcus faecalis V583* | 37.0 | 37.53 | 52.72 |
| *Erwinia carotovora atroseptica SCRI1043* | 28.0 | 50.97 | 53.86 |
| *Erythrobacter litoralis HTCC2594* | 28.0 | 63.07 | 58.41 |
| *Escherichia coli 536* | 37.0 | 50.52 | 54.82 |
| *Escherichia coli APEC O1* | 37.0 | 50.55 | 54.67 |
| *Escherichia coli CFT073* | 37.0 | 50.48 | 54.68 |
| *Escherichia coli O157H7* | 37.0 | 50.54 | 54.83 |
| *Escherichia coli O157H7 EDL933* | 37.0 | 50.44 | 54.85 |
| *Escherichia coli UTI89* | 37.0 | 50.60 | 54.77 |
| *Escherichia coli W3110* | 37.0 | 50.80 | 53.96 |
| *Francisella tularensis FSC 198* | 37.0 | 32.26 | 51.03 |
| *Francisella tularensis holarctica* | 37.0 | 32.15 | 50.51 |
| *Francisella tularensis holarctica OSU18* | 37.0 | 32.16 | 50.71 |
| *Francisella tularensis novicida U112* | 37.0 | 32.48 | 51.15 |
| *Francisella tularensis tularensis* | 37.0 | 32.26 | 51.04 |
| *Frankia alni ACN14a* | 28.0 | 72.83 | 61.30 |
| *Frankia sp. CcI3* | 27.0 | 70.08 | 61.47 |
| *Fusobacterium nucleatum* | 37.0 | 27.15 | 48.49 |
| *Geobacillus kaustophilus HTA426* | 55.0 | 52.09 | 58.35 |
| *Geobacter metallireducens GS-15* | 30.0 | 59.51 | 55.89 |
| *Geobacter sulfurreducens* | 35.0 | 60.94 | 56.79 |
| *Gloeobacter violaceus* | 27.5 | 62.00 | 57.72 |

| | | | |
|---|---|---|---|
| *Gluconobacter oxydans 621H* | 26.0 | 61.07 | 56.52 |
| *Granulobacter bethesdensis CGDNIH1* | 37.0 | 59.07 | 57.82 |
| *Haemophilus ducreyi 35000HP* | 28.0 | 38.22 | 51.40 |
| *Haemophilus influenzae* | 37.0 | 38.15 | 51.58 |
| *Haemophilus influenzae 86 028NP* | 37.0 | 38.16 | 51.57 |
| *Haemophilus somnus 129PT* | 37.0 | 37.20 | 51.65 |
| *Hahella chejuensis KCTC 2396* | 28.0 | 53.87 | 55.00 |
| *Halorhodospira halophila SL1* | 25.0 | 67.98 | 59.71 |
| *Helicobacter acinonychis Sheeba* | 37.0 | 38.18 | 51.12 |
| *Helicobacter hepaticus* | 37.0 | 35.93 | 52.52 |
| *Helicobacter pylori 26695* | 37.0 | 38.87 | 50.64 |
| *Helicobacter pylori HPAG1* | 37.0 | 39.08 | 50.97 |
| *Helicobacter pylori J99* | 37.0 | 39.19 | 51.15 |
| *Hyphomonas neptunium ATCC 15444* | 26.0 | 61.93 | 56.46 |
| *Idiomarina loihiensis L2TR* | 30.0 | 47.04 | 54.48 |
| *Jannaschia CCS1* | 25.0 | 62.33 | 57.11 |
| *Lactobacillus acidophilus NCFM* | 37.0 | 34.71 | 51.59 |
| *Lactobacillus brevis ATCC 367* | 30.0 | 46.22 | 52.33 |
| *Lactobacillus casei ATCC 334* | 30.0 | 46.62 | 52.94 |
| *Lactobacillus delbrueckii bulgaricus* | 30.0 | 49.72 | 53.53 |
| *Lactobacillus gasseri ATCC 33323* | 37.0 | 35.26 | 51.36 |
| *Lactobacillus johnsonii NCC 533* | 37.0 | 34.61 | 50.93 |
| *Lactobacillus plantarum* | 30.0 | 44.47 | 51.70 |
