LA-UR-

Approved for public release; distribution is unlimited.

Title:	Complete genome of the cellulolytic thermophile Acidothermus cellulolyticus 11B provides insights into its ecophysiological and evolutionary adaptations
Author(s):	Ravi D. Barabote, Gary Xie, David Leu, Philippe Normand, Anamaria Necsulea, Vincent Daubin, Claudine Médigue, William S. Adney, Xin Clare Xu, Alla Lapidus, Chris Detter1, Pierre Pujic, David Bruce, Jean F. Challacombe, Thomas S. Brettin, Paul Richardson, and Alison M. Berry.
Intended for:	Genome Research

Los Alamos
NATIONAL LABORATORY
EST. 1943

Form 836 (7/06)

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 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], Pierre Pujic[3], David Bruce[1], Jean F. Challacombe[1], Thomas S. Brettin[1], Paul Richardson[6], 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, thermophile, glycoside hydrolases, genomic islands, flagellar genes.

ABSTRACT

We present here the complete 2.4 MB genome of the actinobacterial thermophile, *Acidothermus cellulolyticus* 11B, that surprisingly reveals thermophilic amino acid usage in only the cytosolic subproteome rather than its whole proteome. Thermophilic amino acid usage in the partial proteome implies a recent, ongoing evolution of the *A. cellulolyticus* genome since its divergence about 200-250 million years ago from its closest phylogenetic neighbor *Frankia*, a mesophilic plant symbiont. Differential amino acid usage in the predicted subproteomes of *A. cellulolyticus* likely reflects a stepwise evolutionary process of modern thermophiles in general. An unusual occurrence of higher G+C in the non-coding DNA than in the transcribed genome reinforces a late evolution from a higher G+C common ancestor. Comparative analyses of the *A. cellulolyticus* genome with those of *Frankia* and other closely-related actinobacteria revealed that *A. cellulolyticus* genes exhibit reciprocal purine preferences at the first and third codon positions, perhaps reflecting a subtle preference for the dinucleotide AG in its mRNAs, a possible adaptation to a thermophilic environment. Other interesting features in the genome of this cellulolytic, hot-springs dwelling prokaryote reveal streamlining for adaptation to its specialized ecological niche. These include a low occurrence of pseudogenes or mobile genetic elements, a flagellar gene complement previously unknown in this organism, and presence of laterally-acquired genomic islands of likely ecophysiological value. New glycoside hydrolases relevant for lignocellulosic biomass deconstruction were identified in the genome, indicating a diverse biomass-degrading enzyme repertoire several-fold greater than previously characterized, and significantly elevating the industrial value of this organism.

Supplementary figures and tables are included as a separate PDF file.

INTRODUCTION

Acidothermus cellulolyticus is an industrially important, plant biomass degrading, eubacterial organism that 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 for bioconversion (Mohagheghi et. al., 1986). *A. cellulolyticus* produces cellulose-degrading enzymes, many of which are thermostable (Adney et al., 1994 and 1995; Baker et al., 1994; Tucker et al., 1992). One of the endoglucanases, E1 endoglucanase, which has been crystallized, is highly thermostable to 81°C and has very high specific activity on carboxymethylcellulose (Himmel et al., 1994; Sakon et al., 1996; Thomas et al., 1995). Several plant biomass deconstructing glycoside hydrolases from this organism have been patented (Adney et al., 1994; Adney et al., 1998; Clarkson et al., 1999; Ding et al., 2006; Himmel et al., 1994; Himmel et al., 1995; Tucker et al. 1992). Although the glycoside hydrolases of *A. cellulolyticus* are of great potential industrial value because of their thermostability, knowledge of the biology of the organism to date is limited.

A. cellulolyticus is a member of the Frankineae, a high G+C, primarily Gram-positive Actinobacteria group (Rainey and Stackebrandt, 1993). It is a non-sporulating, heterotrophic obligate aerobe (Mohagheghi et. al., 1986). *A. cellulolyticus* 11B is acidophilic (pH 4-6, with optimal pH 5.5) and thermophilic (growth between 37° and 70°C; optimal growth temperature [OGT] is 55°C). None of the characterized strains of *A. cellulolyticus* grow below 37°C (Mohagheghi et al., 1986). This makes the evolutionary context of *A. cellulolyticus* interesting, because its closest known phylogenetic neighbor is the mesophilic actinobacterium, *Frankia,* based on the analysis of the 16S rRNA, *recA*, and *shc* nucleotide sequences (Alloisio et al, 2005; Marechal et al, 2000; Normand et al., 1996). *Frankia* is a mesophilic (OGT of 26°C), nitrogen-fixing soil organism that forms symbiotic root nodule associations with plants (Benson 1988). Considering their contrasting environments, the genetic distance between *A. cellulolyticus* and 3 different *Frankia* strains, ACN14a, CcI3 and EAN1pec, is very small (just 0.050, 0.053, and 0.055 nucleotide substitutions per site, respectively, based on 16S rRNA gene sequence comparison), and comparable to that found between certain strains of *Frankia* species (0.053 between

strains AgB19 and Dryas; Fig. S1). 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). Consequently, genomic comparison of *A. cellulolyticus* with the mesophilic as well as thermophilic actinobacteria could provide useful insights into the nature of adaptation of this aquatic thermophile.

Comparative analyses of in the genomes of hyperthermophilic, thermophilic and mesophilic prokaryotes have facilitated the identification of unique characteristics in the genomes and the predicted proteomes of organisms that grow at high temperatures (Haney et al., 1999; Kreil and Ouzounis, 2001; Pascal et al., 2005; Suhre and Claverie, 2003; Takami et al., 2004). Some of the features that have been identified include, distinct patterns of synonymous codon usage for several amino acids, particularly arginine and leucine, in thermophiles (Singer and Hickey, 2003), increased purine-loading of RNAs in thermophiles (Lao and Forsdyke, 2000), increase in glutamic acid content with concomitant decrease in glutamine along with correlation between the increase in glutamic acid and an increase in the pool of lysine+arginine amino acids (Tekai et al., 2002), increase in the E+K/Q+H amino acids ratio with growth temperature (Farias and Bonato, 2003), and elevated content of hydrophobic amino acids in the predicted proteomes of thermophiles (Leiph et. Al., 2006). Principal component analysis (PCA; or correspondence analysis [CA], a similar method) of the global composition of all twenty individual amino acids in completely sequenced organisms has been widely used to visualize and identify the discriminating patterns in the predicted proteomes. In these analyses, the first principal component (PC1) correlates with the G+C content of the genomes, while the second PC (PC2) correlates with the optimal growth temperature (OGT) of the organisms. Although these powerful analyses have elucidated the influence of environmental temperature on the amino acid frequencies in the organisms, they suffer a weakness when

studying moderate thermophiles. In particular, although the distinction of hyperthermophiles from mesophiles is very clear along PC2 in the analyses, a lack of good separation between moderately thermophilic and mesophilic organisms is a noticeable drawback. In fact, some thermophiles display amino acid usage that is close to the borderline distinguishing thermophiles from mesophiles (Takami et al., 2004), while a few other thermophiles have been found to cluster together with mesophiles (Lobry and Chessel, 2003). Such global analyses of prokaryotes also raise other concerns, such as (a) most of the variance along PC2 is due to the highly skewed amino acid usage in the extreme thermophiles which makes it difficult to visualize and assess the subtle separation between the moderate thermophiles and mesophiles along PC2, (b) the distinct phylogenetic and evolutionary histories of archaea and bacteria raise questions about their inclusion in a single dataset, and (c) the global amino acid usage in an organism may not be truly representative of the amino acid composition of all of its proteins localized to different cellular compartments. This emphasizes the need for adopting improved approaches for understanding adaptation in moderately thermophilic proteomes.

We present the complete genome of *Acidothermus cellulolyticus* 11B (ATCC 43068; Genbank accession NC_008578). The genome reveals a much larger repertoire of glycoside hydrolases indicating a far greater plant biomass deconstructing capability of the organism than previously recognized. Our analysis of the *A. cellulolyticus* provides new information about the physiology and evolutionary aspects of the organism and reveals several attributes which likely contribute to the ecophysiological adaptation of the organism. We discuss these results in light of the phylogenetic relationship of *A. cellulolyticus* to other actinobacteria. In addition, we present an improved approach to increasing the resolution between thermophiles and mesophiles in PCA analyses of amino acid usage. Our study addresses the concern of how analysis of whole proteomes may obscure underlying patterns in a subset of proteins and reveals that analysis of the cytosolic subproteome of *A. cellulolyticus*, rather than its entire proteome, offers deeper understanding of thermoadaptation in the organism.

RESULTS

General Genome Characteristics. An overview of the *A. cellulolyticus* genome features in comparison with the genomes of *Frankia*, *Streptomyces* and *T. fusca* is provided in Table 1. The 2.44 megabase (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). 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 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. One fifth of all the predicted

proteins are annotated as either hypothetical proteins, conserved hypothetical proteins or proteins with unknown function. No identifiable prophages or phage-related proteins are found in the genome and two genes encoding fragments of a single transposase (Acel_1666, Acel1667) were found in the genome. Approximately 8% of the proteins (171 proteins) in the genome do not show sequence similarity to any proteins in the NCBI database and thus appear to be ORFans unique to *A. cellulolyticus* (Fig. S2). Analysis of the phyletic distribution of blast hits (Fig. S2) of the remaining proteins revealed that the majority (approx. 80%) of the *A. cellulolyticus* proteins show highest sequence similarity to proteins from other actinobacteria; a significant proportion of the remainder show highest sequence similarity to proteins from proteobacteria (~6%), green non-sulfur bacteria (~2%), and Firmicutes (~1%). Within the actinobacterial hits, the highest numbers of best blast hits, surprisingly, are to the phylogenetically more remote *Streptomyces* (~18%), more so than to its closest phylogenetic neighbor *Frankia* (~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 S3).

Plant Biomass Degradation Capabilities. *A. cellulolyticus* is known to secrete a full complement of cellulolytic enzymes (Mohagheghi et al. 1986; Adney et al., 1994; Tucker et al., 1992). It is evident from the genome that *A. cellulolyticus* produces numerous additional enzymes for the breakdown of other carbohydrates, such as xylans and possibly chitin. At least 35 genes encoding predicted glycoside hydrolases (GH) could be identified in the genome (Table 2). A detailed analysis of the domains in the predicted proteins (Table 2) revealed that the predicted GH enzymes in *A. cellulolyticus* belong to 17 different GH families (Henrissat, 1991; Coutinho and Henrissat, 1999; the Carbohydrate Active Enzymes database at http://www.cazy.org/). Although, several of these enzymes have been previously characterized (Adney et al., 1994 and 1995; Baker et al., 1994; Sakon et al., 1996; Thomas et al., 1995; Tucker et al., 1989 and 1992), genome analysis has revealed the presence of additional genes potentially involved in plant biomass degradation. The genes are found at various locations in the genome; some of

these genes occur close together on the chromosome while others occur as orphan genes. Like the genome of *T. fusca* the *Acidothermus* genome contains both a reducing-end specific Family 48 exoglucanase and a non-reducing-end specific Family 6 exocellulase (Lykidis et al., 2007). Two xylanase-encoding genes (GH family 10) as well as a xylosidase were identified in the genome. Although xylanases from GH families 5, 7, 8, 10, 11 and 43 have been identified to date, only xylanases from families 10 and 11 have been well studied (Collins et al., 2005). It has been observed that the endo-1,4-b-xylanases may be active on xylans and on low molecular mass cellulose substrates at lower catalytic efficiency (Biely, 2003; and Gilkes, et al., 1991), in particular on aryl-cellobiosides (Biely, et al., 1997; and van Tilbeurgh et al., 1985) and certain cello-oligosaccharides (Claeyssens and Henrissat, 1992; and Biely, 2003). The observed presence in the genome of xylanases, α-amylase genes (GH 13), and the absence of genes for pectin degradation are all supported by previously-reported growth experiments (Mogagheghi et al. 1986). The functions of six predicted chitinase-encoding genes (Table 2) belonging to GH family 18 remain to be confirmed experimentally. The capability to degrade chitin could permit degradation of fungal and insect biomass. Unlike other eukaryotic cell-wall biopolymers, chitin contains nitrogen and hence the organism could use it as a carbon and nitrogen source. The ability to utilize chitin could offer a survival edge under varying nutritional conditions.

Genomic Islands. A sliding window plot of the percent G+C content in the *A. cellulolyticus* genome, together with an analysis of deviations from the genomic signature along the *A. cellulolyticus* chromosome, has revealed three genomic islands with significantly lower G+C content than the rest of the genome and with deviant dinucleotide signature (Fig. 2). The three genomic islands, described below, were analyzed for the possible source and functions of the genes contained within these regions.

Genomic Island 1 (Acel_0569-Acel_0583) consists of 15 genes (Table 3), all of which have lower than average G+C as well as deviant dinucleotide signature. The average %G+C of the genes in this island is approximately 58%. Many of the encoded proteins in this cluster have no recognizable orthologs

in *Frankia*, *Streptomyces*, or other actinobacteria. The first five genes appear to 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 the island (Acel_0576-Acel_0582) contains genes involved in sugar metabolism and uptake.

Genomic Island 2 (Acel_0810-Acel_0825) contains 18 genes with moderately low G+C content (an average of 62.5%) and is flanked by tRNA genes at both ends (Table 3). Several of the proteins encoded in this region do not appear to have homologs in actinomycetes. Especially of note, *Frankia* and *Streptomyces* both lack orthologs of these proteins. About half of the genes do not have a recognizable function. Many of the remaining genes in this region appear to encode putative homologs of the genes *vrlI (Acel_0810)*, *vrlJ (Acel_0811)*, *vrlK (Acel_0812)*, *vrlP (Acel_0817)*, and *vrlQ (Acel_0818)* that are found in the virulence associated locus (*vrl*) that is preferentially associated with the virulent strains of *Dichelobacter nodosus* (Billington et al., 1999). Orthologs of the *vrl* genes have been identified in many bacterial groups and the locus is thought to be laterally transferred via a bacteriophage (Knaust et al. 2007). The *vrlI* and *vrlJ* homologs in *Acidothermus* have DNA-binding and ATPase domains, respectively, while the other putative *vrl* homologs in *Acidothermus* do not have recognizable domains. With respect to the intervening proteins, Acel_0813 protein is a transcriptional regulator containing a helix-turn-helix motif, Acel_0814 shows weak homology to DNA methylases, Acel_0815 is a hypothetical protein, and Acel_0816 protein has a helicase domain and could be a homolog of the VrlO protein although the homology is undetectable at sequence level. The *A. cellulolyticus* proteins do not show clear sequence homology to the *S. coelicolor* phase-variable phage growth limitation (Pgl) system, in contrast to the *D. nodosus vrl* locus (Billington et al. 1999), consistent with the observation that most gene products encoded in this gene cluster show high homology to proteins in low G+C Gram-positives, namely *Bacteroides*, *Nitrosococcus*, and *Thermoanerobacter*, rather than to the close phylogenetic neighbors within the actinobacterial group.

Genomic Island 3 (Acel_1621-Acel_1649). This gene cluster consists of 31 genes (Table 3) and is flanked by $tRNA^{Arg}$ gene upstream and by $tRNA^{His}$ gene downstream. The %G+C of the genes in this island varies from 51 to 70%, with an average of about 61.7%. Approximately a third of the genes in this cluster encode proteins with no functional annotation. Of the remaining genes in this island, three (Acel_1626, Acel_1640, Acel_1641) encode proteins involved in ABC transport, the latter two of which are predicted to be involved in the uptake of amino acids. Acel_1633 - Acel_1639 appears to form an operon of seven genes. Although Acel_1633 is annotated as a PurC domain protein, both Acel_1633 and Acel_1634 appear to encode proteins with unknown function. Acel_1635, Acel_1639 encode enzymes involved in amino acid metabolism. Acel_1636, Acel_1637, and Acel_1638 encode subunits of the carbon monoxide (CO) dehydrogenase family proteins. Acel_1642 - Acel_1645 apparently form an operon of four genes that encode an aldehyde oxidase, a coenzyme A transferase, glutaconate coA-transferase and a luciferase family protein, respectively. Several genes in this genomic island (such as 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. With the exception of Acel_1626, homologs of these above six proteins do not occur in *Frankia* sp., the closest phylogenetic neighbor of *A. cellulolyticus*.

