The role of culture collections and DNA banks in the Genomic Encyclopedia of *Bacteria* and *Archaea* (GEBA)

'how to produce the new wine to fill the old skins'

Hans-Peter Klenk

WFCC ICCC-12 Conference 2010

Florianópolis, Santa Catarina, Brasil, September 29, 2010

Summary

Introduction to the **G**enomic **E**ncyclopaedia of ***B****acteria* & ***A****rchaea*

The GEBA production pipeline: Collection & DNA Bank part

Metadata and disemmination of results: *SIGS*, *Standards in Genomic Sciences* (Genomic Standards Consortium)

First results and status of GEBA: Chapter 1 published

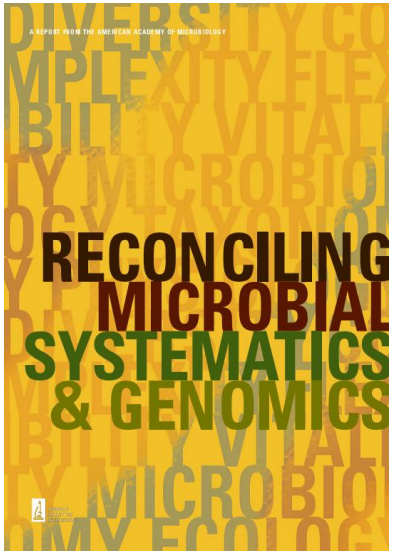
Perspective: The Microbial Earth Project

Recommendations of the American Academy of Sciences Colloquium on *Reconciling Microbial Systematics & Genomics*

... to coordinate an international efford to construct draft genome sequences of each of the roughly 6500 type strains deposited in public strain collections.

... continue with the collection of genome sequences as new species and type strains are described.

... revise and use phenotypic tests to consider genomics data for an improved species definition.



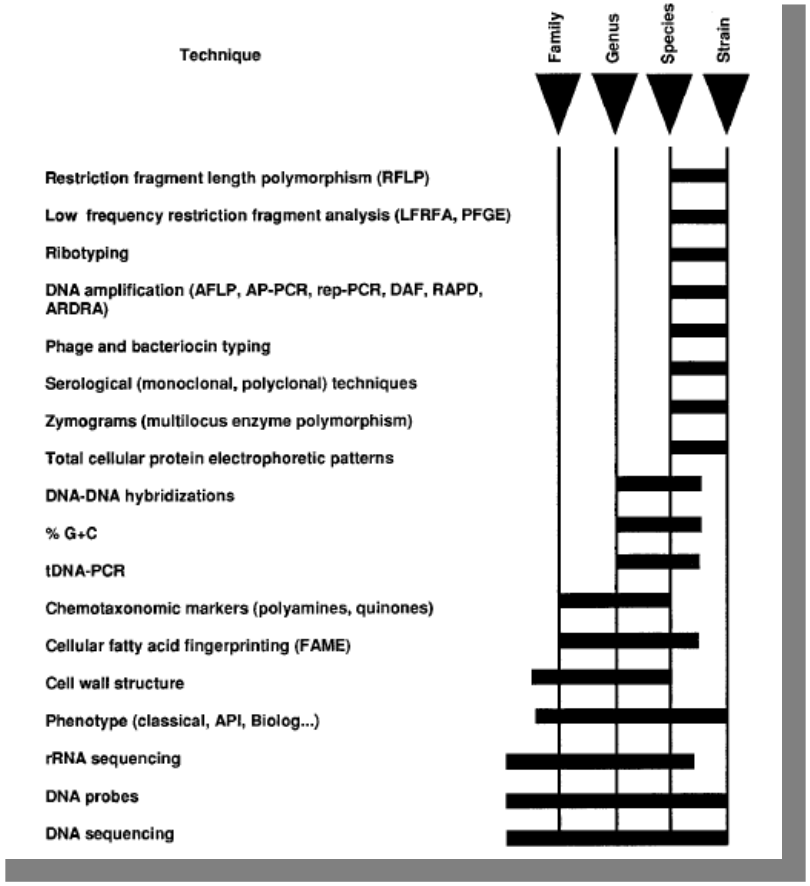

THE HISTORY OF MICROBIAL SPECIES DEFINITIONS, CONVENTIONS (STACKEBRANDT)

Reconciling Microbial Systematics & Genomics

Buckley & Roberts, 2006

Classic Methods for Polyphasic Taxonomy

are based on the analysis of DNA, RNA, proteins, chemotaxonomic markers and expressed features



- laborious
- time consuming
- technically demanding
- standardization required
- not suitable for all organisms

Chemotaxonomy:

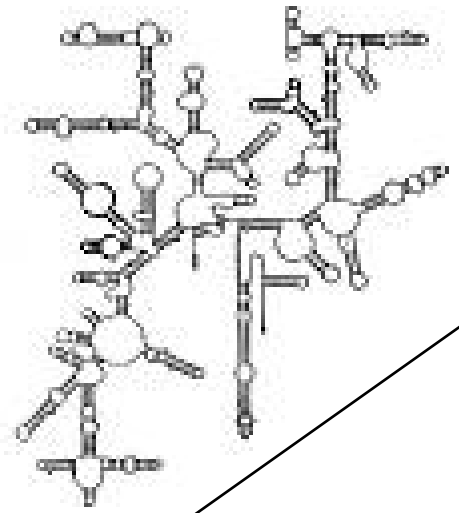
Application of analytical methods to collect information on various chemical constituents of the cells

Vandamme et al. 1996

From Combination of Single Marker Molecules and Physiological Features to the Analysis of Whole Genomes

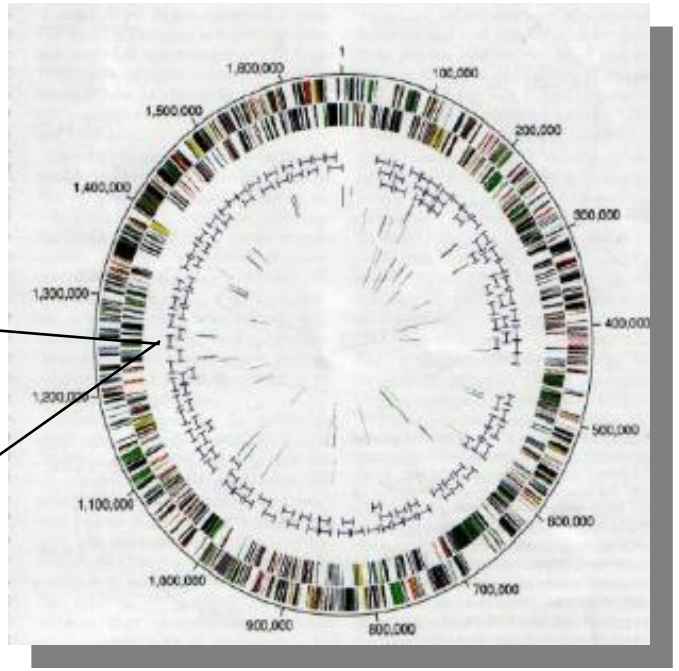
Limitations of 16S rRNA,
the dominant marker molecule

- limited sequence space
- interoperon differences 0-9%
- interstrain differences 0-16%
- seq. identity // strain identity



16S rRNA

ca. 0,1%

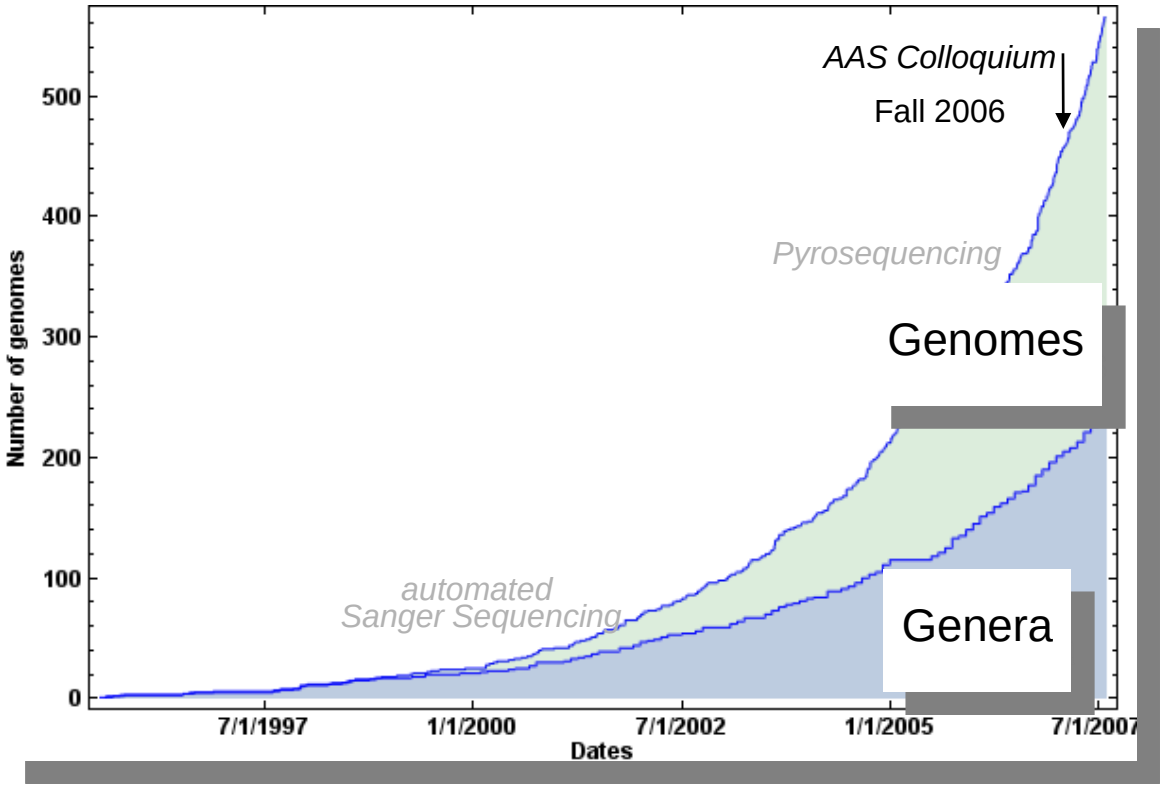


Genome

Growth of Microbial Genome Sequences

September 26th 2010: 1143 bacterial and 94 archaeal completed genome sequences from obout 280 genera
4942 ongoing bacterial genome sequencing projects

data from www.genomesonline.org/gold



Modified from www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_growth.html

Our Aim:

Use of whole-genome sequences to infer the phylogenetic position of organisms in order to support the taxonomic assessment of species

