Material and Methods:

Organism Selection, Cell growth and DNA preparation

Phylogenetic trees of 16S rRNAs from the greengenes database were used to guide the selection of organisms to be sequenced[1]. The GEBA pilot project targeted a broad range of species of archaea and bacteria with very diverse requirements for growth conditions. Table S1C lists 36 different media (DSMZ and ATCC) used for the growth of the 56 organisms whose genomes are analyzed in this paper, with their growth temperature ranging from 28°C to 65°C (1/3 of them grown under anaerobic conditions). Conditions for *efficient* cell lysis of the target organisms vary almost as much as conditions for cell growth. We established a routine for initial small scale tests to explore the required individual procedures for cell lysis, which when successful (microscopic observation of lysis) were scaled up for the large amounts of cell pastes required for the extraction of about 100 µg of high quality genomic DNA (gDNA) (tested by pulse field electrophoresis). Table S1C lists eight cell lysis buffers and varied incubation conditions for lysis, as well as the three procedures used for DNA purification from the lysed cells.

Sequencing and Assembling

Eleven microbes were sequenced by Sanger only approach (Table S2). In this approach, 8kb libraries were constructed from DNA fragments randomly sheared by Hydroshear (GeneMachines) with ends repaired by T4 DNA polymerase and Klenow fragment (New England Biolabs). End repaired fragments were size selected from an agarose gel and purified by using the QIAquick Gel Extraction Kit (Qiagen). Approximately 200 ng of sheared DNA was ligated into 100 ng of linearized, dephosphyorylated pMCL200 vector with kanamycin resistance gene[2]. Ligation mixtures were electroporated into ElectroMAXDH10B cells (Invitrogen), and plated on agar plates containing kanamycin. White colonies were picked by using the robotic QPIX (Genetix) colony picker and grown in media containing glycerol to provide stocks for subsequent sequencing. Fosmid libraries were constructed by using the CopyControl™ Fosmid Production Kit (Epicentre). DNA was sheared and fragments were size-selected on an agarose pulsed-field gel, excised and purified before ligation in the pCCqFos vector. The ligated vector was packaged by using MaxPlax™ Lambda Packaging Extract (Epicentre) and used to transduce *E. coli* (EP300). Cloned genomic fragments were sequenced from both ends to depth of approximately 10x and shotgun reads were automatically assembled with three assemblers (phrap, PGA, Arachne)[3,4]. The best of these draft assemblies was used for annotation and as a starting point for genome completion.

The remaining forty-five genomes were sequenced using various combinations of pyrosequencing and traditional Sanger sequencing (Table S2). Thirty-two of them were drafted by combining in average 30x of standard 454 FLX data with 8x Sanger reads of 8kb and 40kb libraries and thirteen were completed by assembling together 25x of pyrosequence data and 8x of Sanger paired-end sequence from an 8kb-library .

Genome Annotation

Annotation was performed using the IMG system (http://img.jgi.doe.gov/cgi-bin/pub/main.cgi) developed by JGI[5]. Automated metabolic network reconstruction of all 56 genomes was performed using the Pathway Tools platform version 12.5[6]. After generating an initial draft set of pathway genome databases, we manually inspected the report files and collected frequently missed annotations in a local enzyme mapping file, in order to capture any IMG-specific enzyme terminology[7]. The final pathway reconstructions were generated in a manner similar to the "Tier 3" pathway genome databases in BioCyc[8].

"Genome tree' phylogenetic tree building for bacterial genomes

A maximum likelihood phylogenetic tree for bacterial genomes was built upon a concatenated alignment of 31 phylogenetic marker genes. We included 53 GEBA bacteria and 667 bacterial compete genomes from Genbank for the tree building. Protein sequences for 31 phylogenetic marker genes (*dnaG, frr, infC, nusA, pgk, pyrG, rplA, rplB, rplC, rplD, rplE, rplF, rplK, rplL, rplM, rplN, rplP, rplS, rplT, rpmA, rpoB, rpsB, rpsC, rpsE, rpsI, rpsJ, rpsK, rpsM, rpsS, smpB,* and *tsf*) were retrieved, aligned, trimmed, and concatenated using the software AMPHORA[9]. A maximum likelihood tree was then constructed from the concatenated-alignment using PHYML[10]. The model selected based on the likelihood ratio test was the WAG model of amino acid substitution with γ-distributed rate variation (five categories) and a proportion of invariable sites. The shape of the γ-distribution and the proportion of the invariable sites were estimated by the program.

Phylogenetic Diversity (PD) estimation

The maximum likelihood phylogenetic tree was used to estimate the phylogenetic contribution of the GEBA pilot project. PD (phylogenetic diversity) is a measure of the length of the branches in a phylogenetic tree. The PD contribution of GEBA project is the total length of the added branches by the addition of the 53 GEBA bacterial genomes. 100 random sampling of 53 genomes from the genome tree (include or exclude GEBA organisms) were selected to calculate the correspondent PD contributions to compare with the GEBA PD contribution. The same approach was adopted to compare the PD contribution of the 26 GEBA actinobacteria with the other 47 actinobacteria.