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 S4.

Flagella and Motility. Immediately downstream of genomic island 2, we identified a stretch of 37 genes (Acel_0828 – Acel_0864) that do not have any homologs in *Frankia*, *Streptomyces* or *T. fusca*. This region encodes a complete set of genes coding for flagellar biosynthesis and motility (Fig. S3). The genes are organized into two divergent operons. Most of the flagellar structural genes are organized in an operon 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 in the other operon, the last gene in a five-gene operon that is immediately upstream and in the opposite direction to the flagellar biosynthesis

operon. Thus far, only three other actinomycetes, *Nocardioides* sp. JS514, *Kineococcus radiotolerans*, and *Leifsonia xyli*, encode sequence homologs of these genes. The gene content and order of the flagellar operon is highly conserved between *A. cellulolyticus* and *Nocardioides*, while minor differences in gene order are observed in *Kineococcus* (Fig. S3). Many of the 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). 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.

Thermophilic adaptations of the genome. The synonymous as well as global codon usage differences between *Acidothermus* and *Frankia* are very subtle (Supplementary Table S2). Comparison of the closely related thermophilic (*A. cellulolyticus* and *T. fusca*) and mesophilic actinobacteria (*Frankia alni* ACN14a, *Frankia* sp. CcI3, *S. avermitilis*, and *S. coelicolor*) revealed that the differences in the usage of each of the 64 codons do not always follow the differences in G+C content in the coding region of their genomes. However, neither *A. cellulolyticus* nor *T. fusca* show a strong thermophilic codon usage pattern (Supplementary Figs. S4A and S4B). Nevertheless, a comparison of the relative abundances of the four nucleotides at each of the three codon positions revealed 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 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). Such a bias could influence the usage of GNN, ANN, NNA, and NNG codons, which include 48 of the 64 codons. In fact, the most noticeable and evolutionarily significant codon usage differences included some of the above codons, especially GNA and GNN codons. The CNC, CNT, TNC, and TNT codons do not show any preferential bias in these organisms, based on either their OGT or the G+C content of the coding region in their genome. Interesting differences were observed for the GNA and ANG codons. Of the four GNA codons (encoding the four amino acids – E, A, G, and V),

the GAA, GCA, and GGA codons (encoding E, A and G, respectively) showed increased representation in the two thermophiles, with the increase being most prominent for the GAA codon (for glutamate). Of the four ANG codons (for the four amino acids - K, M, R, T), the AGG codon (for arginine) is clearly less preferred in *A. cellulolyticus* and *T. fusca*, unlike other (hyper)thermophiles (Lynn et al., 2002; Lobry and Chessel, 2003; Lobry and Necsulea, 2006; Singer and Hickey, 2003).

Thermophilic adaptations of the proteome. Principal component analysis of the amino acid distribution in 416 prokaryotic genomes showed a separation of hyperthermophiles from the non-hyperthermophiles, along the second principal component axis (PC2), correlating with OGT (Fig. 3). However, the separation between thermophiles and mesophiles was weak. Analysis of the component loadings (data not shown) showed that the separation along the PC2 was mainly due to the differences in the total fraction of IVYWREL amino acids in the organisms. *A. cellulolyticus* was close to the borderline between the thermophiles and mesophiles in this analysis, even though it was evident that *Frankia* and *Streptomyces* tended to be positioned more in the direction of the mesophiles, while *A. cellulolyticus* and *T. fusca* were closer to the thermophiles along PC2.

Using PCA, we further analyzed various "sub-proteomes" from the *A. cellulolyticus* genome, namely the proteins encoded in each of the three genomic islands (Fig. 3A), as well as the predicted cytosolic, membrane, and secretome fractions of the proteome (Fig. 3B). All fractions showed deviations from the *A. cellulolyticus* whole proteome on the PCA plot. Two of the three genomic islands found in *A. cellulolyticus* (Islands 1 and 2; see section below on Genomic Islands) were shifted in the direction of the hyperthermophiles, while genomic island 3 clustered with mesophiles (Fig. 3A). The membrane fraction as well as the cytosolic fraction were shifted noticeably in the direction of the (hyper)thermophiles, while the secretome fraction clustered near the mesophiles (Fig. 3B). The membrane and the secretome fractions are expected to have a skewed amino acid composition compared to the whole proteome because the hydrophobic transmembrane segments in membrane proteins and the signal peptide in secreted proteins as well as the secretion apparatus likely impose constraints on the amino acid usage of

the two respective sub-proteomes. However, the cytosolic fraction showed interesting deviation. The difference between the PC2 value for the cytosolic fraction and that for the whole proteome of *A. cellulolyticus* is 0.8 along the PC2 scale, in agreement with the higher proportion of IVYWREL in its cytosolic proteins as compared to its whole proteome. Although a similar general trend was seen in the cytosolic, membrane, and secretome fractions in *Frankia* and *Streptomyces* on the PCA plot (Supplementary Fig. S5 and S6, respectively), the separation between the cytosolic fraction and the whole proteome was not as prominent in the case of *Frankia* or *Streptomyces*.

Analysis of the amino acid composition of 478 conserved orthologous proteins from *A. cellulolyticus*, *Frankia* sp. (strains ACN14a, CcI3), *S. avermitilis*, *S. coelicolor*, and *T. fusca* further revealed that both *A. cellulolyticus* and *T. fusca* orthologs contain a higher proportion of IVYWREL amino acids (Supplementary Table S5) compared to the two mesophilic organisms. Moreover, an extended analysis of 47 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).

Protein composition, OGT and G+C. The six organisms chosen in our study showed a negative correlation between the total fraction of IVYWREL amino acids in their proteomes and the total chromosomal G+C content (Supplementary Fig. S7A). On the other hand, an equally strong negative correlation was seen between the G+C content and the OGT of these organisms (Supplementary Fig. S7B). Unlike the observed IVYWREL content, the expected (theoretical) total fraction of IVYWREL amino acids, computed based on the G+C content of the coding region in each of the organisms, did not show as strong a correlation with either the chromosomal G+C content (Supplementary Fig. S7C) or the G+C content of the coding regions in the organisms (Supplementary Fig. S7D). Contrary to the established idea that G+C content of the genome is a major determinant of the amino acid usage in an organism, it is possible that a selection pressure at the level of the observed amino acid composition in *A.*

cellulolyticus could be a determining factor in the lower G+C content of its genome as compared to those of *Frankia*. Indeed, when we altered *Frankia* coding DNA *in silico* that consequently encoded a predicted *A. cellulolyticus* proteome, the resulting projected genes contained lower G+C than the G+C content of the coding region in each of the original *Frankia* genomes (Table 5).

DISCUSSION

The relatively small genome of *A. cellulolyticus* with very few pseudogenes or mobile genetic elements (see Table 1) appears to be streamlined for adaptation to its ecological niche, consistent with previous observations that organisms that possess smaller genomes and are adapted to specialized environments usually possess few pseudogenes or mobile genetic elements (Ochman and Davalos, 2006). 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) appear to 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 reflects the role of these mobile elements in their genome expansion, as described for *Frankia* (Normand et al., 2007), and also that genome reduction events accompanied by the loss of mobile elements have resulted in a small genome size of *A. cellulolyticus*.

The *A. cellulolyticus* genome encodes a parsimonious complement of the 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 (citation). 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.

The *A. cellulolyticus* genome reveals several attributes that may enable the organism to adapt to its environment. Many proteins encoded by genes scattered throughout the genome show highest sequence similarity to proteins from distantly related thermophilic bacteria and archaea, supporting the likelihood that they were laterally acquired, and that the assimilation of some of these genes may have facilitated a thermophilic lifestyle. Such a hypothesis is further supported by the elucidation of the three laterally acquired genomic islands that carry eco-physiologically relevant genes with close homologs exclusively in thermophilic bacteria. The three genomic islands in *A. cellulolyticus* are characterized by lower G+C content and deviation from the genomic signature. Genomic signature is a measure of the relative dinucleotide abundances in the genome (Karlin 2001). Regions in the genome that deviate significantly from the average dinucleotide profile of the genome 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 of homologs in other actinobacteria strongly suggests that these DNA regions have been horizontally acquired in *A. cellulolyticus*. Several genes in the three islands show highest sequence similarity to proteins from thermophilic organisms, raising the possibility that these horizontally acquired genes may have played an important role in the evolution and eco-adaptation of this thermophile.

Analysis of the genes encoded within the three genomic islands also suggests a functional role for the acquired genes in the context of the organism's ecology. For example, some of the genes encoded on GI1 may contribute to degradation of plant material and uptake of the products. Aryldialkyl phosphatases catalyze 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). The precise function of each of the *vrl* genes as well as the role of the entire *vrl* locus in the virulence of *Dichelobacter nodosus* is unclear; therefore, it is difficult to understand the functional implication of these genes in *A. cellulolyticus*. However, the homology and predicted annotations of many of these genes in *A. cellulolyticus* suggest that they could be involved in DNA restriction and modification, functions that are important in resistance mechanisms against infection by bacteriophages (Hoskisson and Smith, 2007). Therefore, these genes could offer immunity to *A. cellulolyticus* against phage infection, similar to the phage resistance Pgl system in *S. coelicolor* Pgl (Sumby and Smith, 2002). 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 distinguish whether the various homologs of the CO dehydrogenase family found on GI3 function in carbon fixation or in purine salvage. However, either of these possibilities may add eco-physiological value to *A. cellulolyticus*.

The DNA as well as the proteome of *A. cellulolyticus* reveal characteristics suggesting an ongoing evolution in the thermophilic environment. The relative increase in the G and A nucleotides at the first and third codon positions, respectively, in the *A. cellulolyticus* genes could provide subtle

thermophilic adaptation by increasing the occurrence of AG dinucleotides in the mRNAs, simply due to the likely 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 explains the lack of separation of *A. cellulolyticus* from the mesophiles, along the second axis in our correspondence analysis of global and synonymous codon usage. The AGG codon is known to strongly influence the separation between thermophiles and mesophiles (Lynn et al., 2002; Lobry and Chessel, 2003; Lobry and Necsulea, 2006; Singer and Hickey, 2003). *A. cellulolyticus* is clearly an exception in the use of AGG codons compared to other thermophiles.

Principle component analysis of the usage of all 20 amino acids in whole proteomes segregated hyperthermophilic bacteria from mesophilic bacteria along the second principal axis (Supplementary Fig S8), but the separation of thermophiles from mesophiles using PCA was poor in these analyses. To improve this resolution, our approach combining the amino acid proportions for 7 of the 20 amino acids increased the resolution along PC2, a correlation with OGT shown previously in other organisms (Zeldovitch et al. 2007). However, unlike hyperthermophiles, the amino acid usage in thermophiles still did not appear to be unambiguously separable from that of the mesophiles. Since amino acid usage of an organism is not only influenced by its G+C content but is likely also partially predetermined by its phylogenetic origin, comparing the position of an organism on the PCA plot relative to its phylogenetic neighbors proved to be useful. In this analysis, the two thermophiles with relatively elevated total fraction of IVYWREL, *A. cellulolyticus* and *T. fusca*, were closer towards the thermophiles along PC2, compared to *Frankia* and *Streptomyces* as well as other mesophilic actinobacteria (see Fig. 3). Further, orthologous proteins from *A. cellulolyticus* and *T. fusca* also have higher IVYWREL content compared to *Frankia* and *Streptomyces*.

The six closely-related actinobacterial genomes in our comparative study may pose an apparent paradoxical situation for interpreting thermophilic adaptations since there is an unusually strong correlation between their OGT and the G+C content of their genomes, and it has been well documented

that neither the OGT of prokaryotes nor the fraction of IVYWREL in their proteomes correlate with the G+C content of the organism's genome (Zeldovich et al., 2007). In addition, the expected (theoretical) proportion of IVYWREL amino acids in the six actinobacteria did not show a strong statistically significant correlation ($p = 0.06$) with their genomic G+C content. Therefore, we postulate that the nucleotide content of the *A. cellulolyticus* genome is not solely responsible for the thermophilic amino acid usage in this organism. This supposition is further supported by our observation that orthologous proteins from diverse actinobacteria with varying G+C contents also showed a positive correlation between the IVYWREL content and the OGT.

It is possible that the lower G+C content in *A. cellulolyticus*, compared to the *Frankia* genomes, is a result of evolutionary pressure on the proteome's amino acid usage. Although there is no consensus yet about whether mutational pressure or selective pressure truly determines the nucleotide compositions of genomes, it is likely that different forces shape the nucleotide composition in different organisms. Influence of selection at the amino acid level on the nucleotide composition has been noted (Necsulea and Lobry, 2006). The *A. cellulolyticus* genome also suggests that a selective pressure on the amino acid composition of the proteome may be responsible for the lower G+C composition of the *A. cellulolyticus* genome compared to its closest phylogenetic neighbor *Frankia*. In *A. cellulolyticus*, as well as in most prokaryotes, the G+C composition of the genome is determined mostly by the G+C composition of the coding region, since the coding region constitutes about 90% of the genome. In most organisms the G+C content in the non-coding region is generally lower than the G+C content of the coding genome (Sandberg et al., 2003). On the contrary, the non-coding fraction of the *A. cellulolyticus* genome has higher G+C content than the coding fraction, just opposite to what is seen in *Frankia* and other bacteria (data not shown). This suggests that the ancestral DNA of *A. cellulolyticus* probably had higher G+C and that a selective evolutionary pressure of a hot spring environment on the protein amino acid composition shaped the G+C content of present day *A. cellulolyticus*. Such a probability was modeled *in silico* by altering the coding DNA from the two *Frankia* genomes (ACN14a and CcI3). While maintaining the synonymous codon usage in the respective *Frankia* genomes, we altered the codon frequencies to derive a

predicted protein with the amino acid usage observed in *A. cellulolyticus*. Theoretically, the G+C content the DNA altered in such a way could have one of the three fates: (1) no change in G+C, (2) higher G+C, or (3) lower G+C. The G+C content of the manipulated coding DNA of both *Frankia* sp. was lower than the G+C content observed in the real *Frankia* genomes. This suggests that the lower G+C in *A. cellulolyticus* may be a result of protein evolution favored by adaptation to its ecological niche. Interestingly, genome shrinkage accompanied by accelerated protein evolution and a sharp reduction in G+C has also been observed in a free-living marine bacterium, *Prochlorococcus* spp. (Dufresne et al., 2005).

It is likely that adaptation to thermophily is a slow and on-going process. Consequently, the degree of separation along PC2 in Figure 3 could suggest how recently a thermophile has evolved. Our data would therefore argue that *A. cellulolyticus* is a recent thermophile, as its proteome still shows a meso-thermophilic amino acid usage and that it is slowly but continually evolving to adapt to the thermophilic environment. It is possible that certain proteins evolve faster towards a thermophilic amino acid usage than other proteins in an organism. As we show, the cytosolic fraction in *A. cellulolyticus* shows a greater tendency towards thermophilic amino acid usage than its whole proteome. In addition, the cytosolic fraction in *A. cellulolyticus* showed a greater shift from the whole proteome along PC2, as compared to *Frankia* and *Streptomyces*. This could reflect evolutionary as well as physiological significance, because conceivably, in an extreme environment such as the hot spring, rapid evolution and adaptation of the cytosol, more so than the membrane fraction or secretome, may have a direct and critical influence on the survival of an organism. Thus, although the amino acid usage in the overall proteome of *A. cellulolyticus* shows only a weak thermophilic pattern, a subset of the proteome (cytosolic fraction) has amino acid composition that is more typical of a thermophile. This supports our hypothesis that the *A. cellulolyticus* proteome may be in a process of continual evolution towards thermophilic adaptation. Additionally, it was reported that three strains of *A. cellulolyticus* have different OGT (Mohagheghi et al. 1986), lending further support to this hypothesis. It is conceivable that other, yet unidentified, strains of *A. cellulolyticus* exist that span a range of either lower or higher OGT. 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 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 50-55°C, in LPBM medium (Mohagheghi et al. 1986; also called ATCC medium 1473), pH 5.5, modified such that the carbon source was 0.25 g/l cellobiose + 0.25 g/l glucose, without cellulose.