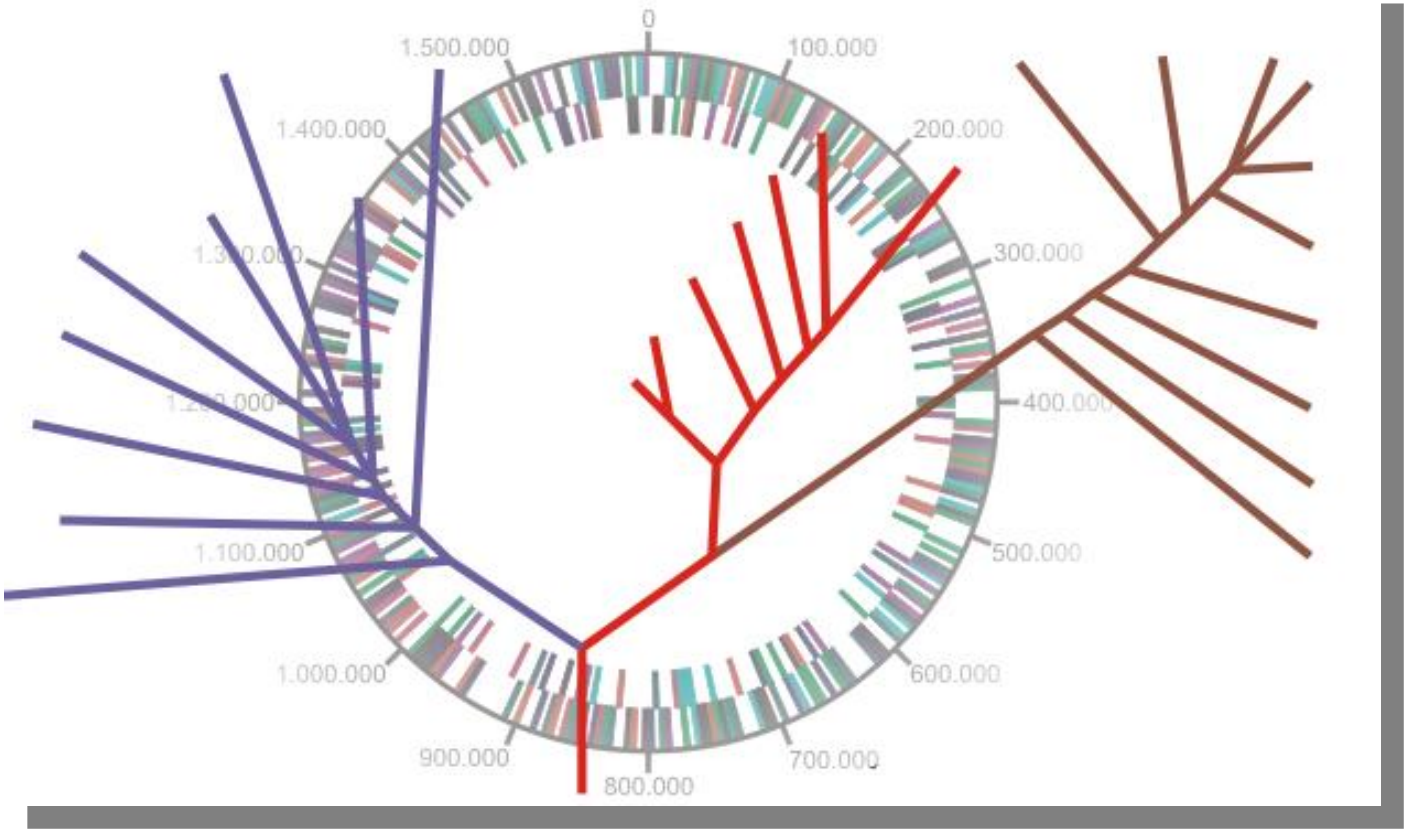


Figure: molecularbiotechnology.ugent.be

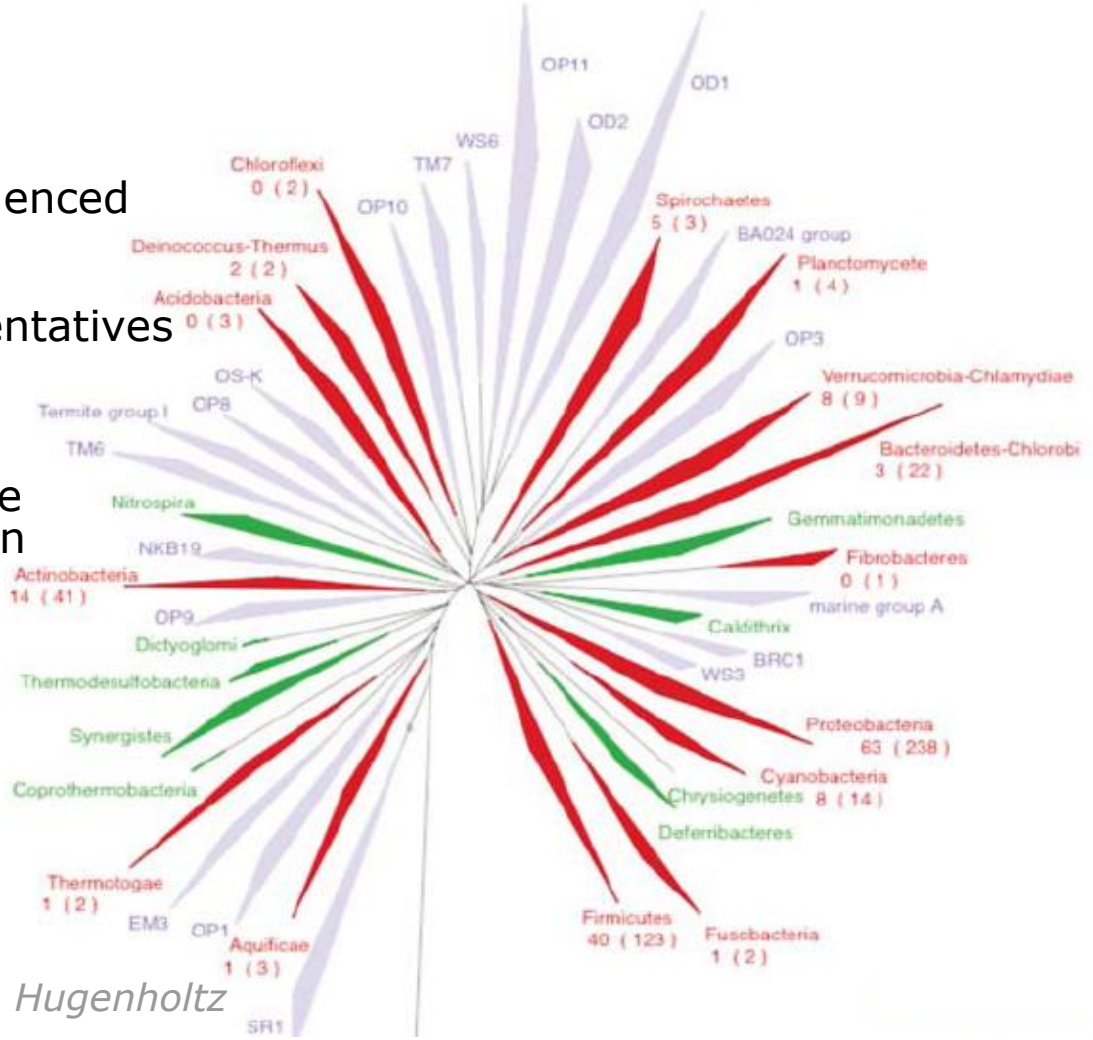
Phylogenetic Tree of the Bacteria Indicating the Status of Known Genome Sequences

pale blue:
no genome sequenced

green:
only few representatives sequenced

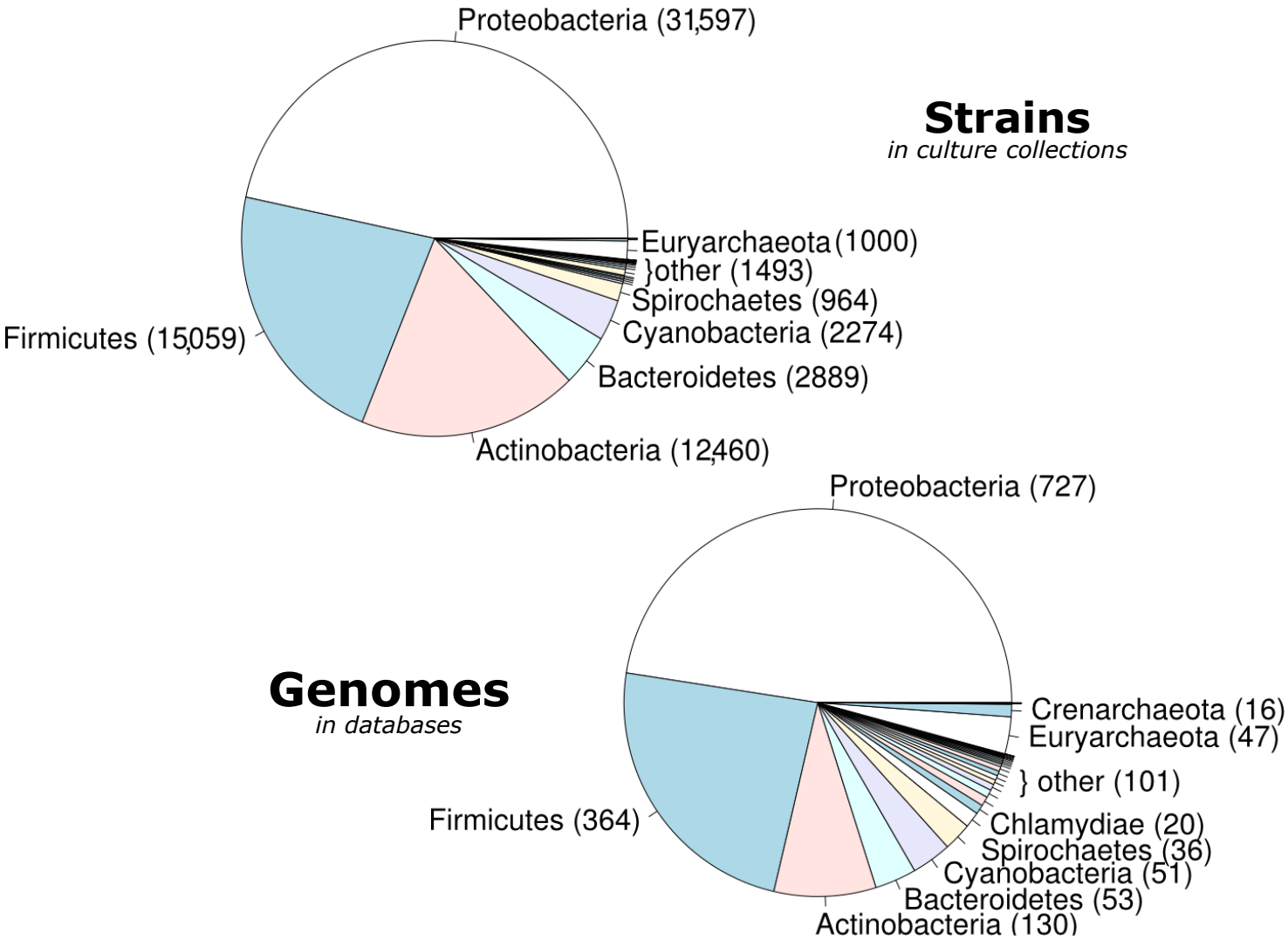
red:
plenty of genome sequences known

Proteobacteria
Firmicutes
Actinobacteria



N.C. Kyrpides & P. Hugenholtz
The Primer, 2007

Diversity of Cultivated Strains and Publicly Available Genome Sequences (Draft and Finished, 2009)



Proposed Solution for the Uneven Sampling of Genome Sequences:

Use a phylogenetic tree for the selection of new genomes to be sequenced

'Key idea of GEBA'

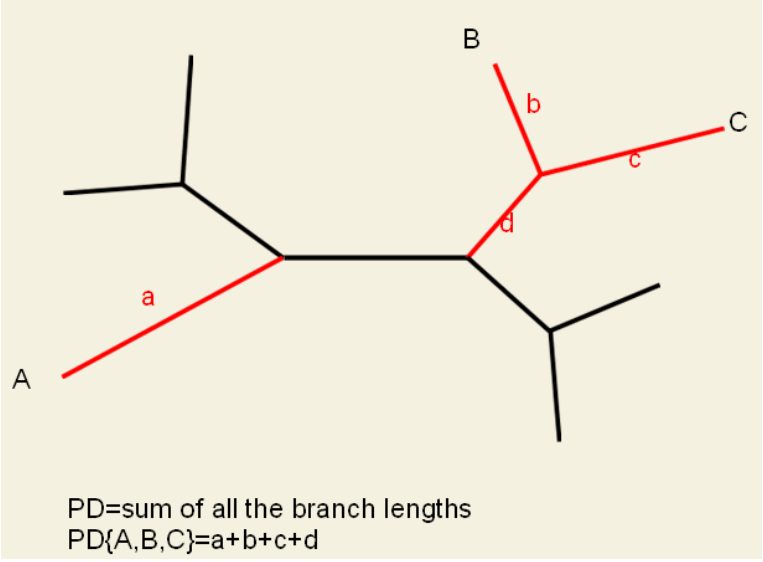
Approach

Take rRNA *Tree of Life*

Overlay finished and in progress genome sequencing projects onto it

Rank phylotypes by how they fill into gaps of the tree

Use phylogenetic diversity as measure for strain selection



Strain selection for GEBA main project
Dongying Wu, JE, PH, NCK, HPK
Phylogenetic Distance (PD) procedure