Rate of gene family discovery in different genome sets

Protein families were identified using the Markov Clustering Algorithm (MCL) [11]. An all versus all BLASTP search was performed for the entire proteomes of the 56 GEBA genomes. All matches with an E-value lower than 1e-10 with matches spanning at least 80% the lengths of both sequences were used as links for MCL clustering. Each cluster is treated as a protein family. Protein families were built separately for the following four groups of genomes: 53 bacterial genomes from the GEBA project, 73 actinobacterial genomes (including 26 GEBA genomes), 40 enterbacteriaceae genomes and eight

Streptococcus agalactiae genomes. For each group, we randomly selected a given number of genomes and calculated the number of protein families present in that set of genomes to estimate the novel gene family discovery rate.

Effect of GEBA genomes on genome analysis

For gene context and fusion analysis we used the corresponding information in the IMG 2.5 database using all the non-eukaryotic publicly available genomes (excluding GEBA genomes). We identified unique chromosomal neighborhoods and fusion events in the 56 GEBA genomes. Chromosomal neighborhood analyses were limited to the identification of new COG combinations found in chromosomal cassettes. A unique fusion event was considered a case where the fusion components are found fused for the first time. Genes that don't belong to any known protein families (COG, Pfam, TIGRfam) were used for protein family link identification. Conservation scores were calculated based upon all vs all BLASTP bit scores according to the equations described in the IMG manual [5]. None-GEBA proteins that exhibit conservation score greater than 0.25 were grouped using a single linkage algorithm. Subsequently we identified the genes in the GEBA genome that have conservation score greater than 0.25 to members of different clusters.

BARP (Bacterial Actin Related Protein) studies

For the ribbon plot in Figure 3, BARP from *H. ochraceum* (GI:227395998) was submitted to SWISS-MODEL (http://swissmodel.expasy.org//SWISS-MODEL.html) for 3D structural modeling. The Protein Data Bank entry 1C0F_A (the actin from *Dictyostelium discoideum*) [12] was the best structure match and was used as the reference to build the 3D structure for BARP by SWISS-MODEL. The secondary structure elements of both the *H. ochraceum* actin and 1C0F_A were defined by DSSP from their 3D models[13]. The two sequences were aligned to each other by MUSCLE[14], and the secondary structure elements were illustrated alongside the alignments by ALINE[15].

For BARP expression study, DNA and RNA from *H. ochraceum SMP-2* were harvested from cells actively growing on VY/4-SWS (DSMZ) agar plates. Cells were scraped and re-suspended in VY/4-SWS medium lacking yeast cell paste. Total RNA was extracted by hot phenol method[16]. For RT-PCR, cDNA was prepared from 10 ug of total RNA using Primer L6actin (5'- GAACCCGGCGAACTGGGCATC -3') as previously described[17] . Chromosomal DNA and cDNA were used to amplify a 1062-bp region using primers RT-D1 (5'-GATAATGAGTTCCGCGCCGTG-3') and RT-U1 (5'-GTATTCTCCAGCCCCATTATC-3'). PCR reactions were performed using the following time settings: 95°C denaturation for 1', 62°C annealing for 30", and a 2' extension at 70°C for 30 cycles. For the RNase control, RNA was digested with RNase for 1 hr prior to cDNA preparation.

For the phylogenetic study of BARP, homologs of BARP were retrieved from the NRAA database by BLASTP, representatives from actin and different actin related protein families were selected. Alignments were built by MUSCLE[14] for BARP , actin related

proteins and MreB. A phylogenetic tree was built by PHYML[10] using settings exactly like the concatenated bacterial genome tree building.

Phylogenetic coverage of current sequencing projects

A phylogenetic tree was built for a combined alignment of SSU rRNA sequences from the published complete bacterial and archaeal genomes, GEBA genomes and a non-redundant subset of greengenes SSU-rRNA[1]. The non-redundant subset of greengenes SSU rRNA was obtained by MCL clustering of SSU rRNA sequences with higher than 99% identity over 80% sequence length span[11]. Alignments of the non-redundant greengenes SSU-rRNAs along with the SSU rRNA from the sequencing projects were retrieved from the greengenes database. The mask defined by the greengenes project was applied to trim the alignments. A neighbor-joining tree was built from this alignment using FastTree[18]. A greedy algorithm was applied to list all the taxa included in the tree in the descending order according to their contributions to the phylogenetic diversity (PD)[19].

1 DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**, 5069-5072 (2006).
2 Nakano, Y., Yoshida, Y., Yamashita, Y. & Koga, T. Construction of a series of pACYC-derived plasmid vectors. *Gene* **162**, 157-158 (1995).
3 de la Bastide, M. & McCombie, W. R. Assembling genomic DNA sequences with PHRAP. *Curr Protoc Bioinformatics* **Chapter 11**, Unit11 14 (2007).
4 Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res* **12**, 177-189 (2002).
5 Markowitz, V. M. *et al.* The integrated microbial genomes (IMG) system. *Nucleic Acids Res* **34**, D344-348 (2006).
6 Karp, P. D., Paley, S. & Romero, P. The Pathway Tools software. *Bioinformatics* **18 Suppl 1**, S225-232 (2002).
7 Green, M. L. & Karp, P. D. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5**, 76, (2004).
8 Caspi, R. *et al.* The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **36**, D623-631(2008).
9 Wu, M. & Eisen, J. A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9**, R151 (2008).
10 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704 (2003).
11 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584 (2002).
12 Matsuura, Y. *et al.* Structural basis for the higher Ca(2+)-activation of the regulated actin-activated myosin ATPase observed with Dictyostelium/Tetrahymena actin chimeras. *J Mol Biol* **296**, 579-595 (2000).