For genomic DNA isolation, 25 ml of bacterial culture was centrifuged at 10,000 rpm for 10 min to collect the cells. The pellet was rinsed in 1X TE buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA), flash frozen in liquid nitrogen and stored in -80°C. 200 µl of 1X TE (50 mM Tris-HCl, pH 8.0, 10 mM EDTA, preheated to 37°C) was added to the cell pellets, followed by10 µl of lysozyme (100 mg/ml, MP Biomedicals). The mixture was 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 once with phenol-chloroform, then chloroform. The upper phase was then transferred into a new tube, and 0.2 volume of NaOAc plus 2 volume of ice-cold ethanol were added. The tube was placed in -20°C for 5 min, 0°C for 5 min, then 4°C overnight. The tube was centrifuged at 10,000 x g for 5 min, the supernatant was carefully pipetted out, and the pellet was washed with 70% ethanol and recentrifuged. The supernatant was carefully decanted and pellet was air-dried. After the pellet was dry, it was resuspended in 50 µl of TE (10 mM Tris-HCl, pH 8.0, 1 mM EDTA) and RNaseI (final concentration at 0.1 mg/ml, Promega), stored in -20°C. The extracted genomic DNA was verified with 0.5% agarose gel containing ethidium bromide, at 10 mV overnight.

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 and Sequence Analyses. The protein sequence fasta files of completely sequenced bacterial genomes were obtained from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The executable BLAST (Altschul et al., 1997) programs as well as the nr database were obtained from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/BLAST/download.shtml). 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. The G+C content of the chromosome and all the genes as well as the codon usage in all predicted CDSs were calculated using short perl codes. The relative proportions of each nucleotide at each codon position were then calculated from the codon usage tables. Genomic signature was calculated as described by Karlin (2001).

In order to test whether the G+C is directly and solely responsible for the observed IVYWREL fractions in these organisms, we computed the expected codon proportions in each of the organisms based on the observed nucleotide composition in their coding DNA. The expected

fraction of the IVYWREL in each organism was calculated from the theoretical frequencies of the respective codons. Regression analysis was performed using the R package (http://www.r-project.org/) using the inbuilt lm function.

For the in silico modeling of *Frankia* DNA evolution experiment, the frequencies of codons in the coding DNA of each of the two *Frankia* DNA were altered such that the predicted proteome had an amino acid usage of *A. cellulolyticus*. However, the synonymous codon usage in the respective *Frankia* DNA was not altered. The percent G+C of the DNA before and after alteration was calculated.

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).

Principal Component Analysis (PCA). The amino acid compositions of 416 genomes were subjected to PCA using the R statistical software (http://www.r-project.org/). The organisms included *A. cellulolyticus*, 17 hyperthermophilic, 15 thermophilic, 4 pyschrophilic, and 254 mesophilic bacteria. The optimal growth temperature of 125 organisms was not available,. Of the 20 amino acids, the relative proportion for 7 amino acids (Ile, Val, Tyr, Trp, Arg, Glu, Leu) were combined into a single value, since the total fraction of these 7 amino acids in an organism's proteome has recently been shown to correlate directly with the optimal growth temperature of the organism (Zeldovich et al., 2007). Values for the other 13 amino acids were left unmerged. Thus, our dataset consisted of [416 organisms x 14 values for amino acid frequencies]. Short perl code was also written and employed to calculate the amino acid compositions of all the organisms, including *A. cellulolyticus*, using the predicted proteomes. For the sub-proteome analysis, proteomes were crudely fractionated into the three sub-proteomes, namely cytosolic, membrane and secretome, based on the number transmembrane segments in a protein. The transmembrane segments in proteins were predicted using HMMTOP (Tusnady and Simon, 1998 and

2001). Proteins with 0 TMSs were designated as cytosolic proteins, proteins with 1 TMS were designated as secreted proteins, and proteins with 2 or more TMSs were designated as membrane proteins.

ACKNOWLEDGMENTS

This work was supported by a Microbial Sequencing Project, U.S. Department of Energy, proposed by AMB. We would like to thank Dr. Charlie Strauss and Dr. Chris Stubben at the Los Alamos National Laboratory for help with the concepts of principal component analysis and with the usage of the R statistical package, respectively.

REFERENCES

1. Adney, W.S., Thomas, S.R., Baker, J.O., Himmel, M. E., and Chou, Y-C. 1998. Method for increasing thermostability in cellulase enzymes. *U.S. Patent No.* **5,712,142.**
2. Adney, W.S., Thomas, S.R., Nieves, R.A., and Himmel, M.E. 1994. Thermostable purified endoglucanase II from *Acidothermus cellulolyticus. U.S. Patent No.* **5,366,884.**
3. Adney, W.S., Tucker, M.P., Nieves, R.A., Thomas, S.R., and Himmel, M. E. 1995. Low molecular weight thermostable b-D-glucosidase from *Acidothermus cellulolyticus. Biotechnology Letters.* **17:** 49-54.
4. 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.
5. 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.

6. 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.

7. Benson, D.R. 1988. The genus *Frankia*: actinomycete symbionts of plants. *Microbiol Sci.* **5:** 9-12.

8. 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.

9. Biely, P. 2003. Diversity of microbial endo-b-1,4-xylanases In *Applications of enzymes to lignocellulosics* (eds. S.D. Mansfield and J.N. Saddler), pp. 361–380. American Chemical Society, Washington.

10. Biely, P., Vrsanska, M., Tenkanen, M., and D. Kluepfel. 1997. Endo-beta-1,4-xylanase families: differences in catalytic properties. *J. Biotechnol.* **57:** 151-166.

11. 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.

12. 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.

13. Claeyssens, M., and Henrissat, B. 1992. Specificity mapping of cellulolytic enzymes: classification into families of structurally related proteins confirmed by biochemical analysis. *Protein Sci.* **1:** 1293-1297.

14. Clarkson, K.A., Morgan, A.J., and Wang, Z.C. 1999. Xylanase from *Acidothermus cellulolyticus*. *U.S. Patent No.* **5,902,581**.

15. Collins, T., Gerday, C., and Feller, G.. 2005. Xylanases, xylanase families and extremophilic xylanases. *FEMS Microbiol. Rev.* **29:** 3-23.

16. 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.

17. Ding, S-Y., Adney, W.S., Vinzant, T.B., and Himmel, M.E. 2006. Thermal tolerant mannanase from *Acidothermus cellulolyticus*. *U.S. Patent No.* **7,112,429** B2.

18. Dufresne, A., Garczarek, L., and Partensky, F.. 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol.* **6:** R14.

19. Farias, S.T., and Bonato, M.C. 2003. Preferred amino acids and thermostability. *Genet Mol Res.* **2:** 383-393.

20. Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783-791.

21. Gilkes, N.R., Claeyssens, M., Aebersold, R., Henrissat, B., Meinke, A., Morrison, H.D., Kilburn, D.G., Warren, R.A., and Miller Jr., R.C. 1991. Structural and functional relationships in two families of beta-1,4-glycanases. *Eur. J. Biochem.* **202:** 367-377.

22. Haney, P.J., Badger, J.H., Buldak, G.L., Reich, C.I., Woese, C.R., and Olsen, G.J. 1999. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. *Proc. Natl. Acad. Sci. U.S.A.* **96:** 3578-3583.

23. Henrissat, B. 1991. A classification of glycosyl hydrolases based on amino-acid sequence similarities. *Biochem. J.* **280:** 309-316.

24. Himmel, M.E., Adney, W.S., Grohmann, K., and Tucker, M.P. 1994. Thermostable Purified Endoglucanase from *Acidothermus cellulolyticus*. *U.S. Patent No.* **5,275,944**.

25. Himmel, M.E., Tucker, M.P., Adney, W.S., and Nieves, R.A. 1995. Low molecular weight thermostable b-D-glucosidase from *Acidothermus cellulolyticus*. *U.S. Patent No.* **5,432,075**.

26. Hoskisson, P.A., and Smith, M.C. 2007. Hypervariation and phase variation in the bacteriophage 'resistome'. *Curr Opin Microbiol.* **10:** 396-400.

27. 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.

28. Karlin, S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol.* **9:** 335-343.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide seqeuences. *J. Mol. Evol.* **16**: 111-130.
Knaust F., Kube M., Reinhardt R., and Rabus R. 2007. Analyses of the *vrl* Gene Cluster in *Desulfococcus multivorans* : Homologous to the Virulence-Associated Locus of the Ovine Footrot Pathogen *Dichelobacter nodosus* Strain A198. *J Mol Microbiol Biotechnol* **13**:156–164.

29.

30. 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.

31. Lao, P.J., and Forsdyke, D.R. 2000. Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Res.* **10:** 228-236.

32. Lieph, R., Veloso, F.A., and Holmes, D.S. 2006. Thermophiles like hot T. *Trends Microbiol.* **14:** 423-426.

33. Lobry, J.R., and Chessel, D. 2003. Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *J Appl Genet.* **44:** 235-261.

34. Lobry, J.R., and Necsulea, A. 2006. Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene.* **385:** 128-136.

35. 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.

36. 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.

37. 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.

38. 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(Database issue):** D344-348.

39. 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.

40. Meyer, O., and Schlegel, H.G. 1983. Biology of aerobic carbon monoxide-oxidizing bacteria. *Annu Rev Microbiol.* **37:** 277-310.

41. 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.

42. Necsulea, A., and Lobry, J. R. 2006. Revisiting the directional mutation pressure theory: The analysis of a particular genomic structure in *Leishmania major*. *Gene* **385:** 28-40.

43. 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.

44. 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.

45. Ochman, H., and Davalos, L.M. 2006. The nature and dynamics of bacterial genomes. *Science.* **311:** 1730-1173.

46. 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.

47. 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.

48. Pascal, G., Médigue, C., and Danchin, A. 2005. Universal biases in protein composition of model prokaryotes. *Proteins: Structure, Function, and Bioinformatics.* **60:** 27-35.

49. Perriere, G., and Gouy, M. 1996. WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie* **78:** 364-369.

50. 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.

51. 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.

52. 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.

53. Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406-425.

54. 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.

55. Sandberg, R., Branden, C.I., Ernberg, I., and Coster, J. 2003. Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. *Gene.* **311:** 35-42.

56. Sharp, P.M., and Li, W-H. 1987. The codon adaptation index – a measure of directional synonymous codon usage bias and its potential applications. *Nucleic Acids Research* **15:** 1281-1295.

57. 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.

58. Suhre, K., and Claverie, J.M. 2003. Genomic Correlates of Hyperthermostability, an Update. *J. Biol. Chem.* **278:** 17198-17202.

59. 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.

60. 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.

61. Tekaia, F., Yeramian, E., and Dujon, B. 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene.* **297:** 51-60.

62. 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.

63. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876-4882.

64. Tucker, M.P., Grohmann, K., Mohagheghi, A., and Himmel, M.E. 1992. Thermostable purified endoglucanase from thermophilic bacterium *Acidothermus cellulolyticus. U.S. Patent No.* **5,110,735**.

65. 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.

66. Tusnady, G.E., and Simon, I. 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol*. **283:** 489-506.

67. Tusnady, G.E., and Simon, I. 2001. The HMMTOP transmembrane topology prediction server. *Bioinformatics*. **17:** 849-850.

68. Vallenet, D., Labarre L., Rouy Z., Barbe V., Bocs S., Cruveiller S., Lajus A., Pascal G., Scarpelli C., and Médigue C. 2006. MaGe - a microbial genome annotation system supported by synteny results. *Nucleic Acids Research* **34:** 53-65.

69. van Tilbeurgh, H., Pettersson, G., Bhikabhai, R., De Boeck, H., and Claeyssens. 1985. Studies of the cellulolytic system of *Trichoderma reesei* QM 9414. Reaction specificity and thermodynamics of interactions of small substrates and ligands with the 1,4-beta-glucan cellobiohydrolase II. *Eur. J. Biochem*. **148:** 329-334.

70. Vernikos, G.S., and Parkhill, J. 2006. Interpolated variable order motifs for the identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. *Bioinformatics* **22:** 2196-2203.

71. 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.

72. 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 (bars), the three genomic islands (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 genomic islands, GI 1, GI 2, and GI 3, respectively. The arrow indicates the location of the flagellar and motility genes.

Figure 3. Principle component analyses (PCA) of amino acid composition. (A) PCA of amino acid usage in 416 prokaryotic organisms as well as the three genomic islands of *A. cellulolyticus*. (B) PCA indicating the amino acid usage in the various "sub-proteomes" of *A. cellulolyticus*. Red: hyperthermophiles, orange: thermophiles, magenta: *A. cellulolyticus*, maroon: *T. fusca*, yellow: two *Frankia* sp. (ACN14a and CcI3), blue: two *Streptomyces* sp. (*S. avermitilis*, *S. coelicolor*), and cyan: other mesophilic actinobacteria.

Mesophiles, psychrophiles and bacteria with unknown optimal growth temperature are all in green. GI 1, GI 2, and GI 3 denote the three genomic islands. C, M, and S indicate the *A. cellulolyticus* cytosolic, membrane, and secretome sub-proteomes, respectively. W denotes the whole proteome of *A. cellulolyticus*.

SUPPLEMENTARY MATERIAL

MATERIALS AND 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.

Correspondence Analysis. The dataset for figures S5 and S6 consisted of 472 complete genome sequences, extracted from the NCBI complete genome database, including sequence data from *Acidothermus cellulolyticus* 11B and two *Frankia* species. Optimum growth temperature information was extracted from the DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH; German Collection of Microorganisms and Cell Cultures). We included in the dataset 18 hyperthermophilic species (OGT greater than or equal to 80°C), 20 thermophilic species (OGT between 55°C and 80°C) and 412 mesophilic prokaryotes (OGT between 20°C and 55°C. We used correspondence analysis to determine patterns of global codon usage and amino acid usage, as well as internal correspondence analysis for the pattern of synonymous codon usage, as described previously (Lobry and Chessel, 2003, Lobry and Necsulea, 2006).

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Distance matrix in substitutions/site of *Acidothermus cellulolyticus* 16S rRNA gene and other Frankinae, Actinobacteria and Firmicutes. Accession numbers: *Frankia alni* CT573213; *Frankia* sp. CcI3, CP000249; *Frankia* sp. EAN1pec, CP000820; *Geodermatophilus obscurus*, L40620; *Sporichthya polymorpha*, AB025317; *Cryptosporangium arvum*, D85465.1; *Modestobacter multiseptatus*, Y18646.1; *Blastococcus aggregatus*, AJ430193.1; *Mycobacterium leprae*, X55022.1; *Streptomyces coelicolor*, AB184800.1; *Bifidobacterium longum*, EF589112.1; *Thermobifida fusca*, AB210960.1; *Leifsonia xyli*, DQ232616.2; *Tropheryma whipplei*, 32447382; *Bacillus subtilis*, Z99104. Sequences were aligned using Clustal X (Thompson et al., 1997) using the No-gap option and Kimura's (1980) correction for multiple substitutions. Then a phylogenetic tree was generated by the Neighbor-Joining method (Saitou and Nei, 1987). Numbers on branch nodes are bootstrap values above 50%. The bar indicates 0.02 nucleotide substitution per site. NJplot software (Perriere and Gouy, 1996) was used to generate a graphic representation of the resulting tree. Bootstrap estimates (Felsenstein, 1985) were obtained from 1000

replicates. Shaded (gray) column are the distances between *A. cellulolyticus* and its neighbors, with at the top *Frankia alni* with 4.2%.