The GEBA Genomes Production Pipeline

Selection of target strains	phylogeny-driven
Growth of cell paste	small scale, 1-3 g, (ATCC 2%)
Extraction of high molecular gDNA	about 20 variations for cell lysis, (ATCC 2%)
QC, confirmation of identity	16S rRNA sequencing & pulse field (ATCC 2%)
Construction of sequenc. libraries	small and large insert size
Production of draft sequences	Sanger -> var. 454 -> Illumina -> PacBio -> ...
Genome sequence assembly	PHRAP, NEWBLER, VELVET, …
Closure of sequencing gaps	primer walking, PCR, subcloning, …
Sequence editing	CONSED
Gene finding/ORF calling	PRODIGAL
Annotation of genes	Oak Ridge annotation pipeline, manual curation
Maintenance of databases	IMG
Deposition of genome with INSDC	after final annotation
Deposition of genome with GOLD	after release by GenBank
Description of organisms biology	original description in IJSEM, PubMed
Completion of metadata tables	following GSC suggestions
Chemotaxonomy and EM image	cell wall, lipids, quinones, …
Insight into the genome stories	genomic basis of reported phenotypes
Whole-genome phylogenies	truly ‘whole genome’, protein and DNA-based
Phenotype-genotype correlation	pathway reconstructions and BIOLOG
Drafting the Genome Report	for SIGS and other journals
Provide gDNA to colleagues	about 250 DNAs on stock

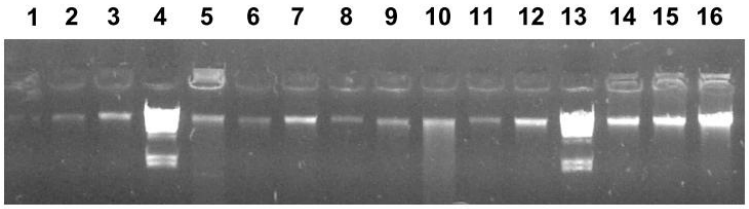
Collection (DSMZ), DNA Bank (DSMZ), Sequencing Center (JGI), jointly JGI-DSMZ

Input of DSMZ to GEBA

Quality controlled high molecular DNAs from type strains

Quantification gel of the genomic DNA isolated from *Veillonella parvula* (DSM 2008T)

Microorganisms



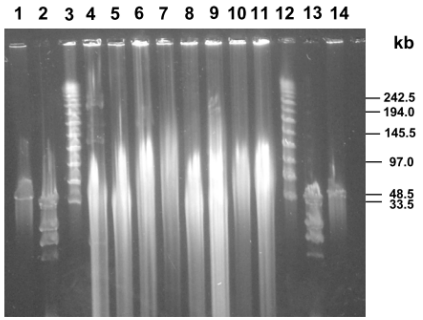
Lane 1: c(λ-Marker)= 15 ng
Lane 2: c(λ-Marker)= 30 ng
Lane 3: c(λ-Marker)= 50 ng
Lane 4: DNA Molecular Weight Marker II (Roche 236250)
Lane 5: DSM 17242, *Alistipes finegoldii;* 5 µl, 1:50
Lane 6: DSM 20460, *Megasphaera elsdenii;* 5 µl, 1:50
Lane 7: DSM 5511, *Haloterrigena turkmenica;* 5 µl, 1:50
Lane 8: DSM 2008, *Veillonella parvula;* 5 µl, 1:50
Lane 9: DSM 20758, *Selenomonas sputigena;* 5 µl, 1:50
Lane 10: DSM 12940, *Halorhabdus utahensis;* 5 µl, 1:50
Lane 11: DSM 1135, *Leptotrichia buccalis;* 5 µl, 1:50
Lane 12: DSM 12286, *Halomicrobium mukohataei;* 5 µl, 1:50
Lane 13: DNA Molecular Weight Marker II (Roche 236250)
Lane 14: c(λ-Marker)= 125 ng
Lane 15: c(λ-Marker)= 250 ng
Lane 16: c(λ-Marker)= 500 ng

Microorganisms

8T) was taken from the German Collection of Microorganisms and Cell Cultures (DSMZ). The
ng the Qiagen Genomic 500 DNA Kit (Qiagen 10262). The genomic DNA was 20-150 kb in size as
el Electrophoresis (PFGE). The bulk of DNA had a size of 50-100 kb (see attached PFGE image).
ng/µl as estimated from the gel. Spectrophotometric measurements yielded a DNA concentration
c DNA are shipped (87,5µg).

DSMZ Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH

Lengths of Genomic DNA Determined by Pulsed Field Gel Electrophoresis (PFGE)



Lane 1: T7 DNA Mass Marker (60 ng)
Lane 2: DNA Molecular Weight Marker II
Lane 3: Mid Range Marker II
Lane 4: DSM 17242, *Alistipes finegoldii*
Lane 5: DSM 20460, *Megasphaera elsdenii*
Lane 6: DSM 5511, *Haloterrigena turkmenica*
Lane 7: DSM 2008, *Veillonella parvula*
Lane 8: DSM 20758, *Selenomonas sputigena*
Lane 9: DSM 12940, *Halorhabdus utahensis*
Lane 10: DSM 1135, *Leptotrichia buccalis*
Lane 11: DSM 12286, *Halomicrobium mukohataei*
Lane 12: Mid Range Marker II
Lane 13: DNA Molecular Weight Marker II
Lane 14: T7 DNA Mass Marker (60 ng)

PFGE of the genomic DNA of the strains was performed in a contour-clamped homogeneous electric field (CHEF) system on a CHEF-DR III device (Bio-Rad Laboratories, Hercules, Calif.) with 1% agarose gels and modified 0.5 TBE buffer (45 mM Tris, 45 mM boric acid, 0.1 mM EDTA) at 14°C. PFGE times used at 200 V (6 V/cm) were 1 to 15 s for 18 h.

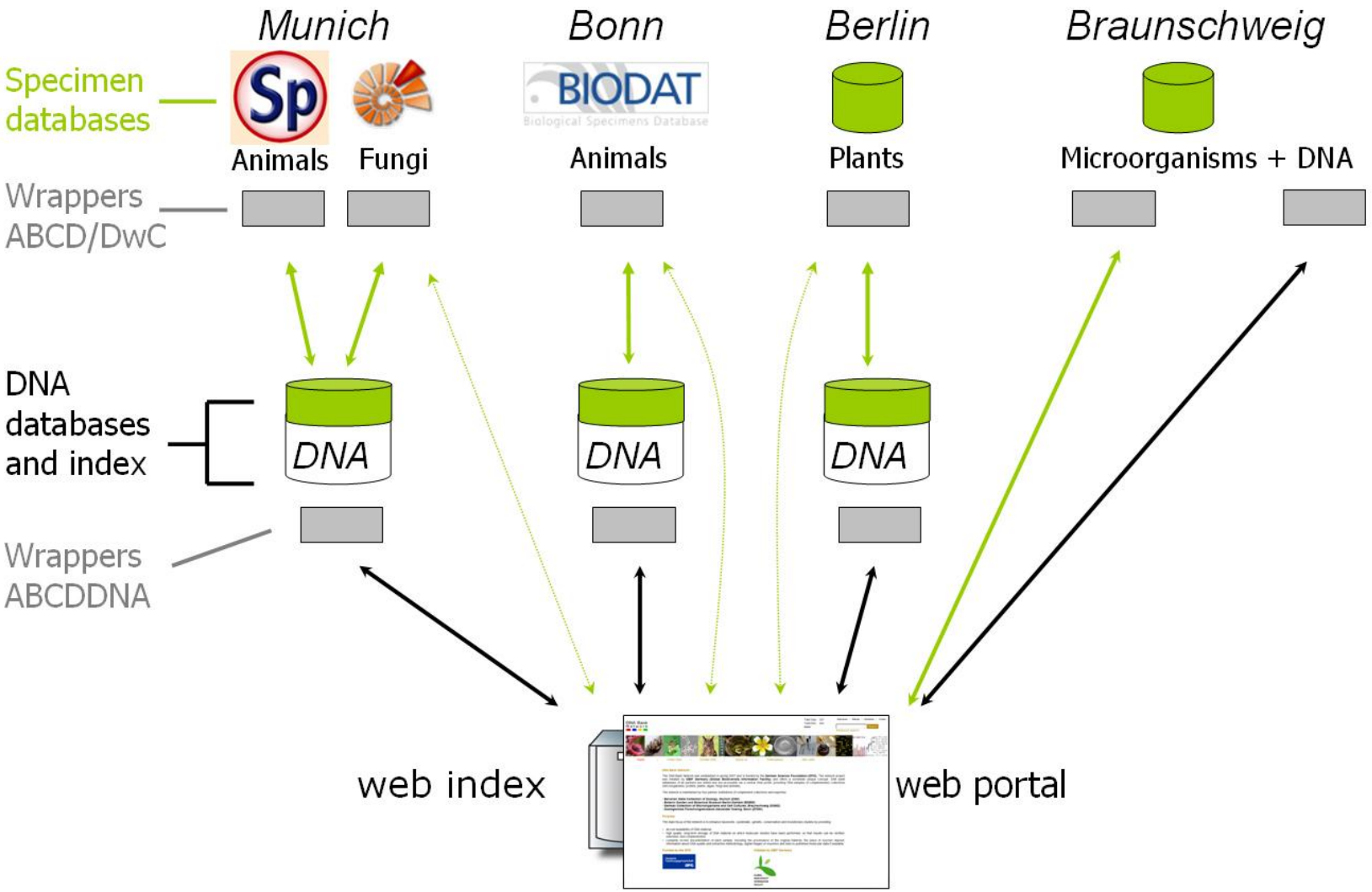
DSMZ Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH

Input of DSMZ to GEBA

Systematic collection of metadata including classi-fication and phylogentic analysis (rRNA and whole-genome based).

The DNA Bank Network http://www.dnabank-network.org/

DNA Bank Data Flow



The GEBA Genomes Production Pipeline

Selection of target strains	phylogeny-driven
Growth of cell paste	small scale, 1-3 g, (ATCC 2%)
Extraction of high molecular gDNA	about 20 variations for cell lysis, (ATCC 2%)
QC, confirmation of identity	16S rRNA sequencing & pulse field (ATCC 2%)
Construction of sequenc. libraries	small and large insert size
Production of draft sequences	Sanger -> var. 454 -> Illumina -> PacBio -> ...
Genome sequence assembly	PHRAP, NEWBLER, VELVET, …
Closure of sequencing gaps	primer walking, PCR, subcloning, …
Sequence editing	CONSED
Gene finding/ORF calling	PRODIGAL
Annotation of genes	Oak Ridge annotation pipeline, manual curation
Maintenance of databases	IMG
Deposition of genome with INSDC	after final annotation
Deposition of genome with GOLD	after release by GenBank
Description of organisms biology	original description in IJSEM, SAM, PubMed
Completion of metadata tables	following GSC suggestions, MIGS
Chemotaxonomy and EM image	cell wall, lipids, quinones, …
Insight into the genome stories	genomic basis of reported phenotypes
Whole-genome phylogenies	truly ‘whole genome’, protein- and DNA-based
Phenotype-genotype correlation	pathway reconstructions and BIOLOG
Drafting the Genome Report	for SIGS and other journals
Provide gDNA to colleagues	about 250 DNAs on stock

Collection (DSMZ), DNA Bank (DSMZ), Sequencing Center (JGI), jointly JGI-DSMZ

Standardisation and Dissemination

Standards in Genomic Sciences

An Open Access Journal of the Genomic Standards Consortium

Standards in Genomic Sciences

In the current issue (published July 28, 2010 - August 31, 2010)

Short Genome Reports (GEBA)