13 Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637 (1983).
14 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).
15 Bond, C. S. & Schuttelkopf, A. W. ALINE: a WYSIWYG protein-sequence alignment editor for publication-quality alignments. *Acta Crystallogr D Biol Crystallogr* **65**, 510-512 (2009).
16 Garza, A. G., Harris, B. Z., Greenberg, B. M. & Singer, M. Control of asgE expression during growth and development of Myxococcus xanthus. *J Bacteriol* **182**, 6622-6629 (2000).
17 Diodati, M. E. *et al.* Nla18, a key regulatory protein required for normal growth and development of Myxococcus xanthus. *J Bacteriol* **188**, 1733-1743 (2006).
18 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. *Mol Biol Evol* (2009).
19 Moulton, V., Semple, C. & Steel, M. Optimizing phylogenetic diversity under constraints. *J Theor Biol* **246**, 186-194 (2007).

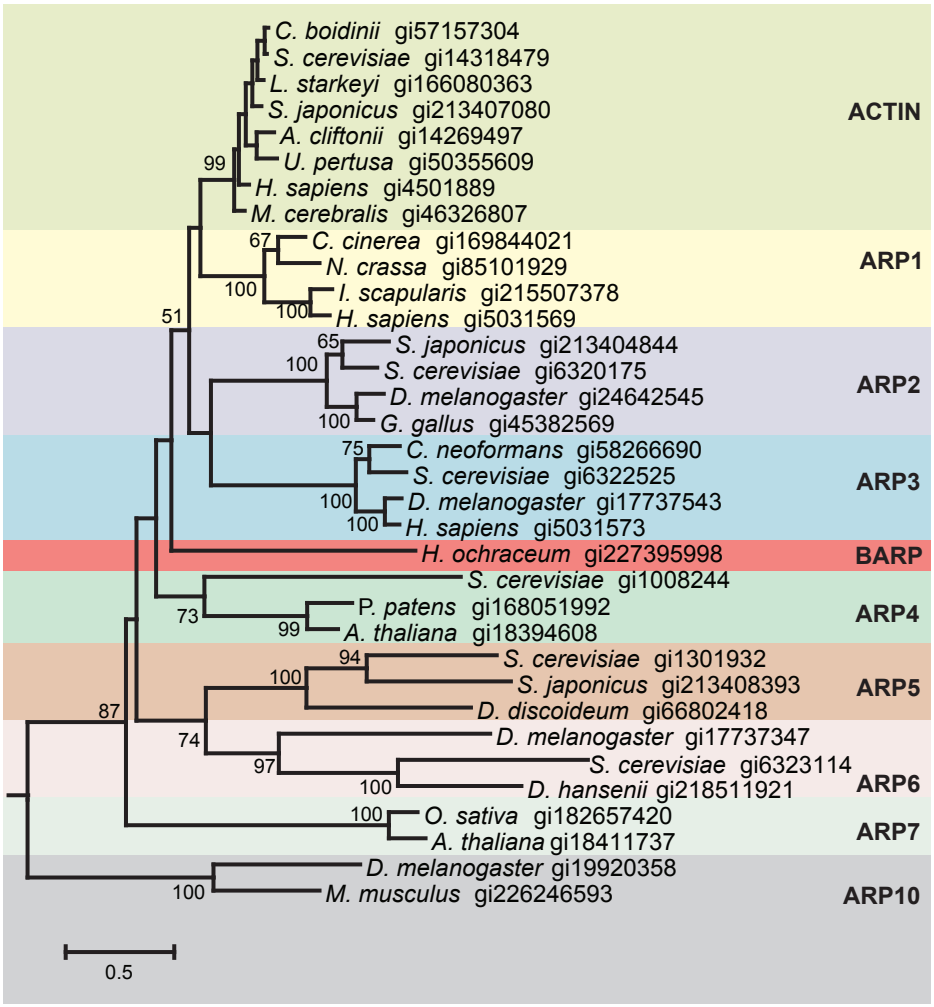

Figure S1. Maximum likelihood tree of the *H. ochraceum* Bacterial Actin Related Protein (BARP) with representatives of eukaryotic actin and eukaryotic actin related protein (ARP) families. The tree was built from a protein multiple sequence alignment using PHYML. The tree was rooted using MreB proteins from *H. ochraceum* and *E. coli* as outgroups (not shown).