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

Figure S3. 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. A, K, L, and N 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 S4. (A) First factorial map for the correspondence analysis on global codon usage. (B) First factorial map for the correspondence analysis on synonymous codon usage. The coordinate on the first factor (horizontal) is positively correlated with the genomic G+C content. Red: hyperthermophiles, orange: thermophiles, green mesophiles, dark blue: psychrophiles, grey: species with unknown optimal growth temperature, cyan: *A. cellulolyticus*, yellow: *Frankia*, magenta: *Thermobifida fusca.*

Figure S5. PCA of *Frankia* CcI3 sub-proteomes. Red: hyperthermophiles, orange: thermophiles, magenta: *A. cellulolyticus*, maroon: *T. fusca*, yellow: two *Frankia* sp. (ACN14a and CcI3), blue: two *Streptomyces* sp. (*S. avermitilis*, *S. coelicolor*), and cyan: other mesophilic actinobacteria. Mesophiles, psychrophiles and bacteria with unknown optimal growth temperature are all in green. C, M, and S

indicate the *Frankia* CcI3 cytosolic, membrane, and secretome sub-proteomes, respectively. W denotes the whole proteome of *Frankia* CcI3.

Figure S6. PCA of *S. coelicolor* sub-proteomes. Red: hyperthermophiles, orange: thermophiles, magenta: *A. cellulolyticus*, maroon: *T. fusca*, yellow: two *Frankia* sp. (ACN14a and CcI3), and blue: two *Streptomyces* sp. (*S. avermitilis*, *S. coelicolor*). Mesophiles, psychrophiles and bacteria with unknown optimal growth temperature are all in grey. C, M, and S indicate the *S. coelicolor* cytosolic, membrane, and secretome sub-proteomes, respectively. W denotes the whole proteome of *S. coelicolor*.

Figure S7. Influence of G+C on the amino acid composition in the six actinobacteria. (A) Correlation between G+C content and total fraction of IVYWREL amino acids, (B) Correlation between G+C content and optimal growth temperature, (C) Correlation between the genomic G+C content and the expected(theoretical) fraction of IVYWREL, and (D) Correlation between the G+C content in the coding DNA and the expected(theoretical) fraction of IVYWREL.

Figure S8. First factorial map for the correspondence analysis on amino acid usage. The coordinate on the first factor (horizontal) is positively correlated with the genomic G+C content. Red: hyperthermophiles, orange: thermophiles, green mesophiles, dark blue: psychrophiles, grey: species with unknown optimal growth temperature, cyan: *A. cellulolyticus* 11B, yellow: *Frankia*, magenta: *Thermobifida fusca*.

SUPPLEMENTARY TABLES

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

Table S2. Comparative analysis of codon usage in six actinobacteria.

Table S3. *A. cellulolyticus* 11B proteins that have best BLAST-hits to Archaea or Eukarya.

Table S4. Genomic regions identified in the genome of *A. cellulolyticus* 11B.

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

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

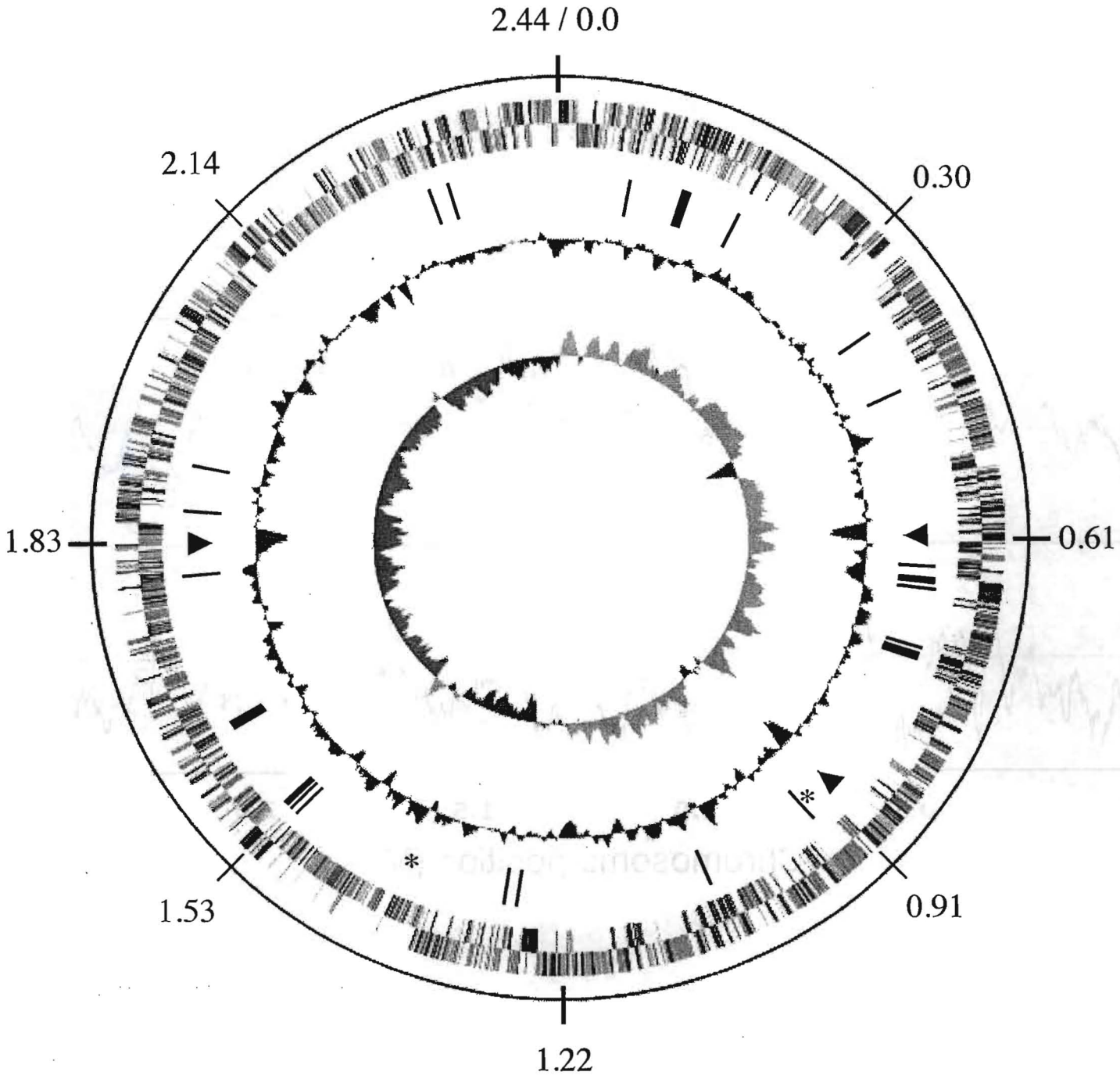


Fig. 1

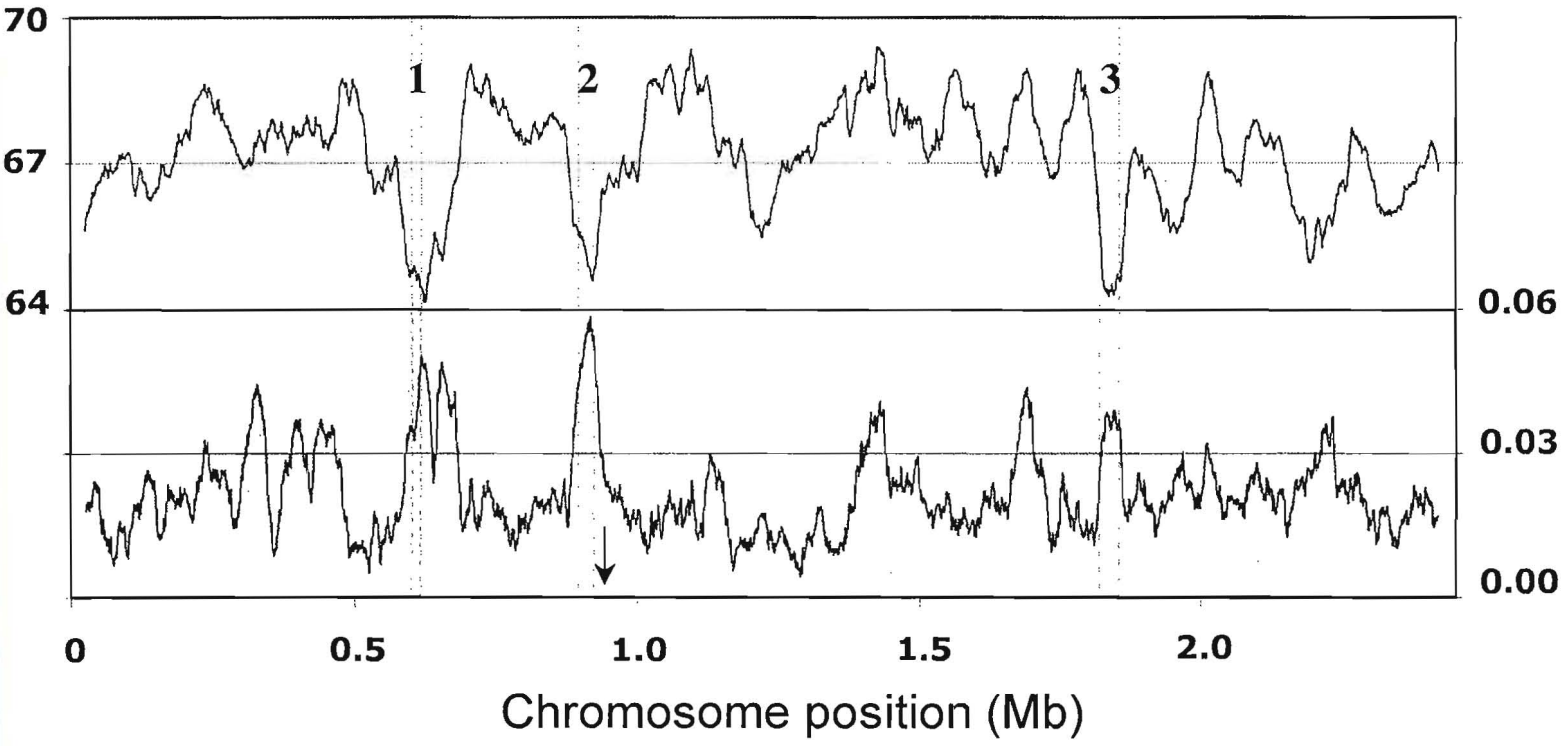
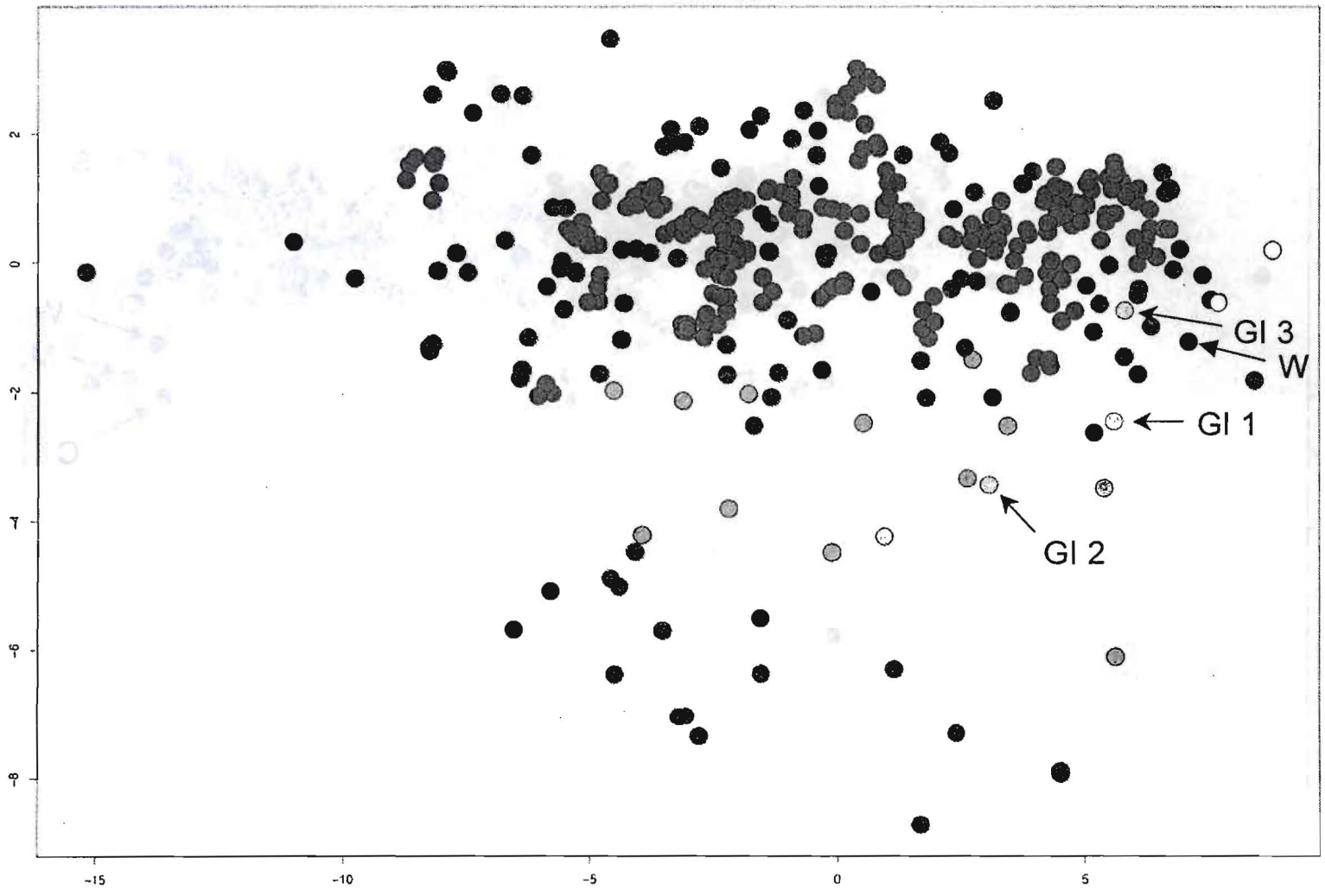
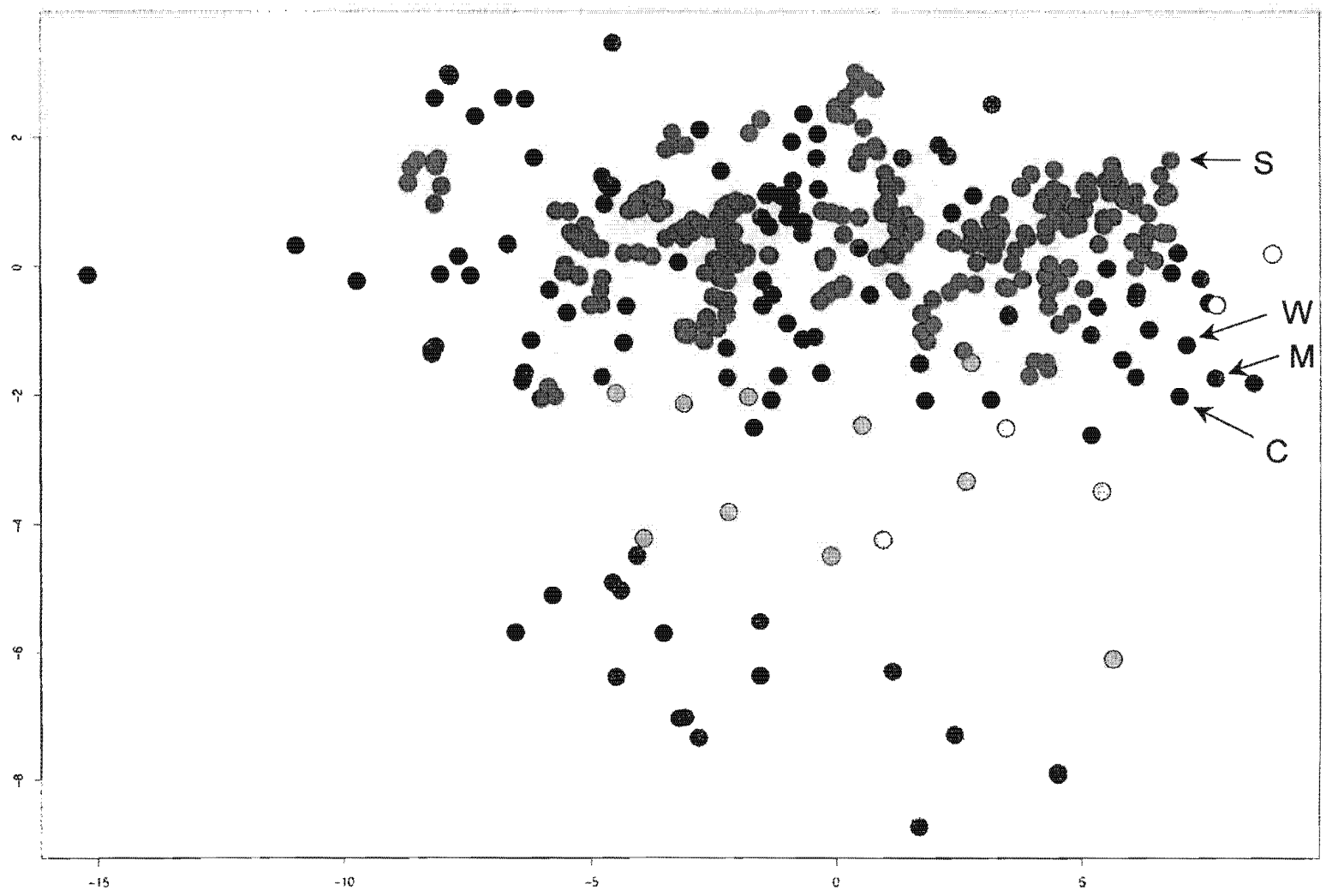