| *Lactobacillus sakei 23K* | 30.0 | 41.26 | 51.20 |
| *Lactobacillus salivarius UCC118* | 37.0 | 32.94 | 51.21 |
| *Lactococcus lactis* | 30.0 | 35.33 | 51.18 |
| *Lactococcus lactis cremoris MG1363* | 30.0 | 35.75 | 51.10 |
| *Lactococcus lactis cremoris SK11* | 30.0 | 35.86 | 51.01 |
| *Lawsonia intracellularis PHE MN1-00* | 37.0 | 33.28 | 53.02 |
| *Legionella pneumophila Lens* | 37.0 | 38.42 | 53.67 |
| *Legionella pneumophila Paris* | 37.0 | 38.37 | 53.78 |
| *Legionella pneumophila Philadelphia 1* | 37.0 | 38.27 | 53.72 |
| *Leifsonia xyli xyli CTCB0* | 29.0 | 67.68 | 58.12 |
| *Listeria innocua* | 37.0 | 37.44 | 53.64 |
| *Listeria monocytogenes* | 37.0 | 37.98 | 53.58 |
| *Listeria monocytogenes 4b F2365* | 37.0 | 38.04 | 53.42 |
| *Listeria welshimeri serovar 6b SLCC5334* | 37.0 | 36.35 | 53.55 |
| *Magnetococcus MC-1* | 25.0 | 54.17 | 55.01 |
| *Magnetospirillum magneticum AMB-1* | 30.0 | 65.09 | 56.94 |
| *Mannheimia succiniciproducens MBEL55E* | 37.0 | 42.54 | 52.75 |
| *Maricaulis maris MCS10* | 26.0 | 62.73 | 57.11 |
| *Marinobacter aquaeolei VT8* | 30.0 | 57.27 | 56.26 |
| *Mesoplasma florum L1* | 30.0 | 27.02 | 50.05 |
| *Mesorhizobium BNC1* | 26.0 | 61.07 | 58.14 |
| *Mesorhizobium loti* | 26.0 | 62.75 | 56.56 |
| *Methylibium petroleiphilum PM1* | 30.0 | 69.20 | 56.45 |
| *Methylobacillus flagellatus KT* | 30.0 | 55.72 | 54.95 |
| *Methylococcus capsulatus Bath* | 37.0 | 63.58 | 56.88 |
| *Moorella thermoacetica ATCC 39073* | 55.0 | 55.79 | 58.53 |
| *Mycobacterium avium 104* | 39.0 | 68.99 | 59.41 |
| *Mycobacterium avium paratuberculosis* | 39.0 | 69.30 | 59.12 |
| *Mycobacterium bovis* | 37.0 | 65.63 | 59.71 |
| *Mycobacterium bovis BCG Pasteur 1173P2* | 37.0 | 65.64 | 59.72 |
| *Mycobacterium JLS* | 30.0 | 68.36 | 58.83 |
| *Mycobacterium KMS* | 30.0 | 68.44 | 58.82 |

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

| | | | |
|---|---|---|---|
| *Prochlorococcus marinus MIT9313* | 30.0 | 50.74 | 54.64 |
| *Prochlorococcus marinus NATL1A* | 30.0 | 34.98 | 55.46 |
| *Prochlorococcus marinus NATL2A* | 30.0 | 35.12 | 55.34 |
| *Propionibacterium acnes KPA171202* | 37.0 | 60.01 | 57.59 |
| *Pseudoalteromonas atlantica T6c* | 20.0 | 44.62 | 52.23 |
| *Pseudomonas aeruginosa* | 30.0 | 66.56 | 55.35 |
| *Pseudomonas aeruginosa UCBPP-PA14* | 30.0 | 66.29 | 54.47 |
| *Pseudomonas entomophila L48* | 30.0 | 64.16 | 54.13 |
| *Pseudomonas fluorescens Pf-5* | 30.0 | 63.30 | 53.85 |
| *Pseudomonas putida KT2440* | 26.0 | 61.52 | 53.61 |
| *Pseudomonas syringae phaseolicola 1448A* | 26.0 | 58.02 | 54.01 |