Table S1 A. Taxonomic information of the organisms in the GEBA pilot project

Name	Domain	Phylum	Class	Subclass	Order	Suborder	Family
Halogeometricum borinquense PR3	Archaea	Euryarchaeota	Halobacteria	-	Halobacteriales	-	Halobacteriaceae
Halomicrobium mukohataei arg-2	Archaea	Euryarchaeota	Halobacteria	-	Halobacteriales	-	Halobacteriaceae
Halorhabdus utahensis AX-2	Archaea	Euryarchaeota	Halobacteria	-	Halobacteriales	-	Halobacteriaceae
Acidimicrobium ferrooxidans ICP	Bacteria	Actinobacteria	Actinobacteria	Acidimicrobidae	Acidimicrobiales	Acidimicrobineae	Acidimicrobiaceae
Catenulispora acidiphila ID 139908	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Catenulisporineae	Catenulisporaceae
Gordonia bronchialis Tsukamura 3410	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Corynebacterineae	Nocardiaceae
Tsukamurella paurometabola	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Corynebacterineae	Tsukamurellaceae
Geodermatophilus obscurus G-20	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Frankineae	Geodermatophilaceae
Nakamurella multipartita Y-104	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Frankineae	Nakamurellaceae
Stackebrandtia nassauensis LLR-40K-21	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Glycomycineae	Glycomycetaceae
Beutenbergia cavernae HKI 0122	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Micrococcineae	Beutenbergiaceae
Cellulomonas flavigena 134	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Micrococcineae	Cellulomonadaceae
Brachybacterium faecium Schefferle 6-10	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Micrococcineae	Dermabacteraceae
Kytococcus sedentarius 541	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Micrococcineae	Dermacoccaceae
Jonesia denitrificans 55134	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Micrococcineae	Jonesiaceae
Xylanimonas cellulosilytica XIL07	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Micrococcineae	Promicromonosporaceae
Sanguibacter keddieii ST-74	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Micrococcineae	Sanguibacteraceae
Kribbella flavida SW-125	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Propionibacterineae	Nocardioidaceae
Actinosynnema mirum 101	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Pseudonocardineae	Actinosynnemataceae
Saccharomonospora viridis P101	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Pseudonocardineae	Pseudonocardiaceae
Thermobispora bispora R51	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Pseudonocardineae	Pseudonocardiaceae
Nocardiopsis dassonvillei subsp. dassonvillei	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Streptosporangineae	Nocardiopsaceae
Streptosporangium roseum NI 9100	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Streptosporangineae	Streptosporangiaceae
Thermomonospora curvata B9	Bacteria	Actinobacteria	Actinobacteria	Actinobacteridae	Actinomycetales	Streptosporangineae	Thermomonosporaceae
Atopobium parvulum IPP 1246	Bacteria	Actinobacteria	Actinobacteria	Coriobacteridae	Coriobacteriales	Coriobacterineae	Coriobacteriaceae
Cryptobacterium curtum 12-3	Bacteria	Actinobacteria	Actinobacteria	Coriobacteridae	Coriobacteriales	Coriobacterineae	Coriobacteriaceae
Eggerthella lenta VPI 0255	Bacteria	Actinobacteria	Actinobacteria	Coriobacteridae	Coriobacteriales	Coriobacterineae	Coriobacteriaceae
Slackia heliotrinireducens RHS 1	Bacteria	Actinobacteria	Actinobacteria	Coriobacteridae	Coriobacteriales	Coriobacterineae	Coriobacteriaceae
Conexibacter woeseii ID 131577	Bacteria	Actinobacteria	Actinobacteria	Rubrobacteridae	Solirubrobacterales	-	Conexibacteraceae
Capnocytophaga ochracea VPI 2845	Bacteria	Bacteroidetes	Flavobacteria	-	Flavobacteriales	-	Flavobacteriaceae
Chitinophaga pinensis	Bacteria	Bacteroidetes	Sphingobacteria	-	"Sphingobacteriales"	-	Chitinophagaceae
Dyadobacter fermentans NS 114	Bacteria	Bacteroidetes	Sphingobacteria	-	"Sphingobacteriales"	-	Cytophagaceae
Spirosoma linguale Claus 1	Bacteria	Bacteroidetes	Sphingobacteria	-	"Sphingobacteriales"	-	Cytophagaceae
Rhodothermus marinus R-10	Bacteria	Bacteroidetes	Sphingobacteria	-	"Sphingobacteriales"	-	Rhodothermaceae
Pedobacter heparinus HIM 762-3	Bacteria	Bacteroidetes	Sphingobacteria	-	"Sphingobacteriales"	-	Sphingobacteriaceae
Sphaerobacter thermophilus S 6022	Bacteria	Chloroflexi	Thermomicrobia	Sphaerobacteridae	"Sphaerobacteriales"	Sphaerobacterineae	Sphaerobacteraceae
Thermobaculum terrenum YNP1	Bacteria	Chloroflexi	Thermomicrobia	-	Thermomicrobiales	-	Thermomicrobiaceae
Denitrovibrio acetiphilus N2460	Bacteria	Deferribacteres	Deferribacteres	-	Deferribacterales	-	Deferribacteraceae
Meiothermus ruber 21	Bacteria	Deinococcus-Thermus	Deinococci	-	Thermales	-	Thermaceae
Meiothermus silvanus VI-R2	Bacteria	Deinococcus-Thermus	Deinococci	-	Thermales	-	Thermaceae
Alicyclobacillus acidocaldarius subsp. acidocaldarius 104-IA	Bacteria	Firmicutes	Bacilli	-	Bacillales	-	Alicyclobacillaceae
Desulfotomaculum acetoxidans 5575	Bacteria	Firmicutes	Clostridia	-	Clostridiales	-	Peptococcaceae
Anaerococcus prevotii PC 1	Bacteria	Firmicutes	Clostridia	-	Clostridiales	-	Peptostreptococcaceae
Veillonella parvula Te3	Bacteria	Firmicutes	Clostridia	-	Clostridiales	-	Veillonellaceae
Leptotrichia buccalis C-1013-b	Bacteria	Fusobacteria	Fusobacteria	-	Fusobacteriales	-	Leptotrichiaceae
Sebaldella termitidis	Bacteria	Fusobacteria	Fusobacteria	-	Fusobacteriales	-	Leptotrichiaceae
Streptobacillus moniliformis 9901	Bacteria	Fusobacteria	Fusobacteria	-	Fusobacteriales	-	Leptotrichiaceae
Planctomyces limnophilus 290	Bacteria	Planctomycetes	Planctomycea	-	Planctomycetales	-	Planctomycetaceae
Desulfohalobium retbaense HR 100	Bacteria	Proteobacteria	Deltaproteobacteria	-	Desulfovibrionales	-	Desulfohalobiaceae
Desulfomicrobium baculatum X	Bacteria	Proteobacteria	Deltaproteobacteria	-	Desulfovibrionales	-	Desulfomicrobiaceae
Haliangium ochraceum SMP-2	Bacteria	Proteobacteria	Deltaproteobacteria	-	Myxococcales	Nannocystineae	Haliangiaceae
Sulfurospirillum deleyianum 5175	Bacteria	Proteobacteria	Epsilonproteobacteria	-	Campylobacteriales	-	Campylobacteraceae
Kangiella koreensis SW-125	Bacteria	Proteobacteria	Gammaproteobacteria	-	Oceanospirillales	-	Incertae sedis
Brachyspira murdochii 56-150	Bacteria	Spirochaetes	Spirochaetes	-	Spirochaetales	-	Brachyspiraceae
Dethiosulfovibrio peptidovorans SEBR 4207	Bacteria	Synergistetes	Synergistia	-	Synergistales	-	Synergistaceae
Thermanaerovibrio acidaminovorans Su883	Bacteria	Synergistetes	Synergistia	-	Synergistales	-	Synergistaceae