Fig.



Fi



Fi

. Genomic overview of the six actinobacterial species used for comparative analysis in this study.

	Acidothermus cellulolyticus 11B	*Frankia* sp. CcI3	*Frankia alni* ACN14a	*Streptomyces avermitilis*	*Streptomyces coelicolor*	*Thermobif fusca*
size (Mb)	2.4	5.4	7.5	9.0	8.7	3.6
f genome	66.9	70.1	72.8	70.7	72.1	67.5
ı DNA	89	84	86	86	88	85
eins	2157	4499	6711	7577	7769	3110
A operons	1	2	2	6	5	4
As	46	46	46	68	64	52
ıdogenes	4	50	12	0	56	7
sposases/IS elements	2	145	33	110	55	5
ge/viral proteins	0	6	24	20	8	3
growth temperature	55	25	25	27	27	57

syl hydrolases encoded in the *A. cellulolyticus* 11B genome.

Location	Strand	Size	GH family	EC#	Domains	Description/annotation
76150..77826	-	558	20	3.2.1.52	GH20b-GH20	Beta-N-acetylhexosaminidase
29121..130548	+	475	3		PRK05337	Glycoside hydrolase family 3 domain protein
30609..132018	-	469	16		GH16-CBDIV	Carbohydrate binding family 6
35878..137314	+	478	1	3.2.1.21	GH1	Beta-glucosidase, Glycosyl Hydrolase family
39082..140491	+	469	6	3.2.1.4	GH6	Cellulase
90901..192937	+	678	10		GH3-CBM3-CBM2	Glycoside hydrolase, family 10
34509..385678	-	389	10	3.2.1.8	GH10	Endo-1,4-beta-xylanase
49974..452355	+	793	18		GH18	Glycoside hydrolase, family 18
39808..641226	-	472	18		GH18	Glycoside hydrolase, family 18
53210..654898	+	562	5	3.2.1.4	Cellulase-CBM2	Endo-1,4-glucanase E1 (Cel5A), glycoside h family 5
55010..658639	+	1209	6 & 12	3.2.1.4	GH6-CBM3-GH12-CBM2	Glycoside hydrolase family 6, endoglucanase
58797..661088	+	763	5		Cellulase-CBM3-CBM2	Cellulose-binding family II, mannanase (Mar
61169..664534	+	1121	48		CBM3-GH48-CBM2	Glycoside hydrolase family 48
64806..668702	+	1298	74		VPS10-CBM3-CBM2	Cellulose-binding family II
69117..670328	+	403	12		GH12-CBM2	Cellulose-binding family II
28911..731094	-	727	13		PulN-GBN-AA-AAC	1,4-Alpha-glucan branching enzyme
32511..734181	-	556	13		AA	Trehalose synthase
34195..736156	-	653	13		AA	Alpha amylase, catalytic region
39202..741304	-	700	13		GdBN-AA / PulA	Glycogen debranching enzyme GlgX
39393..940370	+	325	23		NLPC_P60-LT_GEW	Lytic transglycosylase, catalytic
65840..1068524	-	894	9		GH9-CBM2	Glycoside hydrolase family 9
71105..1272988	-	627	15		GH15	Glycoside hydrolase 15-related
83876..1285078	-	400	23		LT_GEWL	Lytic transglycosylase, catalytic
17759..1518754	+	331	32		GH32N	Glycosyl hydrolase family 32, N-terminal dor protein

27513..1529651	+	712	13		GdBN-AA / PulA	Glycogen debranching enzyme GlgX
29641..1531983	+	780	13		AA / TreY	Malto-oligosyltrehalose synthase
31980..1533713	+	577	13		MTHN-AA / GlgB	Malto-oligosyltrehalose trehalohydrolase
29666..1631048	-	460	18	3.2.1.14	GH18-CBM4_9	Chitinase, Glycosyl Hydrolase family 18
31033..1631818	-	261	18		CBM4_9	Carbohydrate-binding, CenC domain proteir protein
32120..1634411	-	763	18	3.2.1.14	GH18-CBM4_9	Chitinase, Glycosyl Hydrolase family 18
00240..1802540	-	766	77	2.4.1.25	MalQ	4-Alpha-glucanotransferase
66241..1868934	+	897	3	3.2.1.21	GH3-GH3C-PA14-GH3C	Beta-glucosidase
12657..1916070	-	1137	9		GH9-CBM3-CBM3-CBM2	Glycoside hydrolase, family 9
05541..2307316	+	591	18	3.2.1.14	GH18-CBM4_9	Chitinase, Glycosyl Hydrolase family 18
22145..2324598	+	817	3		GH3-GH3C	Glycoside hydrolase family 3 domain proteir

s encoded on the three genomic islands found in the *A. cellulolyticus* 11B genome.

	Start	Stop	Strand	Size	%GC	Annotation	Broad Function
1.							
	599123	600460	+	1338	58.7	fumarate reductase/succinate dehydrogenase flavoprotein	Respiration
	600457	601455	+	999	53.5	aryldialkylphosphatase	Detoxification or organophos
	601452	602315	+	864	57.6	short-chain dehydrogenase/reductase SDR	Metabolism
	602318	603025	+	708	59.7	deoxyribose-phosphate aldolase	Pentose phosphate pathway/Nuclueotide metabo
	603025	604050	+	1026	62.8	ROK family protein	Repressor/Kinase/ORF
	603976	604737	-	762	59.3	transcriptional regulator, GntR family	Regulation
	604831	606093	-	1263	61.9	ROK family protein	Repressor/Kinase/ORF
	606431	607279	+	849	58.2	Uncharacterized protein containing SIS (Sugar ISomerase) phosphosugar binding domain-like	Carbohydrate metabolism
	607358	608434	+	1077	58.9	periplasmic binding protein/LacI transcriptional regulator	ABC transport
	608511	609977	+	1467	59.0	ABC transporter related	ABC transport
	610025	611029	+	1005	58.6	inner-membrane translocator	ABC transport
	611066	612055	+	990	56.6	inner-membrane translocator	ABC transport
	612100	613272	+	1173	58.9	oxidoreductase domain protein	
	613214	614131	+	918	53.8	Xylose isomerase domain protein TIM barrel	Sugar interconversion
	614278	615468	-	1191	59.1	oxidoreductase domain protein	Metabolism
			+				
2.			+				
	895201	895275	+	75	58.7	Xaa tRNA	
	895710	895892	+	183	59.0	DNA binding domain, excisionase family	VrlI homolog of *Dichelobacte nodosus*
	895934	896410	+	477	59.3	conserved hypothetical protein	VrlJ homolog of *Dichelobacte nodosus*
	896448	900179	+	3732	62.6	conserved hypothetical protein	VrlK homolog of *Dichelobacte nodosus*
	900182	901585	+	1404	61.0	putative transcriptional regulator	Transcriptional regulation
	901764	904742	+	2979	60.5	conserved hypothetical protein	
	904746	905549	+	804	51.1	hypothetical protein	

905551	908352	+	2802	64.7	helicase domain protein	VrlO homolog of *Dichelobacte nodosus*?
908345	910378	+	2034	57.5	conserved hypothetical protein	VrlP homolog of Dichelobacte nodosus
910375	911157	+	783	57.0	conserved hypothetical protein	VrlQ homolog of *Dichelobacte nodosus*
911530	911721	+	192	67.2	hypothetical protein	
912180	913517	+	1338	68.8	metallophosphoesterase	DNA exonuclease
913514	916267	+	2754	67.6	SMC domain protein	ATPase involved in DNA repa
916437	917942	+	1506	66.7	acyltransferase 3	Metabolic enzyme
917849	919492	-	1644	66.5	diguanylate cyclase/phosphodiesterase	Metabolic enzyme
919731	920096	-	366	65.0	hypothetical protein	
920160	920777	-	618	66.2	protein of unknown function DUF421	
921140	921213	+	74	66.2	Met tRNA	Protein synthesis
		+				
3.		+				
1824619	1824691	+	73	68.5	Arg tRNA	Protein synthesis
1825076	1825351	+	276	51.1	hypothetical protein	
1825356	1825841	+	486	62.8	hypothetical protein	
1826168	1826434	+	267	64.0	transcriptional regulator, XRE family	Transcriptional regulation
1826461	1826988	+	528	55.7	hypothetical protein	
1827069	1827608	+	540	66.5	hypothetical protein	
1827605	1828294	+	690	63.3	ABC transporter related	Transport
1828319	1829596	+	1278	65.8	protein of unknown function DUF214	
1829664	1830167	+	504	63.7	methylglyoxal synthase	Enzyme
1830331	1831779	-	1449	64.9	methyl-accepting chemotaxis sensory transducer	Chemotaxis
1832014	1832652	-	639	65.9	conserved hypothetical protein	
1832649	1833722	-	1074	65.5	protein of unknown function DUF182	
1833937	1834560	-	624	54.8	conserved hypothetical protein	
1834844	1836649	-	1806	58.9	purine catabolism PurC domain protein	Nucleotide metabolism
1836652	1837632	-	981	59.6	conserved hypothetical protein	
1837635	1838843	-	1209	61.7	Pyridoxal-5'-phosphate-dependent enzyme, beta subunit	Metabolic enzyme

1838855	1839568	-	714	62.2	carbon monoxide dehydrogenase subunit G	CO fixation
1839571	1840056	-	486	59.7	(2Fe-2S)-binding domain protein	
1840053	1840940	-	888	61.4	Carbon-monoxide dehydrogenase (acceptor)	CO fixation
1840960	1841652	-	693	59.2	Asp/Glu racemase	Amino acid metabolism
1841870	1843549	-	1680	58.9	polar amino acid ABC transporter, inner membrane subunit	Amino acid transport
1843660	1844568	-	909	57.2	extracellular solute-binding protein, family 3	Solute uptake
1845097	1847445	-	2349	61.0	aldehyde oxidase and xanthine dehydrogenase	Metabolic enzyme
1847442	1848227	-	786	60.4	coenzyme A transferase	Metabolic enzyme
1848224	1849177	-	954	59.6	Glutaconate CoA-transferase	Metabolic enzyme
1849177	1850124	-	948	55.6	luciferase family protein	Metabolic enzyme
1850459	1851148	+	690	62.8	NADPH-dependent F420 reductase	Metabolic enzyme
1852077	1853591	+	1515	67.7	Malate dehydrogenase (oxaloacetate-decarboxylating)	Metabolic enzyme
1853620	1854708	+	1089	66.1	molybdenum cofactor biosynthesis protein A	
1854843	1855652	+	810	69.5	Exonuclease, RNase T and DNA polymerase III	Metabolic enzyme
1855719	1855794	+	76	59.2	His tRNA	Protein synthesis

elative proportions of each nucleotide at each of the three codon poistions.

	Optimal growth temperature	Nucleotide and Codon base position										
		Position 1				Position 2				Position 3		
		A	C	G	T	A	C	G	T	A	C	G
cus 11B	55	0.362	0.280	0.425	0.235	0.457	0.291	0.213	0.533	0.181	0.429	0.362
14	25	0.388	0.267	0.413	0.255	0.514	0.278	0.213	0.621	0.098	0.455	0.374
	25	0.382	0.277	0.408	0.247	0.487	0.282	0.216	0.580	0.131	0.441	0.376
s	27	0.384	0.261	0.412	0.274	0.518	0.269	0.206	0.617	0.098	0.469	0.382
	27	0.381	0.258	0.417	0.275	0.534	0.264	0.208	0.644	0.086	0.478	0.375
	57	0.357	0.272	0.424	0.256	0.481	0.265	0.212	0.591	0.163	0.463	0.364
quared value		0.961	0.297	0.864	0.271	0.594	0.030	0.018	0.445	0.783	0.129	0.818
lue less than		0.001	0.27	0.007	0.3	0.08	0.74	0.8	0.15	0.02	0.49	0.014