Chang et al., Complete genome sequence of *Acidaminococcus fermentans* type strain (VR4^T)
Abt et al., Complete genome sequence of *Cellulomonas flavigena* type strain (134^T)
Tindall et al., Complete genome sequence of *Meiothermus ruber* type strain (21^T)
Sikorski et al., Complete genome sequence of *Meiothermus silvanus* type strain (VI-R2^T)
LaButti et al., Complete genome sequence of *Planctomyces limnophilus* type strain (Mü 290^T)
Sikorski et al., Complete genome sequence of *Acetohalobium arabaticum* type strain (Z-7288^T)
Göker et al., Complete genome sequence of *Ignisphaera aggregans* type strain (AQ1.S1^T)
Göker et al., Complete genome sequence of *Olsenella uli* type strain (VPI D76D-27C^T)
LaButti et al., Permanent draft genome sequence of *Dethiosulfovibrio peptidovorans* type strain (SEBR 4207^T)

Meeting Reports

Kyrpides et al., Meeting Report from the Genomic Standards Consortium (GSC) Workshop 8

The GEBA Report Series in *Stand Genomic Sci*

Standards in Genomic Sciences, Vol 1, No 1 (2009)

SIGS *Stand. Genomic Sci.* 2009 1:1
ISSN 1944-3277
doi:10.4056/sigs.1463

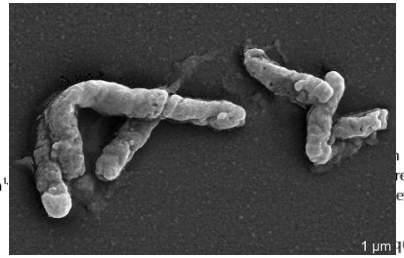
Complete genome sequence of *Acidimicrobium ferrooxidans* type strain (ICP^T)

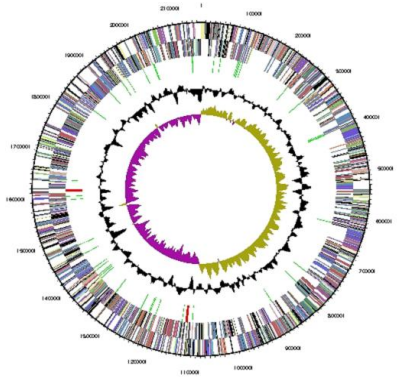
Alicia Clum[1], Matt Nolan[1], Elke Lang[2], Tijana Glavina Del Rio[1], Hope Tice[1], Alex Copeland[1], Jan-Fang Cheng[1], Susan Lucas[1], Feng Chen[1], David Bruce[3], Lynne Goodwin[3], Sam Pitluck[1], Natalia Ivanova[1], Konstantinos Mavrommatis[1], Natalia Mikhailova[1], Amrita Pati[1], Amy Chen[4], Krishna Palaniappan[4], Markus Göker[2], Stefan Spring[2], Miriam Land[5], Loren Hauser[5], Yun-Juan Chang[5], Cynthia C. Jeffries[5], Patrick Chain[1] Jim Bristow[1], Jonathan A. Eisen[1,7], Victor Markowitz[4], Philip Hugenholtz[1], Nikos C. Kyrpides[1], Hans-Peter Klenk[2], Alla Lapidus[1*]

[1] DOE Joint Genome Institute, Walnut Creek, California, USA
[2] DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany
[3] Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico USA
[4] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[5] Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[6] Lawrence Livermore National Laboratory, Livermore, California, USA
[7] University of California Davis Genome Center, Davis, California, USA

* Corresponding author: AllaLapidus

Print publication date: July 20, 2009.





Abstract

Acidimicrobium ferrooxidans (Clark and Norris 1996) is the sole and type species of only genus within the actinobacterial family *Acidimicrobiaceae* and in the order *Acidi* pyrite during autotrophic growth in the absence of an enhanced CO_2 concentration is we describe the features of this organism, together with the complete genome sequ complete genome sequence of the order *Acidomicrobiales*, and the 2,158,157 bp lo protein coding and 54 RNA genes is part of the *Genomic Encyclopedia of Bacteria ar*

Keywords: Moderate thermophile, ferrous-iron-oxidizing, acidophile, *Acidomicrobiales*.

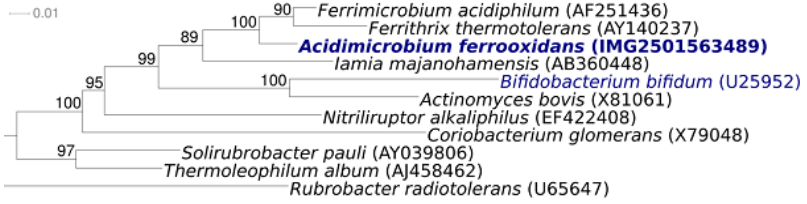
Table 1. Classification and general features of *A. ferrooxidans* ICP^T based on MIGS recommendations [11]

MIGS ID	Property	Term	Evidence code[a,b]
	Current classification	Domain *Bacteria*	
		Phylum *Actinobacteria*	TAS [12]
		Class *Actinobacteria*	TAS [13]
		Order *Acidimicrobiales*	TAS [13]
		Suborder *Acidimicrobineae*	
		Family *Acidomicrobiaceae*	TAS [13]
		Genus *Acidimicrobium*	TAS [1]
		Species *Acidimicrobium ferrooxidans*	TAS [1]
		Type strain ICP	
		positive	TAS [1]
		rod shaped	TAS [1]
		motile	TAS [1]
		nonsporulating	TAS [1]
	re range	moderate thermophile, 45-50°C	TAS [1]
	emperature	48°C	TAS [1]
		not reported	
	uirement	aerobic	TAS [1]
	Carbon source	CO_2 (autotrophic), yeast extract (heterotrophic)	TAS [1]
	Energy source	autotrophic: oxidation of ferrous iron with oxygen as the electron acceptor; heterotrophic: yeast extract	TAS [1]
MIGS-6	Habitat	warm, acidic, iron-, sulfur-, or mineral-sulfide rich environments	TAS [1]
MIGS-15	Biotic relationship	free living	NAS
MIGS-14	Pathogenicity	none	NAS
	Biosafety level	1	TAS [14]
	Isolation	hot springs	TAS [2]
MIGS-4	Geographic location	Krísuvik geothermal area, Iceland	TAS [2]
MIGS-5	Sample collection time	before 1993	TAS [1]
MIGS-4.1 MIGS-4.2	Latitude – Longitude	63.93, -22.1	TAS [2]
MIGS-4.3	Depth	not reported	
MIGS-4.4	Altitude	not reported	

Evidence codes - IDA: Inferred from Direct Assay (first time in publication); TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project [15]. If the evidence code is IDA the property was directly observed for a live isolate by one of the authors or an expert mentioned in the acknowledgements.

Table 2. Genome sequencing project information

MIGS ID	Property	Term
MIGS-31	Finishing quality	Finished
		One Sanger library: 8kb pMCL200
MIGS-28	Libraries used	One 454 pyrosequence standard library and one Illumina library
MIGS-29	Sequencing platforms	ABI3730, 454 GS FLX, Illumina GA
MIGS-31.2	Sequencing coverage	6.8 x Sanger; 52.9 x pyrosequence
MIGS-30	Assemblers	Newbler, Arachne
MIGS-32	Gene calling method	Prodigal
	INSDC / Genbank ID	CP001631
	Genbank Date of Release	not available
	GOLD ID	Gc01023
	Database: IMG-GEBA	2501533204
MIGS-13	Source material identifier	DSM 10331
	Project relevance	Tree of Life, GEBA



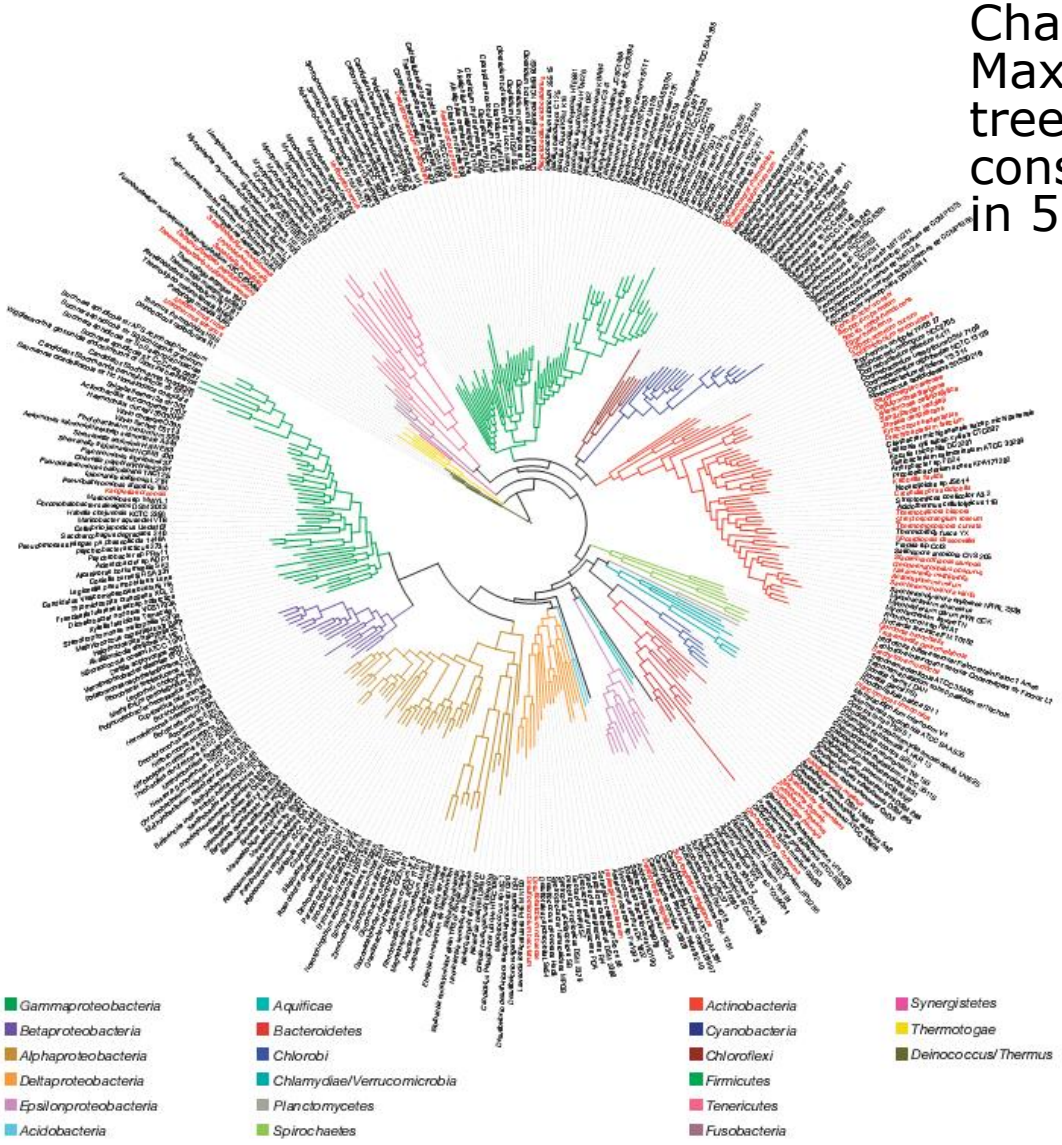
Novel Approach Towards Research Coordination
Genomics Standards Consortium

GSC Mission

Community-driven standards have the best chance of success if developed within the auspices of international working groups. Participants in the GSC include biologists, computer scientists, those building genomic databases and conducting large-scale comparative genomic analyses, and those with experience of building community-based standards.