Table S1 B. Characteristics of the organisms in the GEBA pilot project and their genomes.

Name	Collection ID	Number of replicons	G+C(%)	Genome Size (bp)	protein coding genes	RNA genes	16S	CRISPR	Optimum growth temperature	O_2 requirements	Motility	Habitat
Halogeometricum borinquense PR3	DSM 11551	6	59.94	3,944,467	3937	57	2	1	40°C	aerobe	motile	solar saltern
Halomicrobium mukohataei arg-2	DSM 12286	2	65.53	3,332,349	3475	59	3	2	45°C	fac. anaerobe	motile	salt marsh, soil
Halorhabdus utahensis AX-2	DSM 12940	1	62.90	3,116,795	3027	49	1	2	50°C	aerobe	motile	salt lake sediment
Acidimicrobium ferrooxidans ICP	DSM 10331	1	68.29	2,158,157	2038	54	2	2	45°C	aerobe	motile	hot spring
Catenulispora acidiphila ID 139908	DSM 44928	1	66.77	10,467,782	9056	69	3	4	22-28°C	aerobe	nonmotile	soil
Gordonia bronchialis Tsukamura 3410	DSM 43247	n.d.	67.05	draft 5.3 Mb	5167	65	6	0	28°C	aerobe	nonmotile	human pathogen
Tsukamurella paurometabola	DSM 20162	n.d.	68.39	draft 4.5 Mb	4423	65	2	0	10-35°C	aerobe	nonmotile	soil, sludge
Geodermatophilus obscurus G-20	DSM 43160	n.d.	73.99	draft 5.3 Mb	5283	71	3	0	28°C	aerobe	nonmotile	soil, desert
Nakamurella multipartita Y-104	DSM 44233	n.d.	70.90	draft 6.0 Mb	5594	64	2	10	28°C	aerobe	nonmotile	sludge
Stackebrandtia nassauensis LLR-40K-21	DSM 44728	n.d.	68.14	draft 6.8 Mb	6568	53	2	3	28°C	aerobe	nonmotile	soil
Beutenbergia cavernae HKI 0122	DSM 12333	1	73.12	4,669,183	4225	53	2	1	28°C	aerobe	nonmotile	soil
Cellulomonas flavigena 134	DSM 20109	n.d.	74.14	draft 4.0 Mb	3755	52	4	0	30°C	aerobe	motile	soil
Brachybacterium faecium Schefferle 6-10	DSM 4810	1	72.05	3,614,992	3129	69	3	0	25-30 °C	aerobe	nonmotile	soil
Kytococcus sedentarius 541	DSM 20547	1	71.63	2,785,024	2639	64	2	0	28-36 °C	aerobe	nonmotile	skin flora & marine
Jonesia denitrificans 55134	DSM 20603	1	58.42	2,749,646	2558	77	5	0	37 °C	fac. anaerobe	motile	ox blood
Xylanimonas cellulosilytica XIL07	DSM 15894	2	72.45	3,831,380	3485	61	3	0	28 °C	aerobe	nonmotile	soil
Sanguibacter keddieii ST-74	DSM 10542	1	71.89	4,253,413	3735	70	4	0	25-30 °C	fac. anaerobe	motile	bovine blood
Kribbella flavida SW-125	DSM 17836	n.d.	70.57	draft 7.6 Mb	7262	59	2	0	28 °C	aerobe	nonmotile	soil
Actinosynnema mirum 101	DSM 43827	1	73.71	8,248,144	7100	77	5	0	28 °C	aerobe	nonmotile	soil
Saccharomonospora viridis P101	DSM 43017	1	67.32	4,308,349	3906	64	3	9	45 °C	aerobe	nonmotile	soil
Thermobispora bispora R51	DSM 43833	1	72.43	4,189,976	3596	63	4	0	55-65 °C	aerobe	nonmotile	rotten mixed manure
Nocardiopsis dassonvillei subsp. dassonvillei	DSM 43111	n.d.	72.74	draft 6.5 Mb	5634	72	5	8	28 °C	aerobe	nonmotile	blood
Streptosporangium roseum NI 9100	DSM 43021	n.d.	70.77	draft 10.1 Mb	9430	76	7	0	28 °C	aerobe	nonmotile	soil
Thermomonospora curvata B9	DSM 43183	n.d.	71.56	draft 5.6 Mb	5047	86	3	13	50 °C	aerobe	nonmotile	mixed manure
Atopobium parvulum IPP 1246	DSM 20469	1	45.69	1,543,805	1369	50	1	0	37 °C	anaerobe	nonmotile	oral microflora
Cryptobacterium curtum 12-3	DSM 15641	1	50.91	1,617,804	1364	58	3	0	37 °C	obl. anaerobe	nonmotile	oral microflora
Eggerthella lenta VPI 0255	DSM 2243	1	64.20	3,632,260	3123	58	3	1	37 °C	anaerobe	nonmotile	intestinal microflora
Slackia heliotrinireducens RHS 1	DSM 20476	1	60.21	3,165,038	2798	58	2	0	37 °C	anaerobe	nonmotile	sheep rumen