| *Pseudomonas syringae pv B728a* | 26.0 | 59.23 | 54.08 |
| *Pseudomonas syringae tomato DC3000* | 26.0 | 58.40 | 54.03 |
| *Psychrobacter arcticum 273-4* | 10.0 | 42.80 | 52.08 |
| *Psychrobacter cryohalolentis K5* | 23.0 | 42.29 | 52.06 |
| *Psychromonas ingrahamii 37* | 10.0 | 40.09 | 51.63 |
| *Ralstonia solanacearum* | 28.0 | 67.04 | 54.40 |
| *Rhizobium etli CFN 42* | 30.0 | 61.27 | 55.79 |
| *Rhizobium leguminosarum bv viciae 3841* | 26.0 | 61.09 | 55.66 |
| *Rhodococcus RHA1* | 30.0 | 67.52 | 57.46 |
| *Rhodoferax ferrireducens T118* | 25.0 | 59.88 | 54.53 |
| *Rhodopseudomonas palustris BisA53* | 25.0 | 64.44 | 56.73 |
| *Rhodopseudomonas palustris BisB18* | 25.0 | 64.96 | 57.29 |
| *Rhodopseudomonas palustris BisB5* | 25.0 | 64.81 | 58.00 |
| *Rhodopseudomonas palustris CGA009* | 25.0 | 65.04 | 55.92 |
| *Rhodopseudomonas palustris HaA2* | 25.0 | 66.04 | 58.41 |
| *Rhodospirillum rubrum ATCC 11170* | 25.0 | 65.45 | 57.60 |
| *Rickettsia bellii RML369-C* | 37.0 | 31.65 | 50.87 |
| *Rickettsia conorii* | 37.0 | 32.44 | 50.47 |
| *Rickettsia felis URRWXCal2* | 37.0 | 32.45 | 50.46 |
| *Rickettsia prowazekii* | 37.0 | 29.00 | 50.27 |
| *Rickettsia typhi wilmington* | 37.0 | 28.92 | 49.62 |
| *Roseobacter denitrificans OCh 114* | 20.0 | 58.97 | 55.95 |
| *Rubrobacter xylanophilus DSM 9941* | 60.0 | 70.48 | 64.37 |
| *Saccharophagus degradans 2-40* | 28.0 | 45.83 | 54.26 |
| *Salinibacter ruber DSM 13855* | 37.0 | 66.22 | 59.15 |
| *Salmonella enterica Choleraesuis* | 37.0 | 52.16 | 55.10 |
| *Salmonella enterica Paratypi ATCC 9150* | 37.0 | 52.16 | 54.42 |
| *Salmonella typhi* | 37.0 | 52.09 | 54.75 |
| *Salmonella typhimurium LT2* | 37.0 | 52.22 | 54.73 |
| *Shewanella amazonensis SB2B* | 37.0 | 53.59 | 54.24 |
| *Shewanella ANA-3* | 30.0 | 48.05 | 53.62 |
| *Shewanella baltica OS155* | 25.0 | 46.28 | 53.51 |
| *Shewanella denitrificans OS217* | 28.0 | 45.15 | 53.16 |
| *Shewanella frigidimarina NCIMB 400* | 20.0 | 41.58 | 52.66 |
| *Shewanella MR-4* | 30.0 | 47.89 | 53.60 |
| *Shewanella MR-7* | 30.0 | 47.87 | 53.63 |
| *Shewanella oneidensis* | 30.0 | 45.96 | 53.55 |
| *Shewanella W3-18-1* | 30.0 | 44.63 | 53.70 |
| *Shigella boydii Sb227* | 37.0 | 51.21 | 54.89 |
| *Shigella dysenteriae* | 37.0 | 51.25 | 54.71 |
| *Shigella flexneri 2a* | 37.0 | 50.89 | 54.79 |
| *Shigella flexneri 2a 2457T* | 37.0 | 50.91 | 54.48 |
| *Shigella flexneri 5 8401* | 37.0 | 50.92 | 54.87 |
| *Shigella sonnei Ss046* | 37.0 | 51.01 | 54.98 |

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

| | | | |
|---|---|---|---|
| *Thermosynechococcus elongatus* | 55.0 | 53.92 | 55.94 |