The mission of the GSC is to work with the wider community towards:

* the implementation of new genomic **standards**
* methods of capturing and exchanging **metadata**
* harmonization of metadata collection and analysis efforts across the wider genomics community



Chapter I (December 24th 2009) Maximum-likelihood phylogenetic tree based on 31 concatenated conserved protein-coding genes in 56 genomes

JGI DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

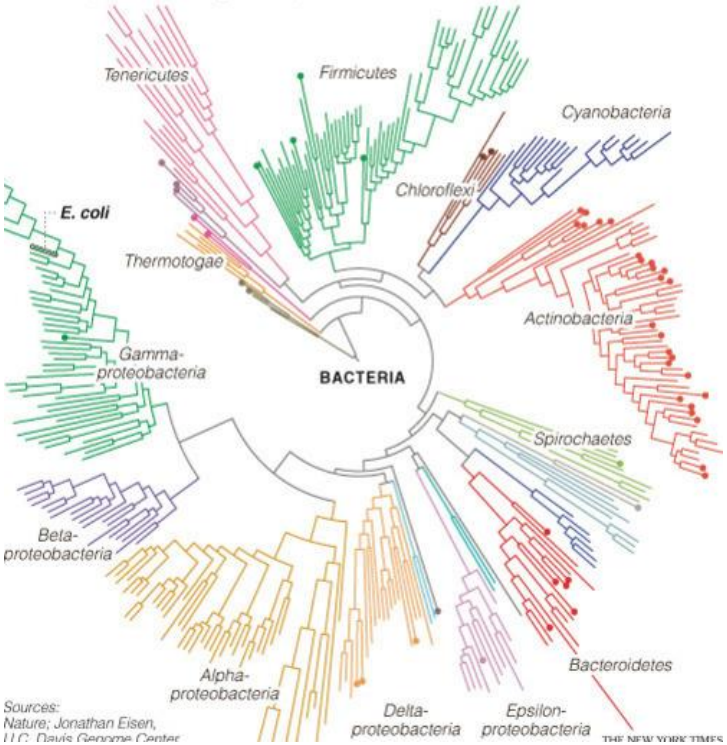
Status September 26th 2010

published:	65
submitted:	12
finished:	44
in closure:	46
in sequencing:	~100
total:	~300

First 56 GEBA Genomes Published …

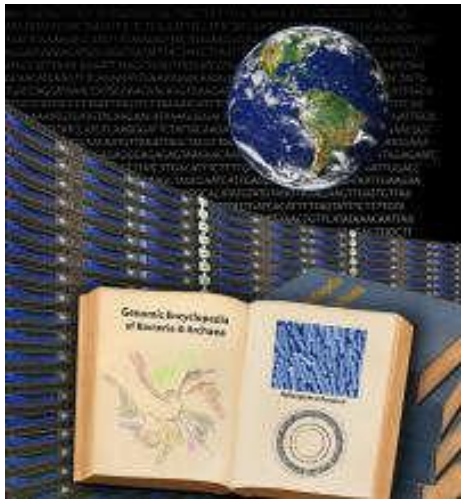
Filling Out the Branches

This "genome tree" shows relationships among the different species of bacteria that have had their genomes sequenced to date, with major phyla shown in different colors. A new project intended to expand the range and variety of sequenced microbes has completed its first 56 species, including the 53 species of bacteria marked below with dots.



Sources: Nature; Jonathan Eisen, U.C. Davis Genome Center

THE NEW YORK TIMES



nature news

Published online 23 December 2009 | Nature | doi:10.1038/news.2009.1161

News

Microbial encyclopaedia guided by evolution

Sequencing project reveals microbial cache of protein families.

December 29, 2009

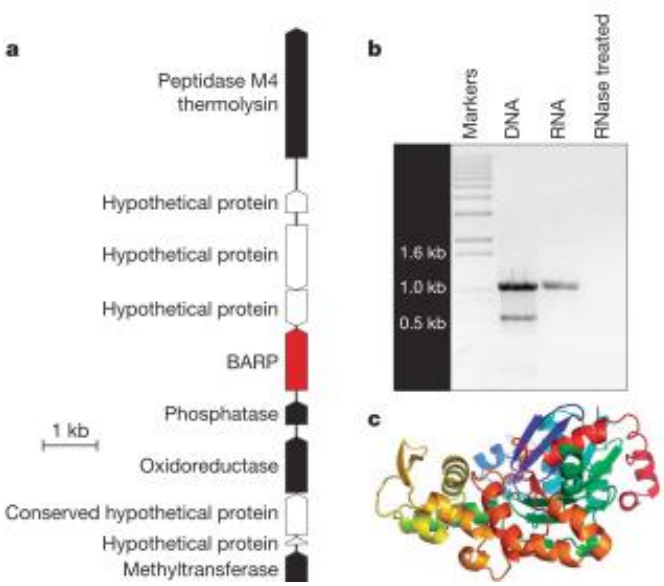
Scientists Start a Genomic Catalog of Earth's Abundant Microbes

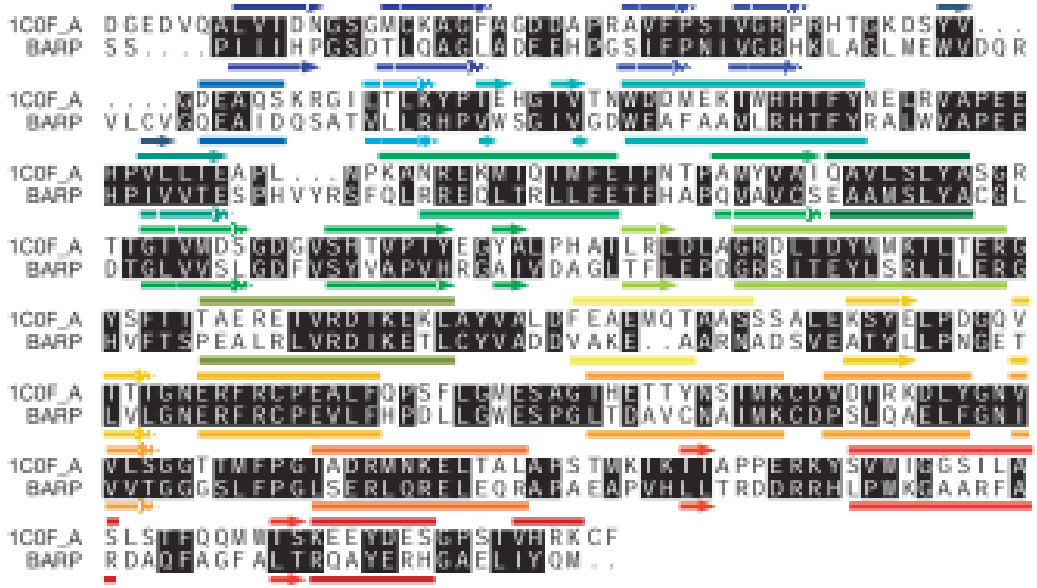
Results: Effect of 16S rRNA Tree-Based Selection of Organisms on Comparative Genomics Metrics

Comparative genomic metric	GEBA set	Random sets (number of resamplings)	Fold improvement
Genome tree phylogenetic diversity[17]			
Bacteria (domain)	11.0	3.2 ± 0.7 (100)	2.8–4.4
Actinobacteria (phylum)	4.3	1.4 ± 0.3 (100)	2.5–3.9
New protein family links	46	3 ± 4 (5)	6.6 to >15.3
Genes in new chromosomal cassettes	71,579	16,579 ± 5,523 (20)	3.2–6.5
New gene fusions	433	65 ± 31 (20)	4.5–12.7

GEBA genomes were compared to equivalently sized random sets of reference genomes to quantify the effect of phylogenetic selection.

Results: **A Bacterial Homologue of Actin**





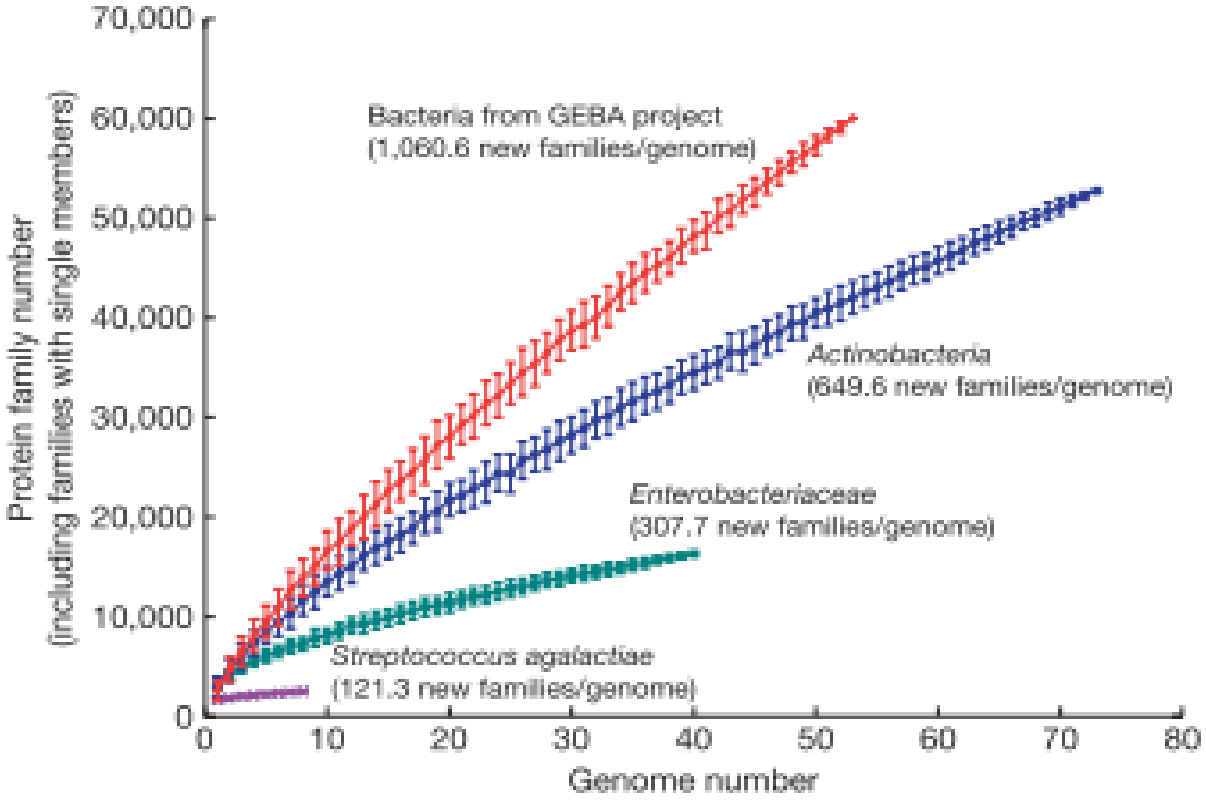
a) Genomic context of the bacterial actin-related protein (BARP) gene within the genome of the marine Deltaproteobacterium *H. ochraceum*. Red, gene encoding BARP; white, genes encoding hypothetical proteins; black, genes with functional annotations.

b) RT–PCR demonstration of expression of the gene encoding BARP in *H. ochraceum*.

c) Ribbon plot of the putative structure of BARP.

d) Alignment of BARP with actin from *Dictyostelium discoideum* with similarities in black shaded text. Secondary structure elements (arrows, beta-strands; bars, alphahelices) are colour-coded as in c

Results: Rate of Discovery of Protein Families as a Function of Phylogenetic Breadth of Genomes