Conexibacter woeseii ID 131577	DSM 14684	n.d.	72.63	draft 6.3 Mb	5979	47	1	0	28 °C	aerobe	motile	soil
Capnocytophaga ochracea VPI 2845	DSM 7271	1	39.59	2,612,925	2193	59	4	1	36 °C	fac. anaerobe	gliding	oral microflora
Chitinophaga pinensis	DSM 2588	1	45.23	9,127,347	7302	95	6	0	23 °C	aerobe	motile	pine litter
Dyadobacter fermentans NS 114	DSM 18053	1	51.54	6,967,790	5804	50	4	0	28-35 °C	aerobe	nonmotile	soil, Zea mays
Spirosoma linguale Claus 1	DSM 74	n.d.	50.12	draft 8.5 Mb	3550	52	1	1	20-30 °C	aerobe	motile	soil, fresh water
Rhodothermus marinus R-10	DSM 4252	2	64.30	3,386,737	2914	48	1	10	65-80 °C	aerobe	nonmotile	marine, hot spring
Pedobacter heparinus HIM 762-3	DSM 2366	1	42.05	5,167,383	4287	57	3	0	25-30 °C	aerobe	motile	soil
Sphaerobacter thermophilus S 6022	DSM 20745	n.d.	68.11	3,988,823	3550	52	1	1	55 °C	aerobe	nonmotile	sludge
Thermobaculum terrenum YNP1	ATCC BAA-798	2	53.51	3,101,581	2872	58	2	0	65-92 °C	aerobe	nonmotile	soil
Denitrovibrio acetiphilus N2460	DSM 12809	n.d.	42.51	draft 3.2 Mb	3183	51	1	0	36 °C	anaerobe	motile	marine and fresh water
Meiothermus ruber 21	DSM 1279	1	63.38	3,097,457	3052	53	2	0	60 °C	aerobe	nonmotile	hot spring
Meiothermus silvanus VI-R2	DSM 9946	n.d.	62.66	draft 3.7 Mb	3624	64	2	8	50 °C	aerobe	nonmotile	hot spring
Alicyclobacillus acidocaldarius subsp. acidocaldarius 104-IA	DSM 446	n.d.	62.08	draft 3.1 Mb	3110	97	3	4	55 °C	aerobe	nonmotile	hot spring
Desulfotomaculum acetoxidans 5575	DSM 771	1	41.55	4,545,624	4370	100	10	0	37 °C	obl. anaerobe	motile	divers
Anaerococcus prevotii PC 1	DSM 20548	2	35.64	1,998,633	1852	61	4	0	37 °C	anaerobe	nonmotile	human microflora
Veillonella parvula Te3	DSM 2008	1	38.63	2,132,142	1859	61	4	0	37 °C	anaerobe	nonmotile	human microflora
Leptotrichia buccalis C-1013-b	DSM 1135	1	29.65	2,465,610	2306	61	5	4	37 °C	anaerobe	nonmotile	blood, oral microflora
Sebaldella termitidis	ATCC 33386	n.d.	33.35	draft 4.5 Mb	4633	55	2	1	mesophile	anaerobe	nonmotile	termite intestine
Streptobacillus moniliformis 9901	DSM 12112	2	26.28	1,673,280	1511	55	5	0	mesophile	fac. anaerobe	nonmotile	human
Planctomyces limnophilus 290	DSM 3776	2	54.68	5,460,085	4304	66	2	1	mesophile	aerobe	motile	marine and fresh water
Desulfohalobium retbaense HR 100	DSM 5692	n.d.	57.29	draft 2.9 Mb	2638	62	2	2	35 °C	anaerobe	motile	saline lake sediment
Desulfomicrobium baculatum X	DSM 4028	1	58.65	3,942,657	3494	72	2	0	30 °C	anaerobe	motile	fresh water/anoxic sedim.
Haliangium ochraceum SMP-2	DSM 14365	n.d.	69.49	draft 9.4 Mb	7063	52	2	5	30-34 °C	aerobe	gliding	costal sands
Sulfurospirillum deleyianum 5175	DSM 6946	n.d.	38.96	draft 2.3 Mb	2412	49	2	6	20-36 °C	anaerobe	motile	anoxic mud
Kangiella koreensis SW-125	DSM 16069	1	43.69	2,852,073	2647	48	2	0	30-37 °C	fac. anaerobe	nonmotile	marine
Brachyspira murdochii 56-150	DSM 12563	n.d.	27.74	draft 3.2 Mb	3016	42	1	1	37°C	fac. anaerobe	motile	intestinal microflora
Dethiosulfovibrio peptidovorans SEBR 4207	DSM 11002	3	54.42	2,576,359	2458	59	5	2	42 °C	obl. anaerobe	motile	oil fields
Thermanaerovibrio acidaminovorans Su883	DSM 6589	1	63.79	1,848,474	1765	60	3	0	55 °C	obl. anaerobe	motile	sludge

obl. =obligate, obligate anaerobes will die when exposed to atmospheric levels of oxygen;

fac. = facultative, facultative anaerobes can use oxygen when present.