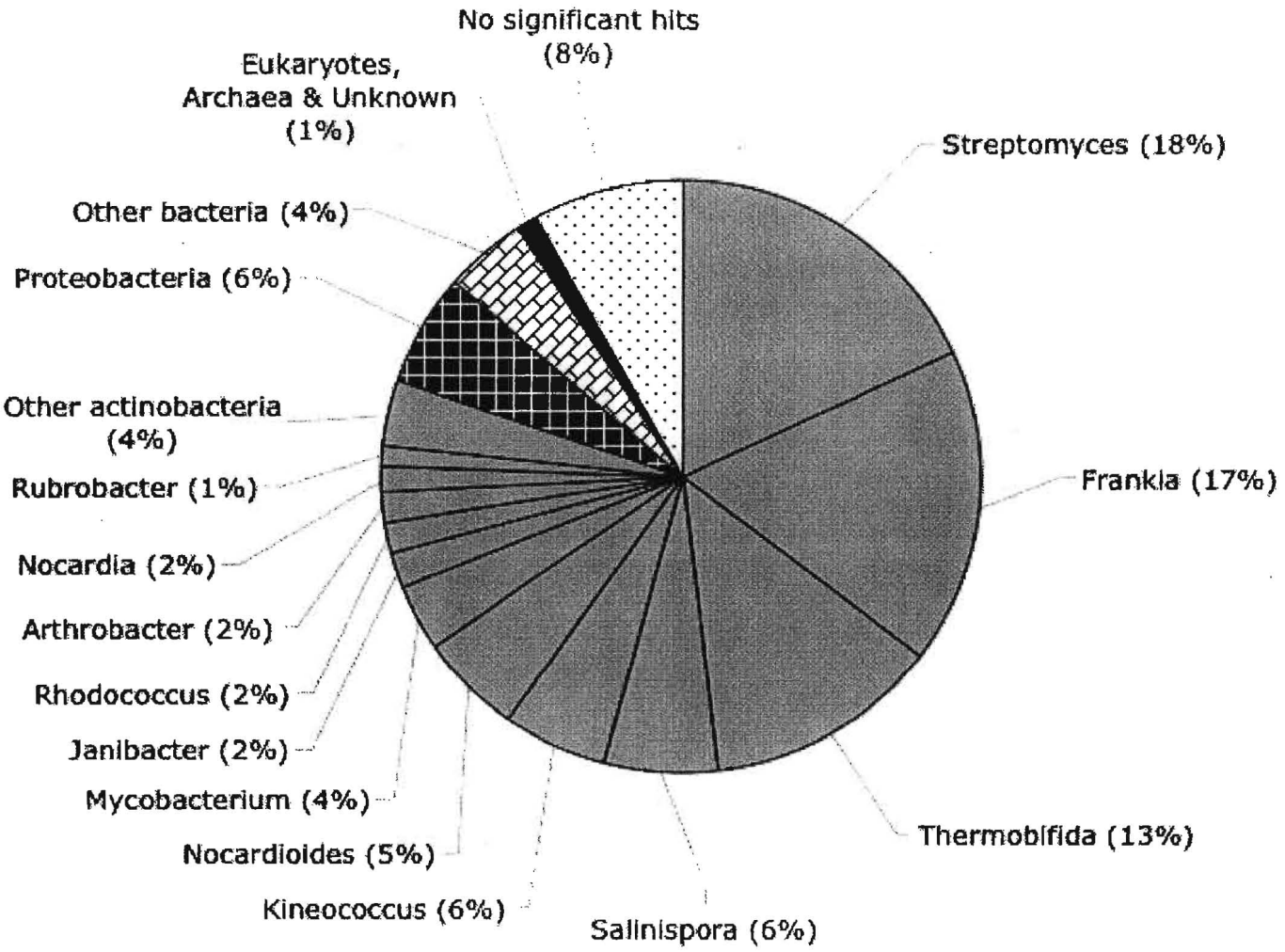
Percent G+C of *in silico* altered coding DNA from two *Frankia* genomes.

Source DNA	Organism	
	Frankia CcI3	*Frankia* ACN14a
Original	70.54	73.02
In silico altered	69.47	71.10

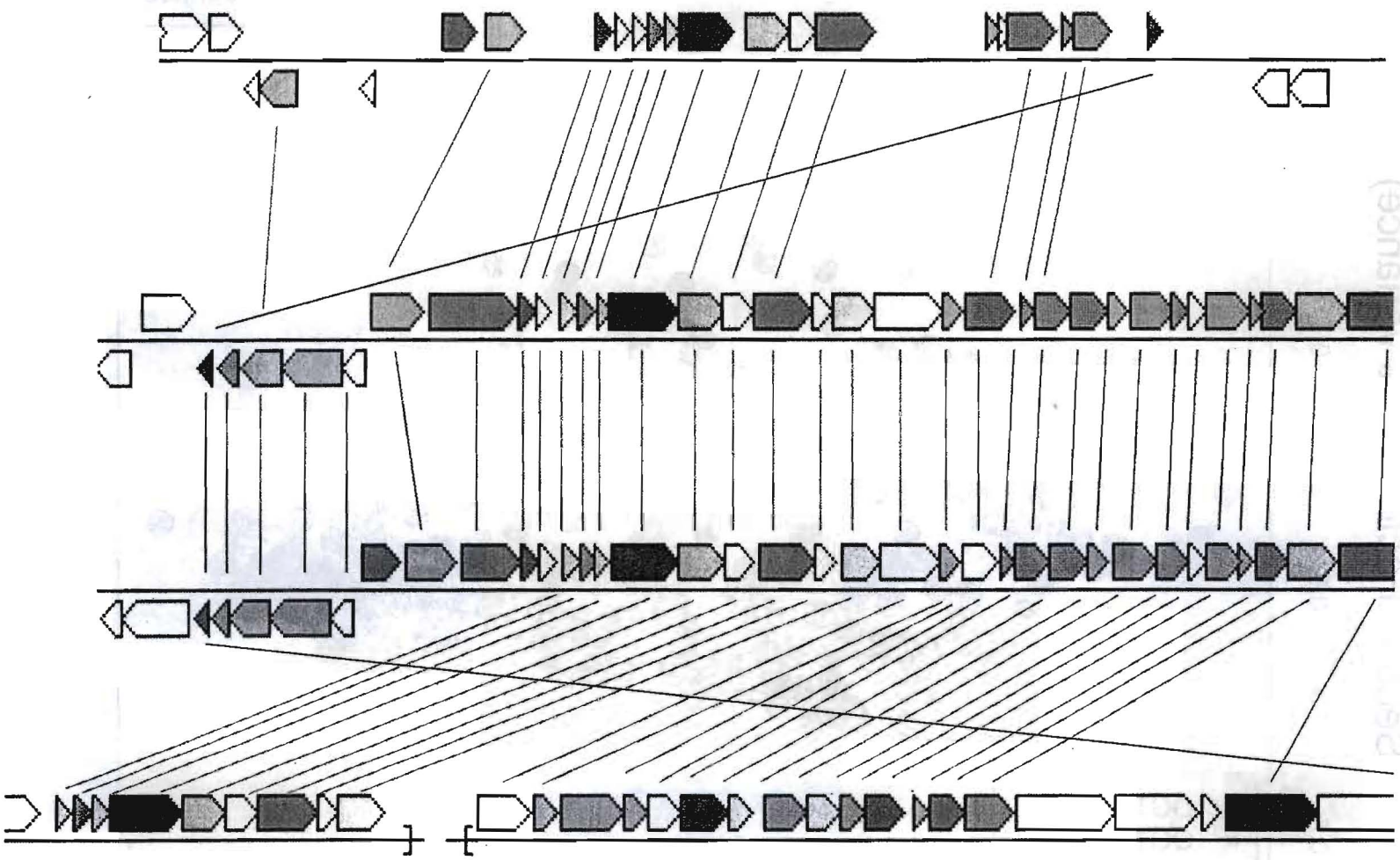
Acidothermus cellulolyticus	0																
Frankia alni	0.050	0															
Frankia sp CcI3	0.053	0.009	0														
Frankia sp. EaN1pe	0.055	0.020	0.021	0													
Geodermatophilus obscurus	0.073	0.071	0.070	0.079	0												
K. radiotolerans	0.093	0.086	0.089	0.092	0.078	0											
Sporichthya polymorpha	0.068	0.064	0.067	0.065	0.065	0.082	0										
Modestobacter multiseptatus	0.080	0.067	0.073	0.073	0.043	0.094	0.069	0									
Cryptosporangium arvum	0.085	0.071	0.072	0.072	0.074	0.084	0.065	0.092	0								
Thermobifida fusca	0.116	0.111	0.104	0.111	0.101	0.111	0.110	0.120	0.106	0							
Mycobacterium leprae	0.138	0.131	0.132	0.130	0.140	0.139	0.144	0.142	0.139	0.147	0						
Streptomyces coelicolor	0.107	0.102	0.103	0.100	0.113	0.115	0.092	0.110	0.106	0.127	0.144	0					
Tropheryma whipplei	0.122	0.112	0.115	0.116	0.123	0.120	0.101	0.127	0.116	0.158	0.171	0.129	0				
Leifsonia xyli	0.124	0.108	0.112	0.110	0.109	0.101	0.100	0.104	0.114	0.144	0.151	0.124	0.091	0			
Nocardioides JS614ctg204	0.112	0.106	0.101	0.106	0.108	0.090	0.112	0.111	0.102	0.121	0.138	0.115	0.124	0.113	0		
Bifidobacterium longum	0.168	0.176	0.176	0.174	0.179	0.172	0.178	0.189	0.181	0.204	0.211	0.201	0.174	0.184	0.189	0	
Bacillus subtilis	0.208	0.206	0.205	0.202	0.217	0.214	0.201	0.210	0.201	0.234	0.260	0.221	0.220	0.227	0.238	0.261	0

Fig. S1

R.D. Baral


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%)
Thermobifida (13%)
Nocardioides (5%)
Kineococcus (6%)
Salinispora (6%)

Fi



F

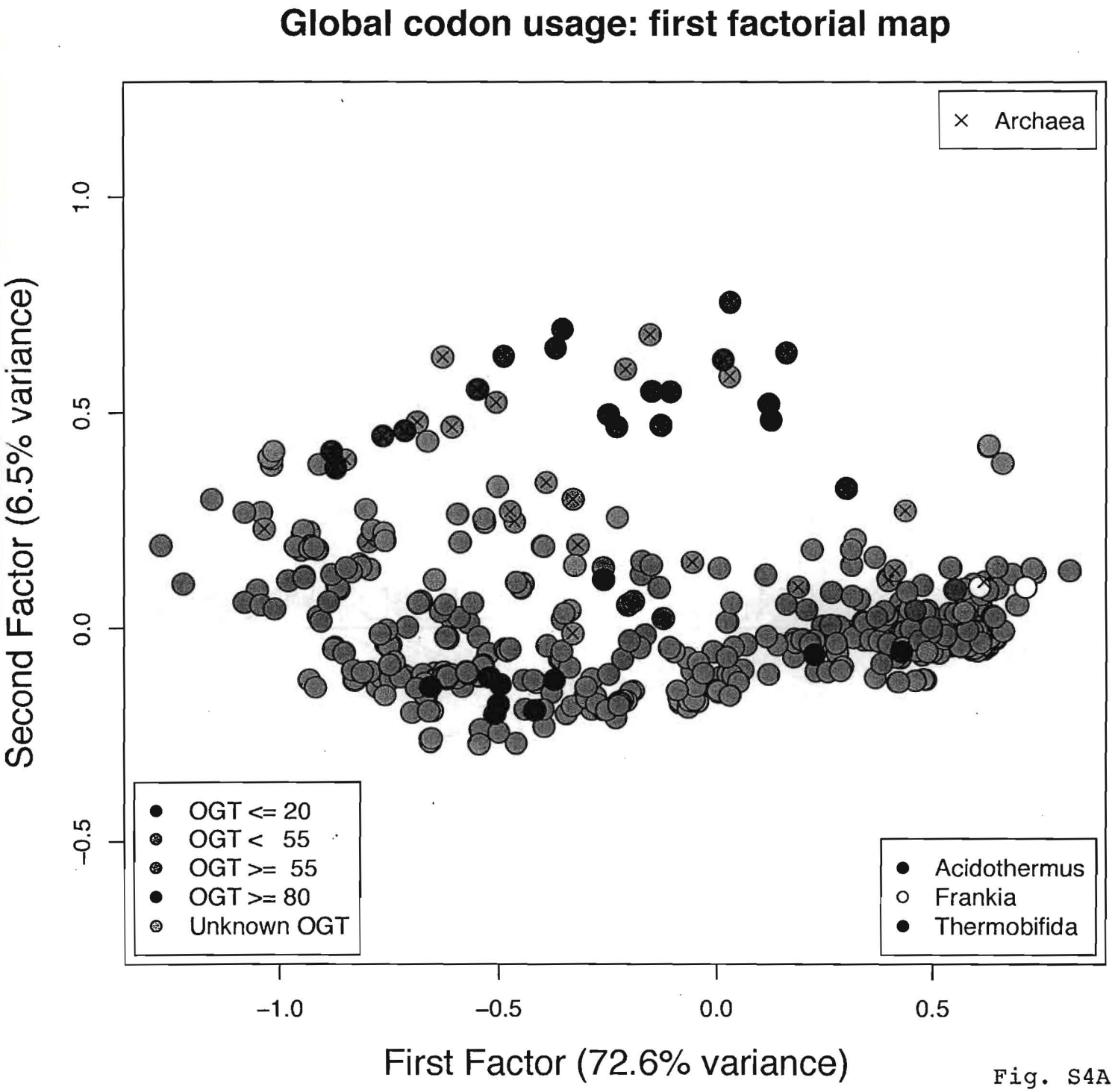

Global codon usage: first factorial map
Archaea
OGT <= 20
OGT < 55
OGT >= 55
OGT >= 80
Unknown OGT
Acidothermus
Frankia
Thermobifida
Second Factor (6.5% variance)
First Factor (72.6% variance)
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5