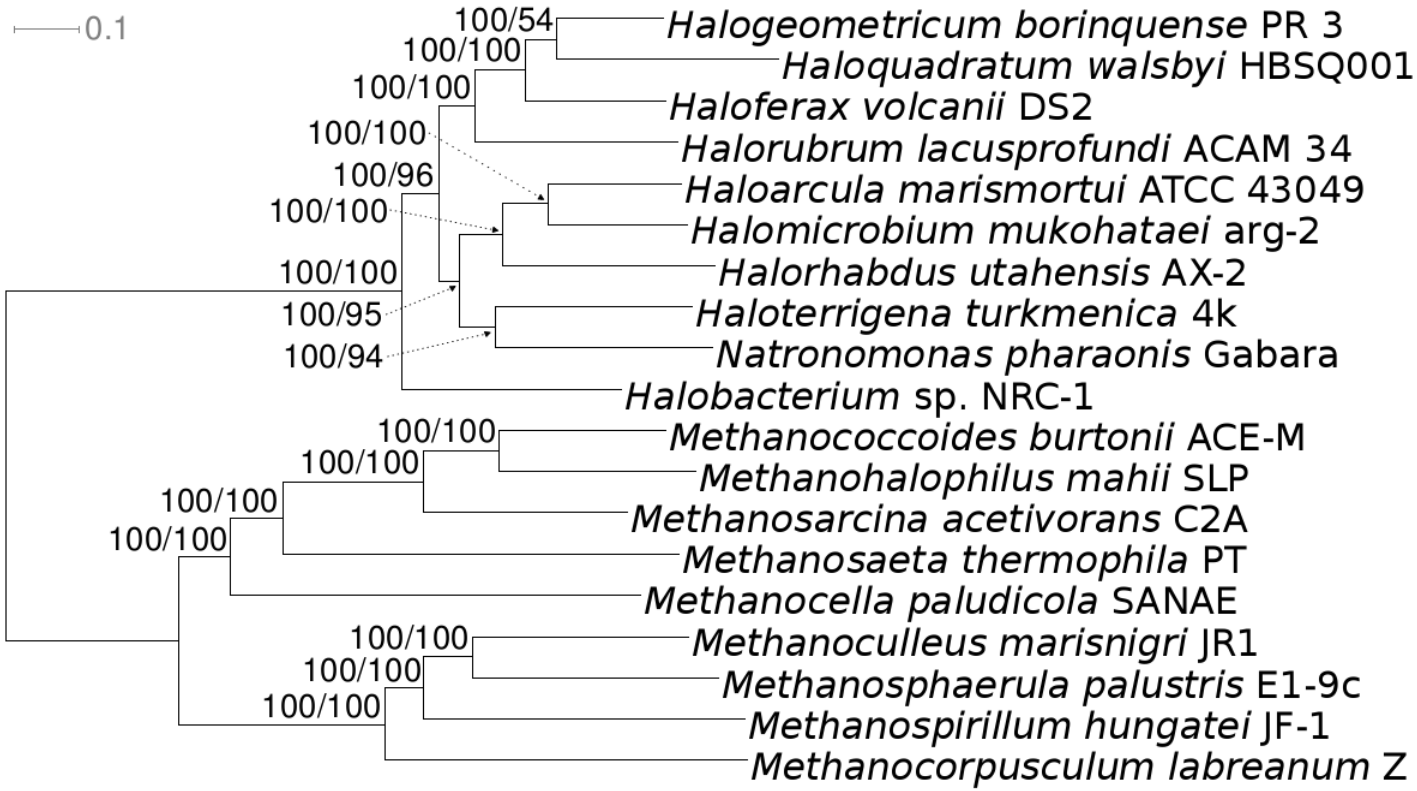


For each of four groupings (species, different strains of ***Streptococcus agalactiae***; family, ***Enterobacteriaceae***; phylum, ***Actinobacteria***; domain, **GEBA-Bacteria**), all proteins from that group were compared to each other to identify protein families. Then the total number of protein families was calculated as genomes were progressively sampled from the group (starting with one genome until all were sampled). This was done multiple times for each of the four groups using random starting seeds; the average and standard deviation were thenplotted.

Whole-genome-sequence analysis software

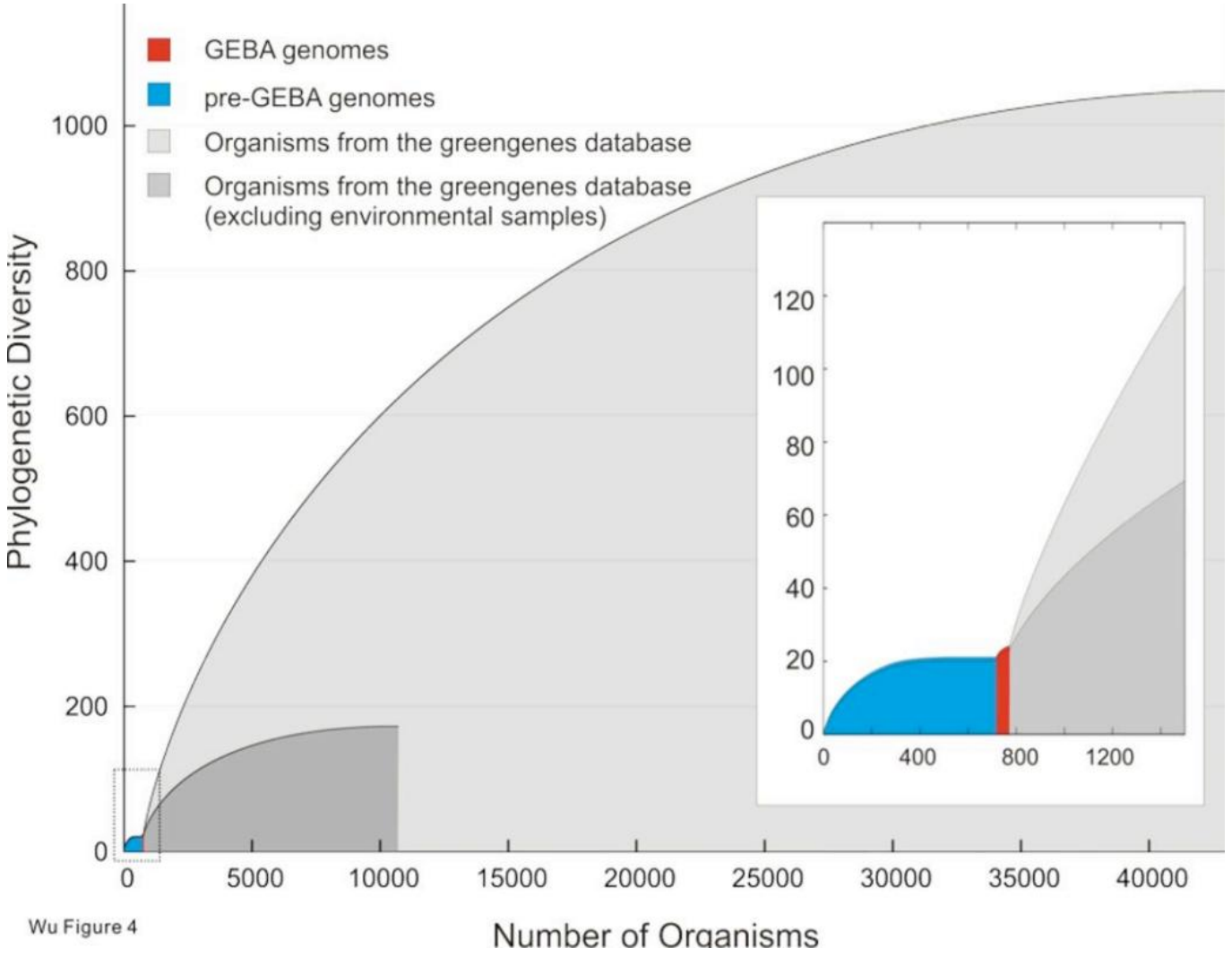
Extremely Well Supported Phylogenies From Genome Sequences and Proteomes

- generated by automated data handling pipelines



Halobacteriaceae

Results: Estimation of Number of Genomes to be Required for a Overview on Genomic Diversity



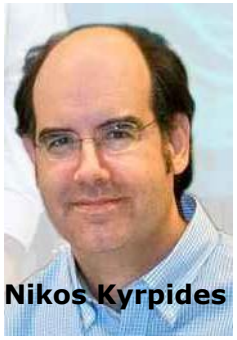
Wu Figure 4

Perspective: The Microbial Earth Project

A systematic, genomic exploration
of all validly-named species
of bacteria and archaea.

The ambitious but assuredly tractable* **goal** of the project, is **to sequence the genome of at least one representative of every bacterial and archaeal species** that has a validly published name in conformance with the Bacteriological Code.

* given an international effort and current technologies

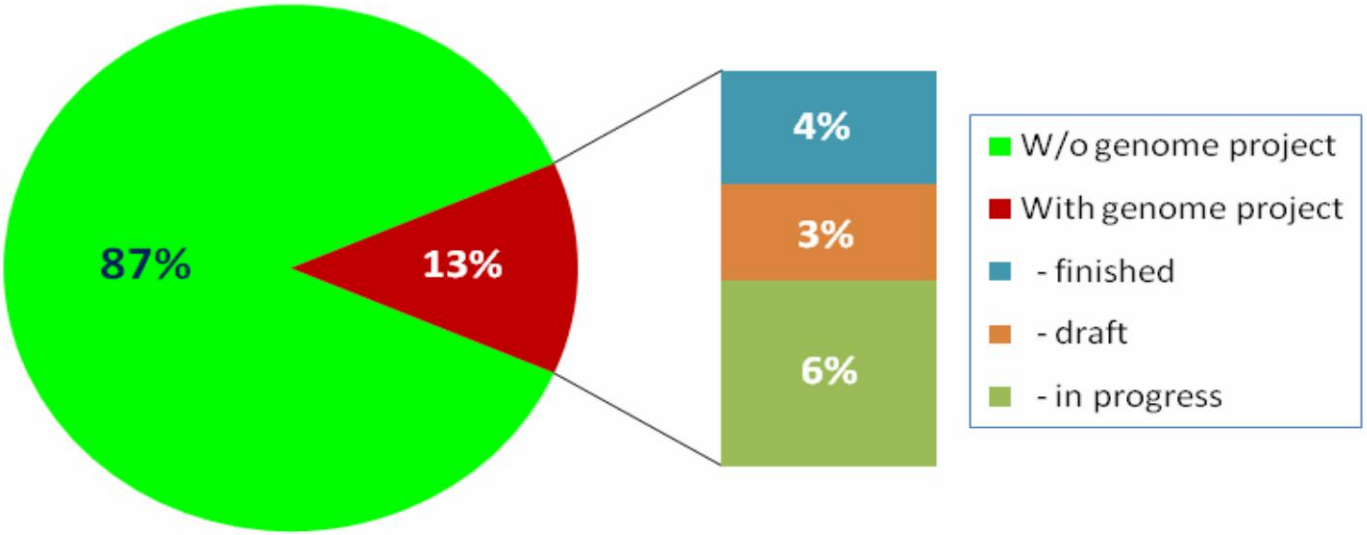

Nikos Kyrpides


Philip Hugenholtz

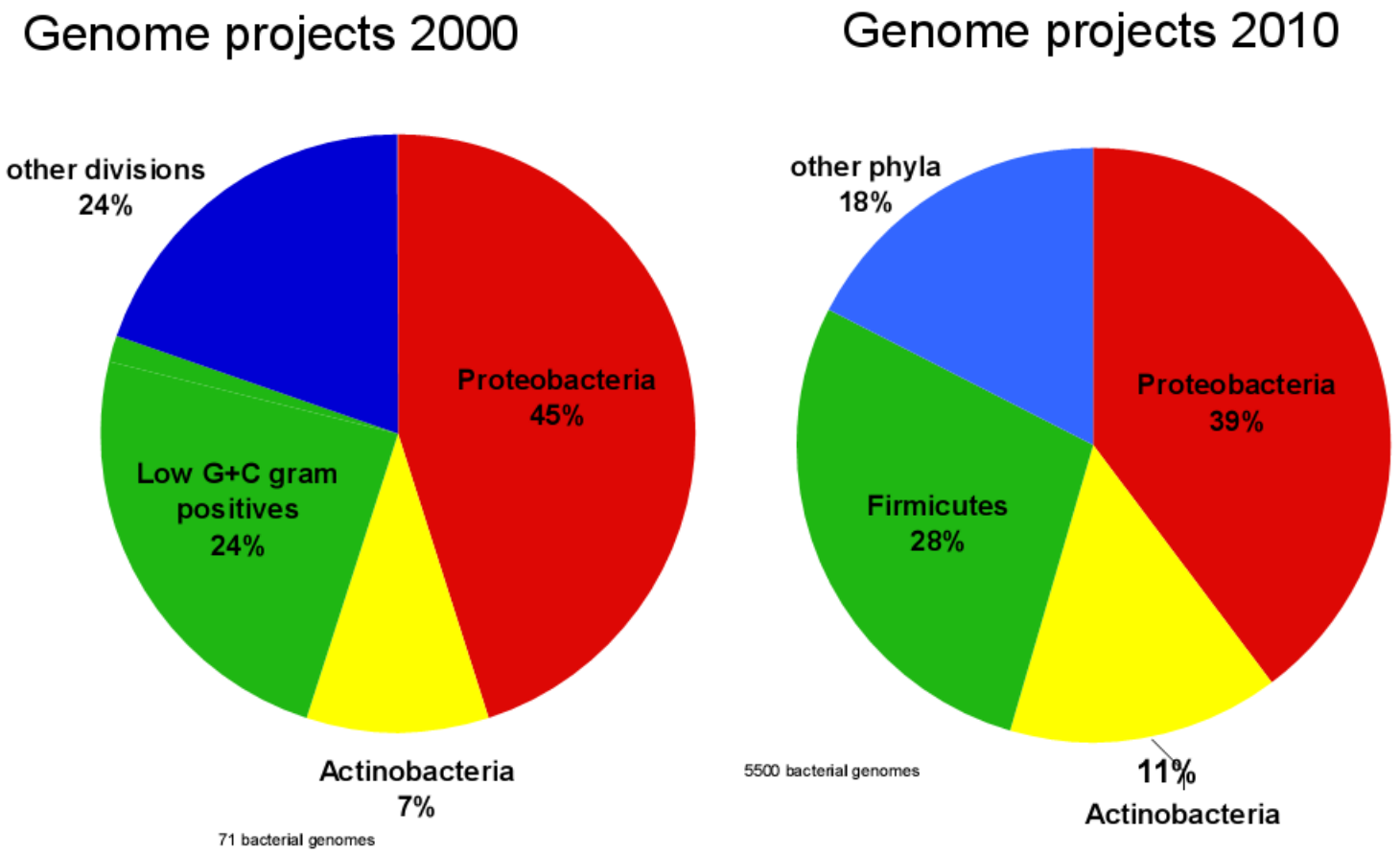

George Garrity

Status: Genome Project Coverage of Bacterial and Archaeal Type Strains

Phase I:
Sequence one representative from every characterized microbial **type species**