All numbers (e.g., number of replicons or genome size) are values as of time of submission of the paper.

Table S1 C. Cell growth condition and DNA preparation for the GEBA pilot project.

	Cell mass production			gDNA isolation			
Species and strain	**Medium**	**Temp.**	**Oxygen**	**Cell lysis**	**Incubation**	**DNA purification**	**DSM ID**
Halogeometricum borinquense PR3	DSMZ 372	35°C	aerobic	st/LALMP	1 hour, 37°C	QIAGEN	11551
Halomicrobium mukohataei arg-2	DSMZ 372	35°C	aerobic	st/L	1 hour, 37°C	QIAGEN	12286
Halorhabdus utahensis AX-2	DSMZ 927	40°C	aerobic	standard	overnight, 35°C	QIAGEN	12940
Acidimicrobium ferrooxidans ICP	DSMZ 709	45°C	aerobic	st/L	1 hour, 37°C	QIAGEN	10331
Catenulispora acidiphila ID 139908	DSMZ 65	28°C	aerobic	st/ALM	overnight, 35°C	JGI CTAP	44928
Gordonia bronchialis Tsukamura 3410	DSMZ 535	28°C	aerobic	st/LALMP	overnight, 35°C	QIAGEN	43247
Tsukamurella paurometabola	DSMZ 535	30°C	aerobic	st/LALMice	1 hour, 37°C	MasterPure	20162
Geodermatophilus obscurus G-20	DSMZ 65	28°C	aerobic	st/LALMP	overnight, 35°C	QIAGEN	43160
Nakamurella multipartita Y-104	DSMZ 553	28°C	aerobic	st/FT	overnight, 35°C	QIAGEN	44233
Stackebrandtia nassauensis LLR-40K-21	DSMZ 553	28°C	aerobic	st/DALM	1 hour, 37°C	QIAGEN	44728
Beutenbergia cavernae HKI 0122	DSMZ 736	28°C	aerobic	st/FT	overnight, 35°C	QIAGEN	12333
Cellulomonas flavigena 134	DSMZ 53	30°C	aerobic	standard	30 min., 37°C	QIAGEN	20109
Brachybacterium faecium Schefferle 6-10	DSMZ 92	28°C	aerobic	standard	30 min., 37°C	QIAGEN	4810
Kytococcus sedentarius 541	DSMZ 92	30°C	aerobic	st/FT	overnight, 35°C	QIAGEN	20547
Jonesia denitrificans 55134	DSMZ 215	37°C	aerobic	st/ALM	overnight, 35°C	JGI CTAP	20603
Xylanimonas cellulosilytica XIL07	DSMZ 92	28°C	aerobic	standard	30 min., 37°C	QIAGEN	15894
Sanguibacter keddieii ST-74	DSMZ 92	30°C	aerobic	standard	1 hour, 37°C	QIAGEN	10542
Kribbella flavida SW-125	DSMZ 830	28°C	aerobic	st/FT	overnight, 35°C	QIAGEN	17836
Actinosynnema mirum 101	DSMZ 535	28°C	aerobic	st/DALM	1 hour, 37°C	QIAGEN	43827
Saccharomonospora viridis P101	DSMZ 535	45°C	aerobic	st/FT	overnight, 35°C	QIAGEN	43017
Thermobispora bispora R51	DSMZ 65	55°C	aerobic	st/FT	overnight, 35°C	QIAGEN	43833
Nocardiopsis dassonvillei subsp. dassonvillei	DSMZ 535	28°C	aerobic	st/DALM	30 min., 37°C	QIAGEN	43111
Streptosporangium roseum NI 9100	DSMZ 535	28°C	aerobic	st/ALM	overnight, 35°C	JGI CTAP	43021
Thermomonospora curvata B9	DSMZ 553	45°C	aerobic	st/LALM	4 hours, 37°C	MasterPure	43183
Atopobium parvulum IPP 1246	DSMZ 104	37°C	anaerobic	st/ALM	overnight, 35°C	JGI CTAP	20469
Cryptobacterium curtum 12-3	DSMZ 78	37°C	anaerobic	st/FT	overnight, 35°C	QIAGEN	15641
Eggerthella lenta VPI 0255	DSMZ 209	37°C	anaerobic	standard	30 min., 37°C	QIAGEN	2243
Slackia heliotrinireducens RHS 1	DSMZ 104	37°C	anaerobic	st/FT	overnight, 35°C	QIAGEN	20476
Conexibacter woeseii ID 131577	DSMZ 92	28°C	aerobic	st/FT	overnight, 35°C	QIAGEN	14684
Capnocytophaga ochracea VPI 2845	DSMZ 340	37°C	anaerobic	st/L	1 hour, 37°C	QIAGEN	7271
Chitinophaga pinensis	DSMZ 67	22°C	aerobic	st/LALMP	overnight, 35°C	QIAGEN	2588
Dyadobacter fermentans NS 114	DSMZ 830	28°C	aerobic	st/FT	overnight, 35°C	QIAGEN	18053
Spirosoma linguale Claus 1	DSMZ 7	28°C	aerobic	st/L	1 hour, 37°C	QIAGEN	74
Rhodothermus marinus R-10	DSMZ 630	65°C	aerobic	st/LALMP	overnight, 35°C	QIAGEN	4252
Pedobacter heparinus HIM 762-3	DSMZ 1	25°C	aerobic	st/LALMP	overnight, 35°C	QIAGEN	2366
Sphaerobacter thermophilus S 6022	DSMZ 467	55°C	aerobic	st/FT	overnight, 35°C	QIAGEN	20745
Thermobaculum terrenum YNP1	ATCC 1981	55°C	aerobic	-	-	-	-
Denitrovibrio acetiphilus N2460	DSMZ 881*	30°C	anaerobic	st/L	30 min., 37°C	QIAGEN	12809
Meiothermus ruber 21	DSMZ 256	50°C	aerobic	st/L	1 hour, 37°C	QIAGEN	1279
Meiothermus silvanus VI-R2	DSMZ 86	50°C	aerobic	st/LALMP	overnight, 35°C	QIAGEN	9946
Alicyclobacillus acidocaldarius subsp.	DSMZ 402	60°C	aerobic	st/L	1 hour, 37°C	QIAGEN	446
Desulfotomaculum acetoxidans 5575	DSMZ 124	37°C	anaerobic	st/LALMP	overnight, 35°C	QIAGEN	771
Anaerococcus prevotii PC 1	DSMZ 104	37°C	anaerobic	st/LALMP	overnight, 35°C	QIAGEN	20548
Veillonella parvula Te3	DSMZ 104	37°C	anaerobic	st/L	1 hour, 37°C	QIAGEN	2008
Leptotrichia buccalis C-1013-b	DSMZ 104	37°C	anaerobic	st/L	1 hour, 37°C	QIAGEN	1135
Sebaldella termitidis	ATCC 1490	37°C	anaerobic	unknown	-	-	-
Streptobacillus moniliformis 9901	DSMZ 429	37°C	aerobic	st/L	1 hour, 37°C	QIAGEN	12112
Planctomyces limnophilus 290	DSMZ 621	28°C	aerobic	standard	1 hour, 37°C	QIAGEN	3776
Desulfohalobium retbaense HR 100	DSMZ 499	35°C	anaerobic	standard	30 min., 37°C	QIAGEN	5692
Desulfomicrobium baculatum X	DSMZ 63	30°C	anaerobic	st/LALMP	overnight, 35°C	QIAGEN	4028
Haliangium ochraceum SMP-2	DSMZ 958#	30°C	aerobic	st/LALMP	overnight, 35°C	QIAGEN	14365
Sulfurospirillum deleyianum 5175	DSMZ 541	29°C	anaerobic	st/L	30 min., 37°C	QIAGEN	6946
Kangiella koreensis SW-125	DSMZ 514	28°C	aerobic	st/L	1 hour, 37°C	QIAGEN	16069
Brachyspira murdochii 56-150	DSMZ 840	37°C	anaerobic	st/L	30 min., 37°C	QIAGEN	12563
Dethiosulfovibrio peptidovorans SEBR 4207	DSMZ 786	42°C	anaerobic	st/FT	overnight, 35°C	QIAGEN	11002
Thermanaerovibrio acidaminovorans Su883	DSMZ 104	55°C	anaerobic	standard	30 min., 37°C	QIAGEN	6589