Fig. S4A

Synonymous codon usage: first factorial map

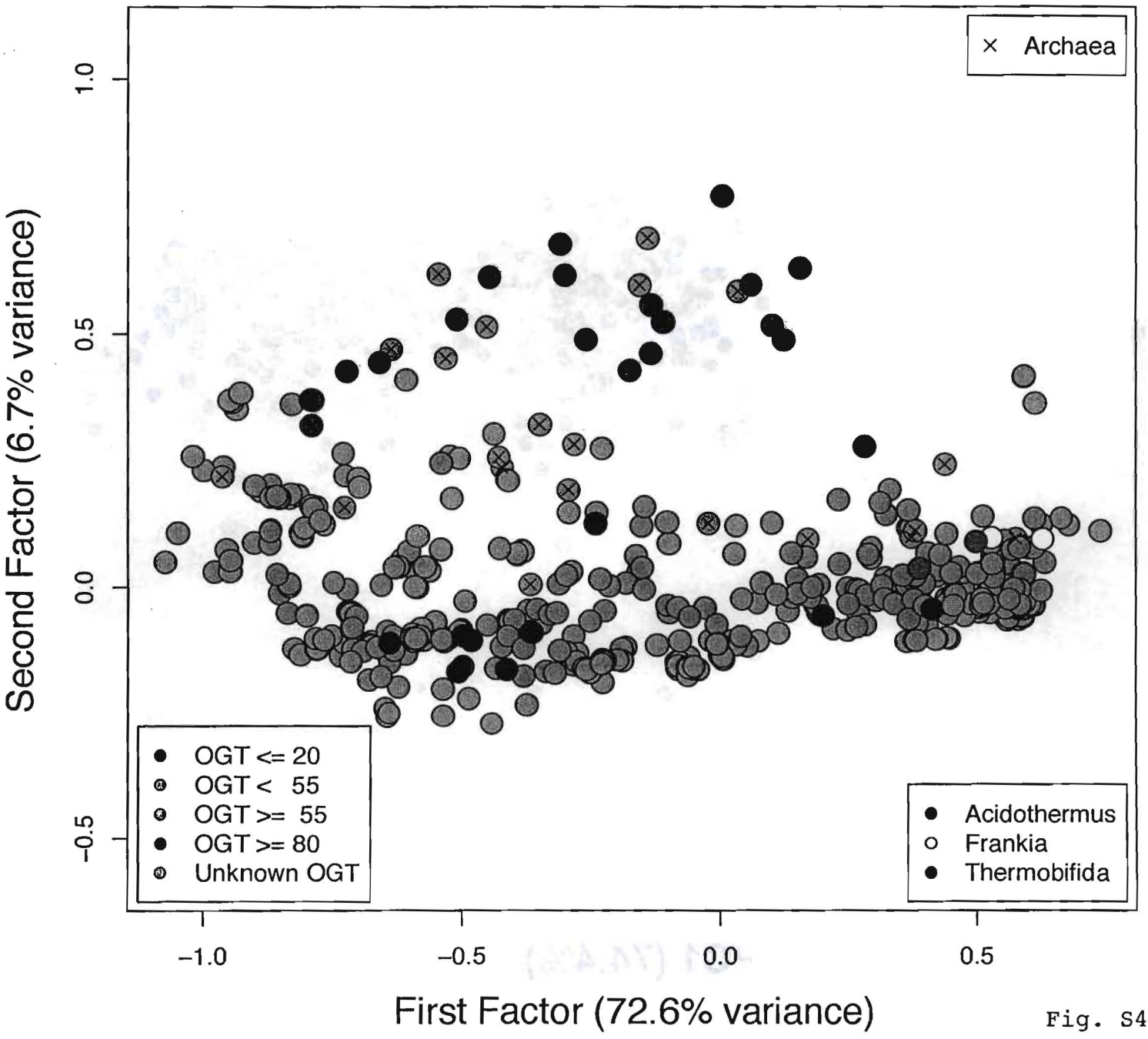
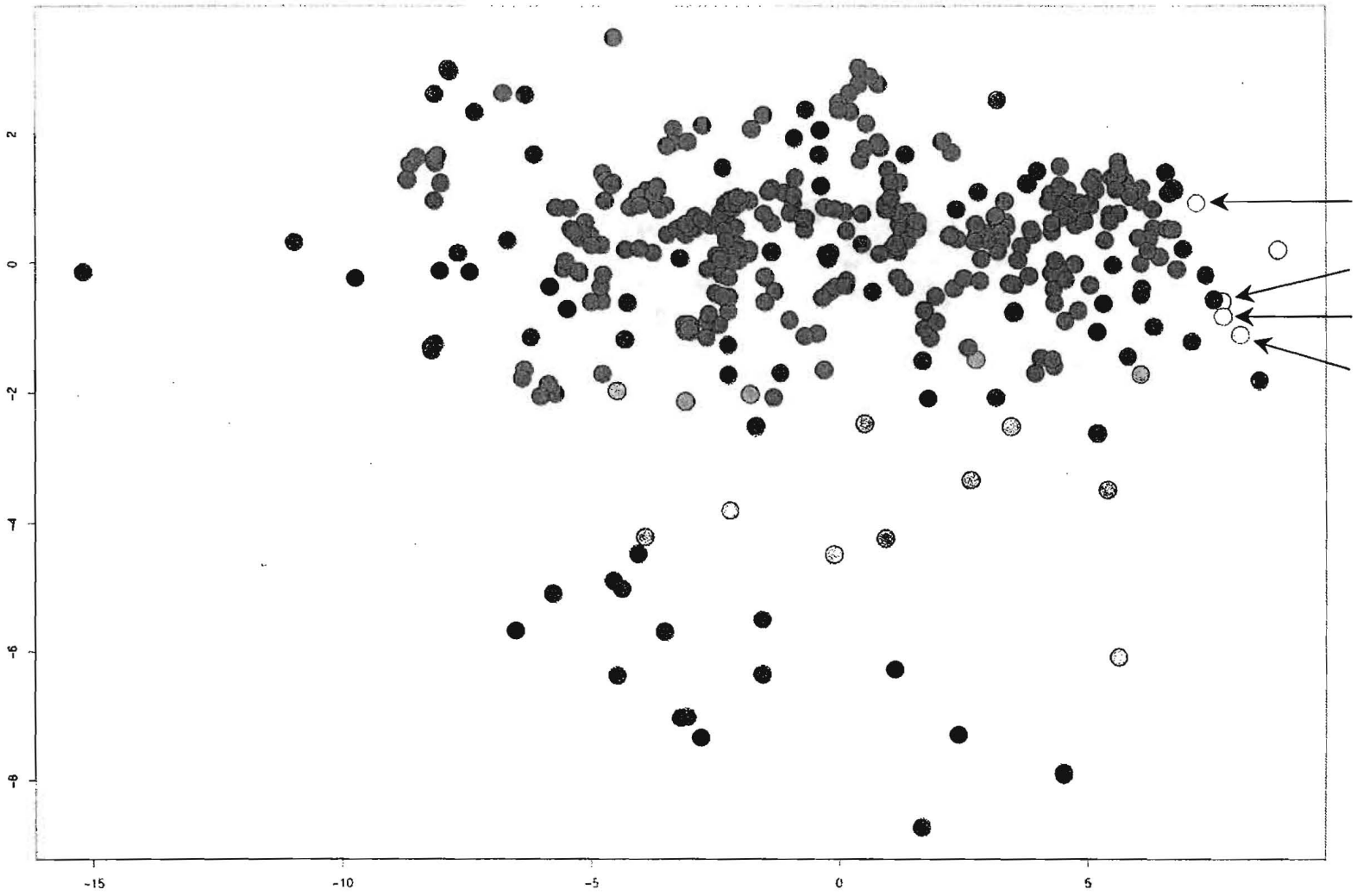
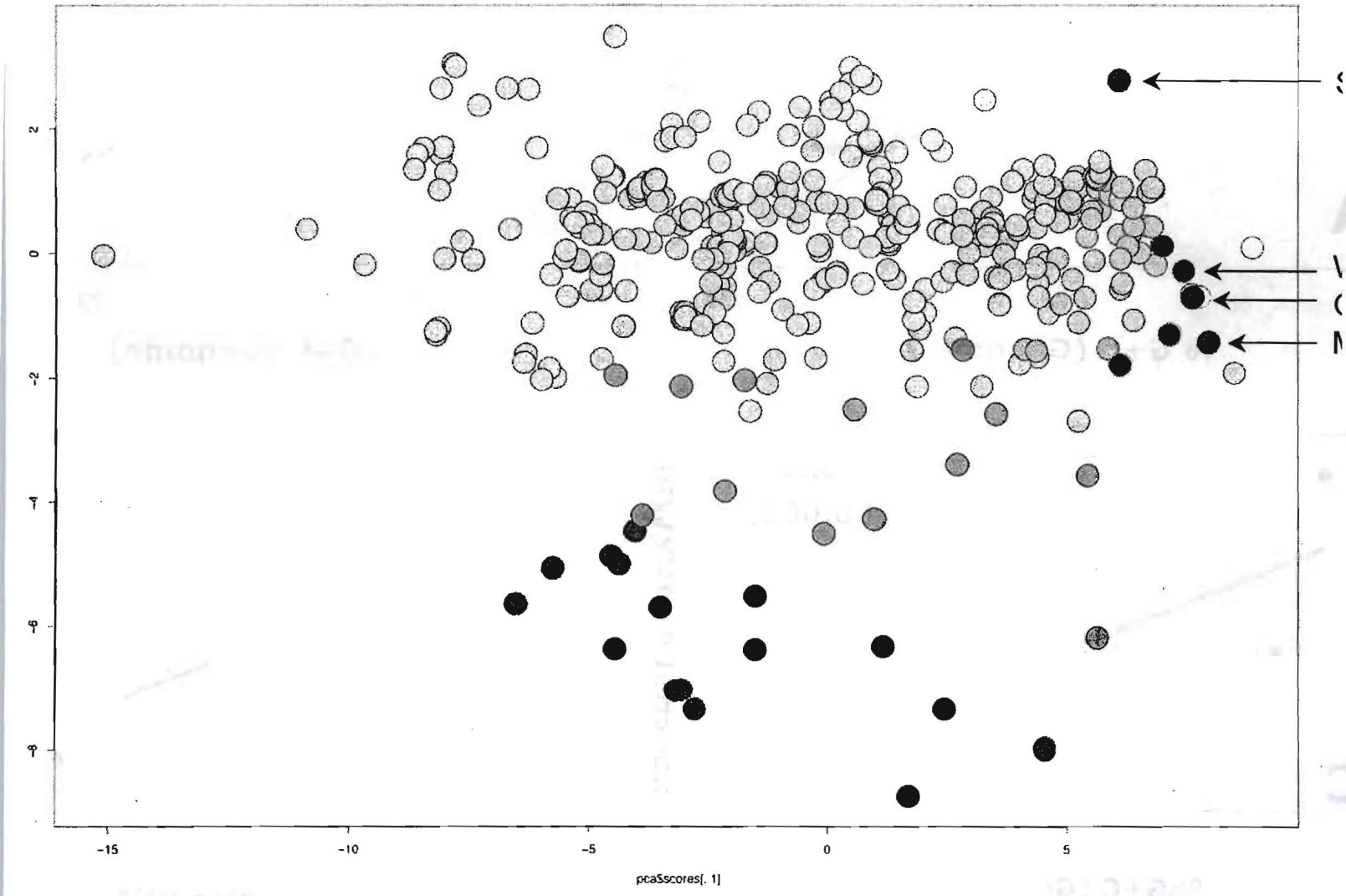
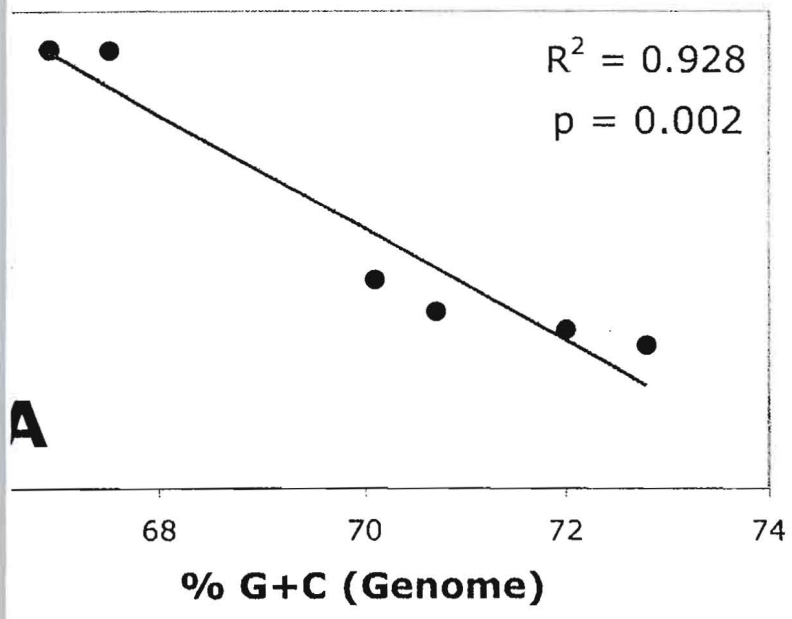


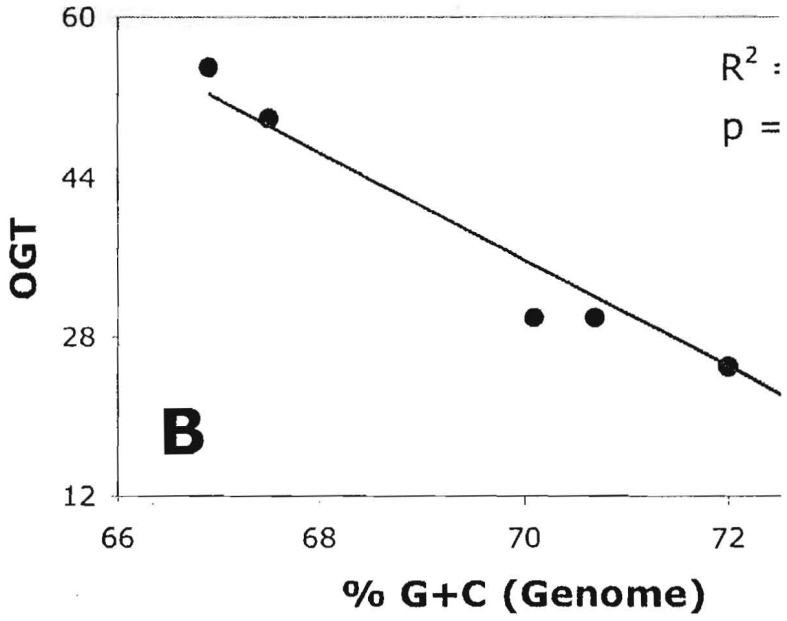
Fig. S4B

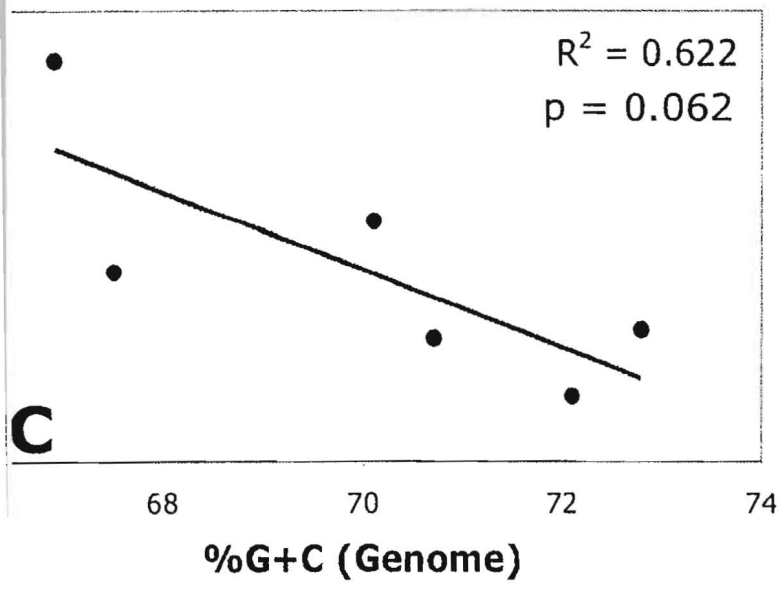

2
0
-2
-4
-6
-8
-15
-10
-5
0
5
PC1 (74.4%)

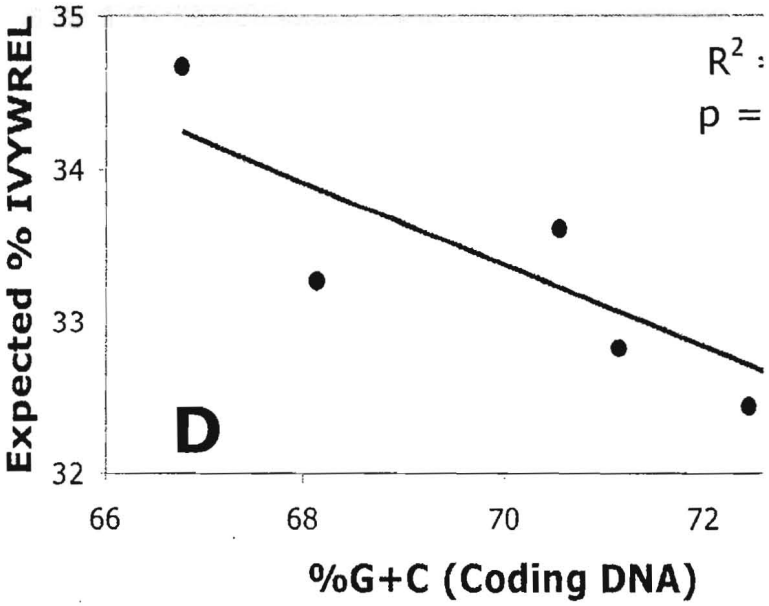


PC1 (74.4%)


R^2 = 0.928
p = 0.002
A
68
70
72
74
% G+C (Genome)


60
44
28
12
OGT
R^2 =
p =
B
66
68
70
72
% G+C (Genome)


R^2 = 0.622
p = 0.062
C
68
70
72
74
%G+C (Genome)


35
34
33
32
Expected % IVYWREL
R^2 =
p =
D
66
68
70
72
%G+C (Coding DNA)