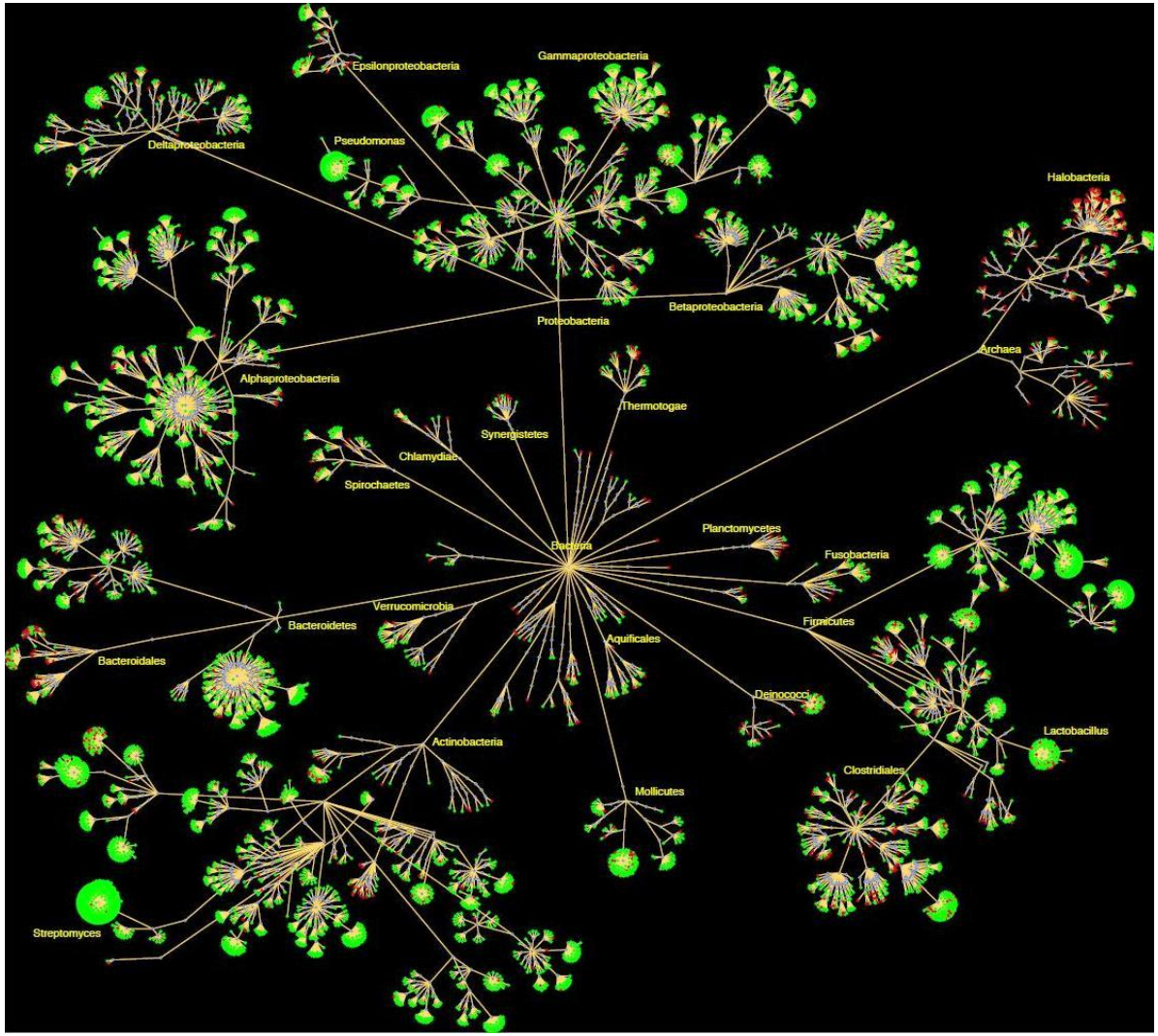


What we Need to Change …



… with a systematic approach.

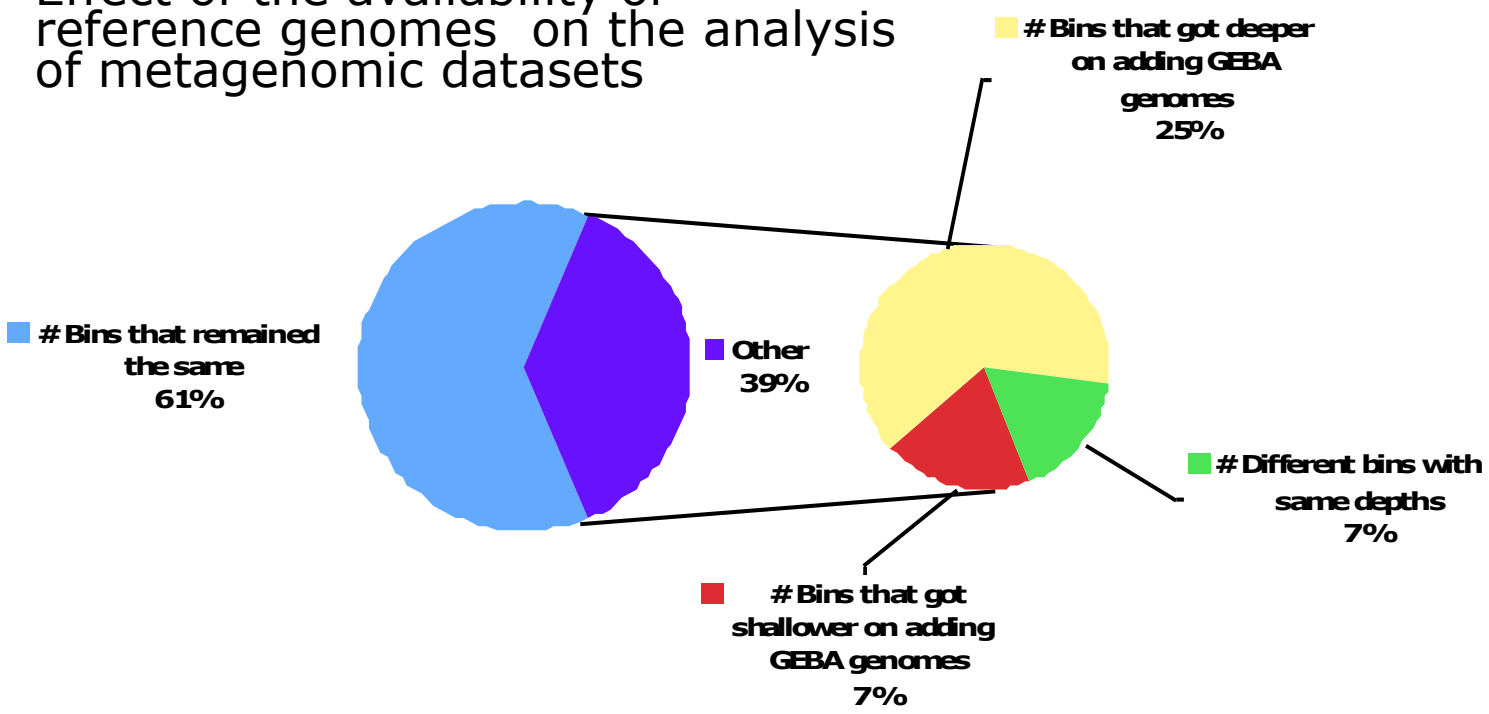
Graphical Representation of a Tree of DSMZ strains with (red) and without (green) genome sequencing projects



Victor Kunin
MEP (JGI)

Binning Changes in the Soil Metagenome on Adding GEBA Genomes

Effect of the availability of reference genomes on the analysis of metagenomic datasets



Without the GEBA/MEP framework,
the exploration of our microbial planet
is equivalent to navigation without a compass,
a map or stars by which one can fix their position.

The Already Identified Partners For the Microbial Earth Project

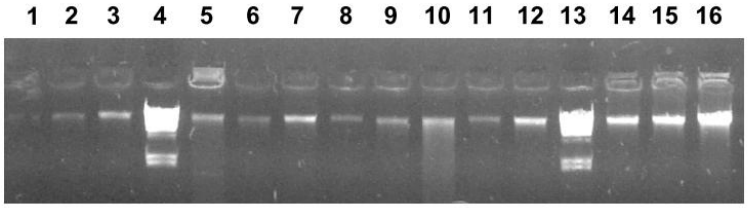
- GSC
- CULTURE COLLECTION CENTERS -> members of WFCC
  - DSMZ [Hans-Peter Klenk]
- ORGANISM SELECTION AND PRIORITIZATION
  - BERGEY's, NAMES4LIFE [George Garrity, Barny Whitman]
- REPRESENTATIVES FROM GRAND CHALLENGE PROJECTS
  - GEBA [Phil Hugenholtz, Jonathan Eisen]
  - TERRAGENOME [Janet Jansson, Jim Tiedje]
  - HMP [George Weinstock, Karen Nelson, Julian Parkhill]
- CORE PARTICIPANTS
  - Large Sequencing Centers [JGI, BGI, JCVI, Sanger, Broad, WashU, Baylor]
  - Annotation/Analysis standards [Owen White, Folker Meyer, & GSC]

FUNDING

What the MEP needs from Culture Collections

Quality controlled high molecular DNAs

Quantification gel of the genomic DNA isolated from *Veillonella parvula* (DSM 2008T)

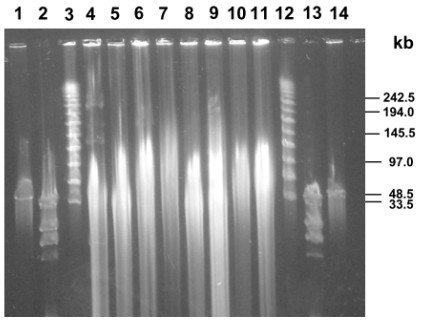


Lane 1: c(λ-Marker)= 15 ng
Lane 2: c(λ-Marker)= 30 ng
Lane 3: c(λ-Marker)= 50 ng
Lane 4: DNA Molecular Weight Marker II (Roche 236250)
Lane 5: DSM 17242, *Alistipes finegoldii;* 5 µl, 1:50
Lane 6: DSM 20460, *Megasphaera elsdenii;* 5 µl, 1:50
Lane 7: DSM 5511, *Haloterrigena turkmenica;* 5 µl, 1:50
Lane 8: DSM 2008, *Veillonella parvula;* 5 µl, 1:50
Lane 9: DSM 20758, *Selenomonas sputigena;* 5 µl, 1:50
Lane 10: DSM 12940, *Halorhabdus utahensis;* 5 µl, 1:50
Lane 11: DSM 1135, *Leptotrichia buccalis;* 5 µl, 1:50
Lane 12: DSM 12286, *Halomicrobium mukohataei;* 5 µl, 1:50
Lane 13: DNA Molecular Weight Marker II (Roche 236250)
Lane 14: c(λ-Marker)= 125 ng
Lane 15: c(λ-Marker)= 250 ng
Lane 16: c(λ-Marker)= 500 ng

Microorganisms

8T) was taken from the German Collection of Microorganisms and Cell Cultures (DSMZ). The
ng the Qiagen Genomic 500 DNA Kit (Qiagen 10262). The genomic DNA was 20-150 kb in size as
el Electrophoresis (PFGE). The bulk of DNA had a size of 50-100 kb (see attached PFGE image).
ng/µl as estimated from the gel. Spectrophotometric measurements yielded a DNA concentration
c DNA are shipped (87,5µg).