| *Thermotoga maritima* | 80.0 | 46.25 | 63.85 |
| *Thermus thermophilus HB27* | 75.0 | 69.44 | 64.53 |
| *Thermus thermophilus HB8* | 75.0 | 69.52 | 64.68 |
| *Thiobacillus denitrificans ATCC 25259* | 30.0 | 66.07 | 55.77 |
| *Thiomicrospira crunogena XCL-2* | 25.0 | 43.13 | 52.20 |
| *Thiomicrospira denitrificans ATCC 33889* | 22.0 | 34.46 | 50.33 |
| *Treponema denticola ATCC 35405* | 37.0 | 37.87 | 52.17 |
| *Treponema pallidum* | 35.0 | 52.78 | 54.27 |
| *Trichodesmium erythraeum IMS101* | 37.0 | 34.14 | 53.19 |
| *Tropheryma whipplei TW08 27* | 37.0 | 46.31 | 57.87 |
| *Tropheryma whipplei Twist* | 37.0 | 46.33 | 57.75 |
| *Ureaplasma urealyticum* | 37.0 | 25.50 | 45.80 |
| *Wigglesworthia brevipalpis* | 27.5 | 22.48 | 48.34 |
| *Wolbachia endosymbiont of Brugia malayi TRS* | 27.5 | 34.18 | 48.40 |
| *Wolbachia endosymbiont of Drosophila melanogaster* | 27.5 | 35.23 | 48.73 |
| *Wolinella succinogenes* | 37.0 | 48.46 | 52.24 |
| *Xanthomonas campestris 8004* | 26.0 | 64.96 | 56.13 |
| *Xanthomonas campestris vesicatoria 85-10* | 26.0 | 64.75 | 56.18 |
| *Xanthomonas citri* | 26.0 | 64.77 | 56.55 |
| *Xanthomonas oryzae KACC10331* | 26.0 | 63.69 | 56.29 |
| *Xanthomonas oryzae MAFF 311018* | 26.0 | 63.70 | 56.22 |
| *Xylella fastidiosa* | 26.0 | 52.67 | 55.16 |
| *Xylella fastidiosa Temecula1* | 26.0 | 51.78 | 55.37 |
| *Yersinia pestis Antiqua* | 37.0 | 47.70 | 53.75 |
| *Yersinia pestis CO92* | 37.0 | 47.64 | 53.90 |
| *Yersinia pestis KIM* | 37.0 | 47.64 | 54.11 |
| *Yersinia pestis Nepal516* | 37.0 | 47.58 | 53.78 |
| *Yersinia pseudotuberculosis IP32953* | 37.0 | 47.61 | 53.75 |
| *Zymomonas mobilis 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 Acel_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 | cysteinyl-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 Acel_0155 |
| 117927376 | UDP-glucose pyrophosphorylase |
| 117927377 | molybdopterin molybdochelatase |
| 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 Acel_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, SecG 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 | ferrochelatase |
| 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 ChlI |
| 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 | phosphoribosylformylglycinamidine cyclo-ligase |
| 117929279 | amidophosphoribosyltransferase |
| 117929281 | phosphoribosylformylglycinamidine synthase II |
| 117929282 | phosphoribosylformylglycinamidine synthase I |
| 117929283 | phosphoribosylformylglycinamidine 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 |