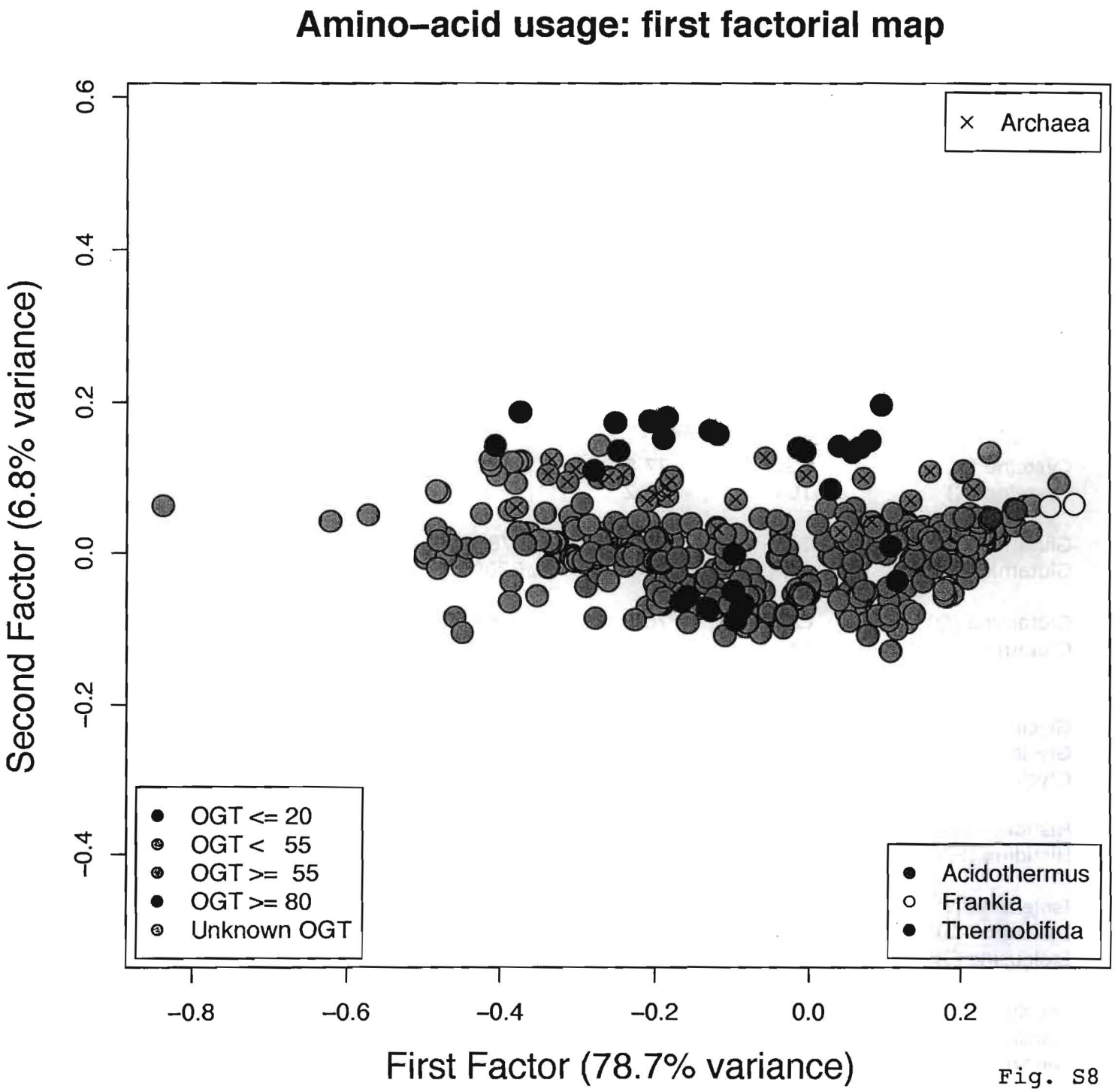

Amino–acid usage: first factorial map
Second Factor (6.8% variance)
First Factor (78.7% variance)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.8
−0.6
−0.4
−0.2
0.0
0.2
× Archaea
OGT <= 20
OGT < 55
OGT >= 55
OGT >= 80
Unknown OGT
Acidothermus
Frankia
Thermobifida

Fig. S8

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

Amino acid	Codon	Synonymous Codon Usage	tRNA Location	Anticodon
Alanine (A)	GCG	40.9	179235..179159	CGC
Alanine (A)	GCC	40.8	842997..843072	GGC
Alanine (A)	GCA	10.1	10372..10447	TGC
Alanine (A)	GCT	8.2	-	-
Arginine (R)	CGG	43.4	677195..677124	CCG
Arginine (R)	CGC	33.4	-	-
Arginine (R)	CGT	11.1	41078..41153	ACG
Arginine (R)	CGA	8.4	-	-
Arginine (R)	AGG	2.3	160793..160721	CCT
Arginine (R)	AGA	1.3	1824619..1824691	TCT
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)	GGC	47.5	1524762..1524834	GCC
Glycine (G)	GGG	19.3	2389324..2389254	CCC
Glycine (G)	GGT	18.8		
Glycine (G)	GGA	14.4	1797034..1797104	TCC
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)	CTC	39.5	1294313..1294229	GAG
Leucine (L)	CTG	34.9	24698..24783	CAG
Leucine (L)	TTG	12.5	1195870..1195798	CAA
Leucine (L)	CTT	10.6		
Leucine (L)	CTA	1.4	1894564..1894636	TAG
Leucine (L)	TTA	1.1	2147532..2147460	TAA
Lysine (K)	AAG	70.4	1861376..1861451	CTT
Lysine (K)	AAA	29.6	2336318..2336246	TTT

Methionine (M)	ATG	100	308433..308509	CAT
Methionine (M)	ATG	100	509883..509959	CAT
Methionine (M)	ATG	100	921140..921213	CAT
Phenylalanine (F)	TTC	79.3	2339803..2339879	GAA
Phenylalanine (F)	TTT	20.7	-	-
Proline (P)	CCG	63.7	2281657..2281581	CGG
Proline (P)	CCC	23.0	1372838..1372762	GGG
Proline (P)	CCA	7.1	1796952..1796877	TGG
Proline (P)	CCT	6.3	-	-
Serine (S)	TCG	29.9	2293483..2293394	CGA
Serine (S)	TCC	26.1	2292585..2292671	GGA
Serine (S)	AGC	25.8	40835..40924	GCT
Serine (S)	AGT	7.8	-	-
Serine (S)	TCA	6.5	2374948..2375032	TGA
Serine (S)	TCT	4.0	-	-
Threonine (T)	ACC	50.3	308349..308424	GGT
Threonine (T)	ACG	37.6	2230630..2230555	CGT
Threonine (T)	ACA	7.1	89585..89660	TGT
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)	GTC	50.9	1525000..1525074	GAC
Valine (V)	GTG	34.9	1523720..1523649	CAC
Valine (V)	GTT	10.4	-	-
Valine (V)	GTA	3.8	974815..974886	TAC
Stop codon	TGA	67.1	-	-
Stop codon	TAG	21.3	-	-
Stop codon	TAA	11.6	-	-
Unknown	?	-	895201..895275	?

Table S2. Comparative analysis of 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

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

Size	Protein description	GI of Best hit	Best hit organism	Blast2Seq score
to Archaea				
245	hypothetical protein Acel_0034	110667166	Haloquadratum walsbyi DSM 16790	224
111	protein of unknown function DUF59	14590230	Pyrococcus horikoshii OT3	264
137	hypothetical protein Acel_0525	15897985	Sulfolobus solfataricus P2	291
204	hypothetical protein Acel_0526	88604270	Methanospirillum hungatei JF-1	111
366	hypothetical protein Acel_0621	110667166	Haloquadratum walsbyi DSM 16790	225
381	UDP-N-acetylglucosamine 2-epimerase	15678857	Methanothermobacter thermautotrophicus str. Delta H	519
283	fructose-bisphosphate aldolase	15922681	Sulfolobus tokodaii str. 7	320
284	ABC transporter related	73668563	Methanosarcina barkeri str. Fusaro	588
245	ABC transporter related	110669070	Haloquadratum walsbyi DSM 16790	620
291	ATP phosphoribosyltransferase (homohexameric)	116753404	Methanosaeta thermophila PT	780
219	hypothetical protein Acel_1310	110667166	Haloquadratum walsbyi DSM 16790	163
230	Asp/Glu racemase	14521305	Pyrococcus abyssi GE5	408
317	Glutaconate CoA-transferase	11498798	Archaeoglobus fulgidus DSM 4304	495
233	Small-conductance mechanosensitive channel-like	48477585	Picrophilus torridus DSM 9790	200
282	ABC-2 type transporter	15899372	Sulfolobus solfataricus P2	221
347	ABC transporter related	119720042	Thermofilum pendens Hrk 5	406
164	Vitamin K epoxide reductase	70606913	Sulfolobus acidocaldarius DSM 639	193
303	hypothetical protein Acel_2067	15898666	Sulfolobus solfataricus P2	355
to Eukarya				
898	hypothetical protein Acel_0064	118129698	Gallus gallus	294
656	esterase, PHB depolymerase family	114324587	Volvariella volvacea	753
439	hypothetical protein Acel_0740	97180301	Contains: Proline-rich peptide SP-A (PRP-	264
160	hypothetical protein Acel_0770	118085709	Gallus gallus	112
206	GPR1/FUN34/yaaH family protein	119178442	Coccidioides immitis RS	354
225	cell wall surface anchor family protein	109658562	Homo sapiens	180
323	hypothetical protein Acel_1712	46119356	Gibberella zeae PH-1	760
219	beta-lactamase domain protein	125820913	Danio rerio	438

Table S4. Salient features of additional genomic regions (GR) identified in the genome of *Acidothermus cellulolyticus* 11B.

GR#	Description and Features
GR0	Mainly hypothetical proteins – specific to *Acidothermus* compared to the 7 selected genomes (see methods).
GR1	Enzymatic activities (ATPase + Kelch repeat possibly in a galactose oxidase).
GR2	Several (conserved) hypothetical protein + enzymatic activities + transporter and 1 regulator.
GR3	Mainly specific (conserved) hypothetical protein in the first part and cellulose transport + metabolism (degradation) shared with *Streptomyces* species, *Frankia* EAN1 and *T. fusca*.
GR4a	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
GR4b	First part, probably nitrate metabolism with transporter and nitrate reductase activity (shared with *S. coelicolor* only). Second part, specific (conserved) hypothetical proteins.
GR6	Cluster of protein/enzymes involved in cellulose degradation (specific to *Acidothermus* although partial homologs exists in the compared species).
GR7	Enzymatic activities (glycosyltransferase, carbamoylphosphate, epimerase, hydrolase)
GR8	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: PorC = Pyruvate oxidoreductase gamma chain (ACICE0782) PorD = Pyruvate oxidoreductase delta chain (very partial match on ACICE0785 but more than 51% identity in aa). PorA = Pyruvate oxidoreductase alpha chain (ACICE0783)

	PorB = Pyruvate oxidoreductase beta chain (ACICE0784)
GR9	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	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	First part, shared with Frankia CcI3 only, enzymatic activities (cytochrome c3 hydrogenase), second part more specific to *Acidothermus* with (conserved) hypothetical proteins.
GR13a	Transport system (type ABC, substrate amino acid ?) shared with *Frankia* species only.
GR13b	Metabolism, very probably degradation (monooxygenase, dioxygenase,...) of glutamine/glutamate ? + regulator (marR family) + transport ? (permease). Highly specific to *Acidothermus*.
GR14	Mainly hypothetical proteins with a putative rRNA methylase and exonuclease. Mainly specific to *Acidothermus*.
GR15	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), ...
GR16	Transport (ABC type, sugar ?) + regulator (lacI family) + enzymatic activities

	(levabase, oxidase)
GR17a	Mainly specific conserved hypothetical proteins
GR17b	Cluster of enzymatic functions involved in aromatic compound degradation (paa genes cluster for phenylacetatic acid degradation) + regulation (*tetR* family). Shared with *Streptomyces* species only.
GR17c	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).
GR19a	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.
GR19b	Mainly (conserved) hypothetical proteins + transport system + regulator (*marR* family) + enzymatic activities (oxidoreductase).
GR20	Mainly specific conserved hypothetical proteins + chitinase (chitin degradation)
GR21a	ONLY specific conserved hypothetical proteins + regulator (marR family) and probable transporter.
GR21b	ONLY (conserved) hypothetical proteins + regulator (fragment) and probable transporter.

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

Organism	% IVYWREL	Optimal growth temperature	%G+C
Acidothermus cellulolyticus 11B	41.76	55	66.9
Frankia sp. ACN14a	39.90	25	72.8
Frankia sp. CcI3	40.32	25	70.1
Streptomyces avermitilis	40.12	27	70.7
Streptomyces coelicolor	40.00	27	72.1
Thermobifida fusca	41.75	57	67.5
	R-squared value	0.966	0.926
	p-value is less than	0.0005	0.0021

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

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

Organism	Genome Size	% IVYWREL	Optimal growth temperature	%G+C
Arthrobacter aurescens TC1	5.23	40.2	30	62.4
Acidothermus cellulolyticus 11B	2.40	42.4	55	66.9
Arthrobacter sp. FB24	5.08	40.1	30	65.4
Bifidobacterium adolescentis ATCC 15703	2.10	39.0	37	59.2
Bifidobacterium longum	2.26	39.0	37	60.1
Corynebacterium diphtheriae	2.49	40.5	37	53.5
Corynebacterium efficiens YS-314	3.15	40.9	37	63.1
Corynebacterium glutamicum ATCC 13032 Bielefeld	3.30	40.7	33	53.8
Corynebacterium glutamicum ATCC 13032 Kitasato	3.30	40.7	33	53.8
Corynebacterium glutamicum R	3.35	40.7	33	54.1
Corynebacterium jeikeium K411	2.48	39.8	37	61.4
Clavibacter michiganensis NCPPB 382	3.40	40.3	28	72.5
Frankia alni ACN14a	7.50	41.4	26	72.8
Frankia sp. CcI3	5.40	41.7	26	70.1
Kineococcus radiotolerans SRS30216	4.81	40.4	32	74.2
Leifsonia xyli subsp. *xyli* CTCB0	2.58	40.4	29	67.7
Mycobacterium avium 104	5.50	40.7	39	69.0
Mycobacterium avium subsp. *paratuberculosis*	4.80	40.7	39	69.3
Mycobacterium bovis	4.35	40.7	37	65.6
Mycobacterium bovis BCG Pasteur 1173P2	4.40	40.7	37	65.6
Mycobacterium gilvum PYR-GCK	5.96	40.3	30	67.7
Mycobacterium leprae	3.27	41.0	37	57.8
Mycobacterium smegmatis MC2 155	7.00	40.6	37	67.4
Mycobacterium sp. JLS	6.00	40.4	30	68.4
Mycobacterium sp. KMS	6.22	40.5	30	68.2
Mycobacterium sp. MCS	5.92	40.5	30	68.4
Mycobacterium tuberculosis CDC1551	4.40	40.8	37	65.6
Mycobacterium tuberculosis F11	4.40	40.7	37	65.6
Mycobacterium tuberculosis H37Ra	4.40	40.7	37	65.6
Mycobacterium tuberculosis H37Rv	4.40	40.7	37	65.6
Mycobacterium ulcerans Agy99	5.60	40.8	32	65.5
Mycobacterium vanbaalenii PYR-1	6.50	40.4	30	67.8
Nocardia farcinica IFM10152	6.29	40.6	37	70.7
Nocardioides sp. JS614	5.31	40.5	30	71.4
Propionibacterium acnes KPA171202	2.56	40.0	37	60.0
Rhodococcus sp. RHA1	9.67	40.4	30	67.0
Rubrobacter xylanophilus DSM 9941	3.23	45.0	60	70.5
Streptomyces avermitilis	9.12	40.4	27	70.7
Streptomyces coelicolor	9.09	40.3	27	72.0
Saccharopolyspora erythraea NRRL 2338	8.20	40.9	28	71.1
Symbiobacterium thermophilum IAM14863	3.60	42.8	60	68.7
Salinispora tropica CNB-440	5.20	41.2	28	69.5
Thermobifida fusca YX	3.60	41.6	57	67.5
Tropheryma whipplei TW08 27	0.93	40.0	37	46.3
Tropheryma whipplei Twist	0.93	39.8	37	46.3
	R-sqaured value		0.4	0.1
	p-value is less than		7.9E-06	0.034