Medium additions: *1% R2A-medium; #yeast cell paste was replaced by 3 g/l casitone and 1 g/l yeast extract

cell lysis procedures (modification):

- **standard (st)** following standard procedures (as described in original literature or in supplier's manual)
- **st / L** add additional 200 µl lysozyme to standard lysis solution
- **st / ALM** add 500 µL achromopeptidase, lysostaphin, mutanolysin, each, to standard lysis solution
- **st / DALM** add 1000 µL achromopeptidase, 500 µL lysostaphin, 500 µL mutanolysin to standard lysis solution
- **st / LALMP** add additional 100 µL lysozyme; 500 µL achromopeptidase, lysostaphin, mutanolysin, each, to standard lysis solution, reduce proteinase K to 160µl, only
- **st/ LALM** add additional 1 µl lysozyme; 5 µl mutanolysin, achromopeptidase, lysostaphine, each, to standard lysis solution
- **st / LALMice** add additional 4 µL lysozyme; 5 µL mutanolysin, achromopeptidase, lysostaphine, each, to standard lysis solution; after MPC-Step incubate 1hour on ice (extra step)
- **st / FT** freeze 20 min. (-70°C), heat 5 min. (98°C), cool 15 min. to 37°C; add 1.5 ml lysozyme (standard: 0.3 ml, only), 1.0 ml achromopeptidase, 0.12 ml lysostaphine, 0.12 ml mutanolysine, 1.5 ml proteinase K (standard: 0.5 ml, only)

DNA purification

- JGI CTAP: www.jgi.doe.gov
- QIAGEN: Qiagen Genomic 500 Kit (Qiagen 10262)
- MasterPure: MasterPure Gram Positive DNA Purification Kit (Epicentre MGP04100)