Lengths of Genomic DNA Determined by Pulsed Field Gel Electrophoresis (PFGE)



Lane 1: T7 DNA Mass Marker (60 ng)
Lane 2: DNA Molecular Weight Marker II
Lane 3: Mid Range Marker II
Lane 4: DSM 17242, *Alistipes finegoldii*
Lane 5: DSM 20460, *Megasphaera elsdenii*
Lane 6: DSM 5511, *Haloterrigena turkmenica*
Lane 7: DSM 2008, *Veillonella parvula*
Lane 8: DSM 20758, *Selenomonas sputigena*
Lane 9: DSM 12940, *Halorhabdus utahensis*
Lane 10: DSM 1135, *Leptotrichia buccalis*
Lane 11: DSM 12286, *Halomicrobium mukohataei*
Lane 12: Mid Range Marker II
Lane 13: DNA Molecular Weight Marker II
Lane 14: T7 DNA Mass Marker (60 ng)

PFGE of the genomic DNA of the strains was performed in a contour-clamped homogeneous electric field (CHEF) system on a CHEF-DR III device (Bio-Rad Laboratories, Hercules, Calif.) with 1% agarose gels and modified 0.5 TBE buffer (45 mM Tris, 45 mM boric acid, 0.1 mM EDTA) at 14°C. PFGE times used at 200 V (6 V/cm) were 1 to 15 s for 18 h.

What the MEP needs from Culture Collections

Metadata of type strains

Table 1. Classification and general features of *A. ferrooxidans* ICP^T based on MIGS recommendations [11]

MIGS ID	Property	Term	Evidence code[a,b]
	Current classification	Domain *Bacteria*	
		Phylum *Actinobacteria*	TAS [12]
		Class *Actinobacteria*	TAS [13]
		Order *Acidimicrobiales*	TAS [13]
		Suborder *Acidimicrobineae*	
		Family *Acidomicrobiaceae*	TAS [13]
		Genus *Acidimicrobium*	TAS [1]
		Species *Acidimicrobium ferrooxidans*	TAS [1]
		Type strain ICP	
	Gram stain	positive	TAS [1]
	Cell shape	rod shaped	TAS [1]
	Motility	motile	TAS [1]
	Sporulation	nonsporulating	TAS [1]
	Temperature range	moderate thermophile, 45-50°C	TAS [1]
	Optimum temperature	48°C	TAS [1]
	Salinity	not reported	
MIGS-22	Oxygen requirement	aerobic	TAS [1]
	Carbon source	CO_2 (autotrophic), yeast extract (heterotrophic)	TAS [1]
	Energy source	autotrophic: oxidation of ferrous iron with oxygen as the electron acceptor; heterotrophic: yeast extract	TAS [1]
MIGS-6	Habitat	warm, acidic, iron-, sulfur-, or mineral-sulfide rich environments	TAS [1]
MIGS-15	Biotic relationship	free living	NAS
MIGS-14	Pathogenicity	none	NAS
	Biosafety level	1	TAS [14]
	Isolation	hot springs	TAS [2]
MIGS-4	Geographic location	Krísuvik geothermal area, Iceland	TAS [2]
MIGS-5	Sample collection time	before 1993	TAS [1]
MIGS-4.1 MIGS-4.2	Latitude – Longitude	63.93, -22.1	TAS [2]
MIGS-4.3	Depth	not reported	
MIGS-4.4	Altitude	not reported	

Evidence codes - IDA: Inferred from Direct Assay (first time in publication); TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project [15]. If the evidence code is IDA the property was directly observed for a live isolate by one of the authors or an expert mentioned in the acknowledgements.

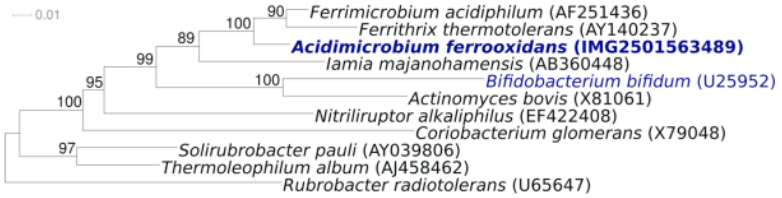
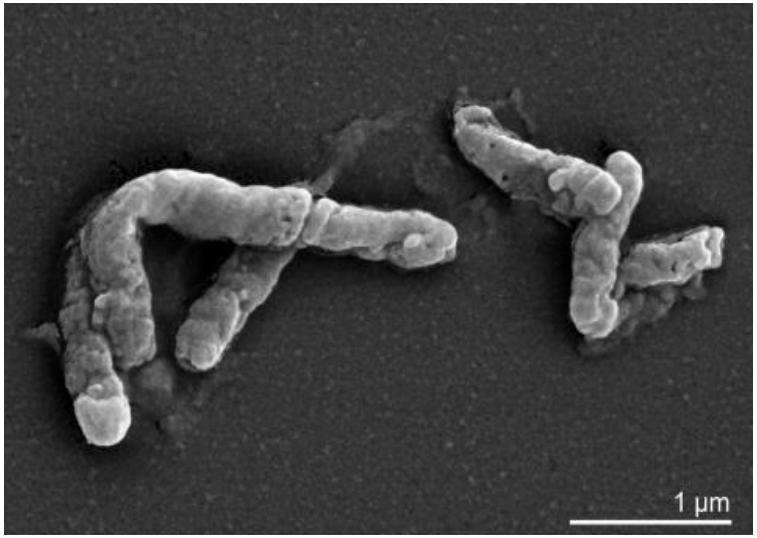


Figure 1. Phylogenetic tree highlighting the position of *A. ferrooxidans* strain ICP^T relative to all other type strains within the *Acidimicrobiales* and the type strains of all other orders within the *Actinobacteria*. The tree was inferred from 1306 aligned characters [7, 8] of the 16S rRNA gene under the maximum likelihood criterion [9] and rooted with *Rubrobacteriales*. The branches are scaled in terms of the expected number of substitutions per site. Numbers above branches are support values from 1000 bootstrap replicates if larger than 60%. Lineages with type strain genome sequencing projects registered in GOLD [10] are shown in blue, published genomes in bold.

Chemotaxonomy

The murein of *A. ferrooxidans* ICP^T contains meso-DAP, like all other characterized type species from the *Acidomicrobineae* [3, 4]. It differs from the other characterized *Acidomicrobineae* strains in MK-9(H_8) being the predominant menaquinone, whereas *F. acidiphilum* has MK-8(H_{10}) as the predominant menaquinone [4], and *I. majanohamensis* possesses a mixture MK-9(H_6), MK-9(H_4), and MK-9(H_8) [3]. The major cellular fatty acids of strain ICP^T are saturated branched acids: iso- (i-) $C_{16:0}$ (83%) and anteiso- (ai-) $C_{17:0}$ (8%) [3], which is more similar to *F. thermotolerans* (90% i-$C_{16:0}$) and *F. acidiphilum* (64% i-$C_{16:0}$ and 11% i-$C_{14:0}$) [4], than to *I. majanohamensis* which predominantly possesses straight chain acids ($C_{17:0}$, $C_{16:0}$ and $C_{15:0}$) [3].



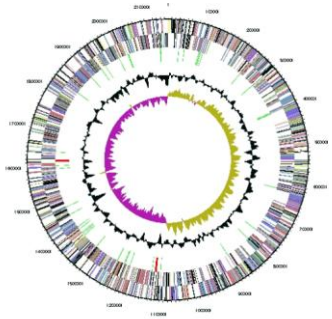
What the MEP gives back to Culture Collections

Genomic data of type strains

Actinosynnema mirum type strain (101T)

Table 4. Number of genes associated with the 21 general COG functional categories

Code	Value	%	Description
J	182	2.6	Translation, ribosomal structure and biogenesis
A	2	0.0	RNA processing and modification
K	607	8.5	Transcription
L	173	2.4	Replication, recombination and repair
B	2	0.0	Chromatin structure and dynamics
D	34	0.5	Cell cycle control, mitosis and meiosis
Y	0	0.0	Nuclear structure
V	96	1.4	Defense mechanisms
T	389	5.5	Signal transduction mechanisms
M	210	3.0	Cell wall/membrane biogenesis
N	45	0.6	Cell motility
Z	1	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	46	0.6	Intracellular trafficking and secretion
O	149	2.1	Posttranslational modification, protein turnover, chaperones
C	306	4.3	Energy production and conversion
G	441	6.2	Carbohydrate transport and metabolism
E	425	6.0	Amino acid transport and metabolism
F	108	1.5	Nucleotide transport and metabolism
H	223	3.1	Coenzyme transport and metabolism
I	226	3.2	Lipid transport and metabolism
P	241	3.4	Inorganic ion transport and metabolism
Q	265	3.7	Secondary metabolites biosynthesis, transport and catabolism
R	670	9.4	General function prediction only
S	328	4.6	Function unknown
-	2613	36.8	Not in COGs



, Graphical circular map of the genome. From outside to the center: Genes on forward olor by COG categories), Genes on reverse strand (color by COG categories), RNA genes reen, rRNAs red, other RNAs black), GC content, GC skew.

Table 3. Genome Statistics

Attribute	Value	% of Total
Genome size (bp)	2,158,157	100.00%
DNA Coding region (bp)	1,988,736	92.15%
DNA G+C content (bp)	1,473,791	68.29%
Number of replicons	1	
Extrachromosomal elements	0	
Total genes	2092	100.00%
RNA genes	54	2.58%
rRNA operons	2	
Protein-coding genes	2038	97.42%
Pseudo genes	74	3.54%
Genes with function prediction	1584	75.72%
Genes in paralog clusters	1969	9.37%
Genes assigned to COGs	1526	72.94%
Genes assigned Pfam domains	1603	76.63%
Genes with signal peptides	591	28.25%
Genes with transmembrane helices	436	20.84%
CRISPR repeats	2	

Figure 4. Schematic cellular overview of all pathways of the *B. faecium* strain Schefferle 6-10T metabolism. Nodes represent metabolites, with shape indicating class of metabolite Lines represent reactions.

→Genomic data of type strains

Basis for metabolic reconstruction
phenotyping improved culture conditions

The GSC Biocomputing Consortium

ANL JGI CAMERA U. Maryland IGS open

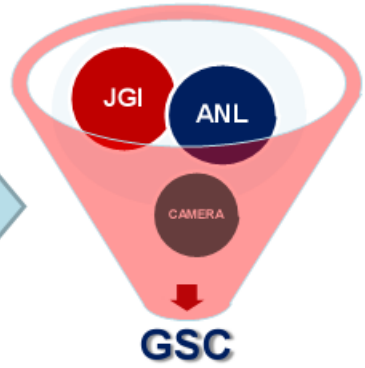
m5

innovation through collaboration

M5: Metagenomics, Metadata, MetaAnalysis, Multiscale-models and ***MetaInfrastructure***

ANL JGI

- develop standards to share data more effectively
- run data-intensive workflows on cloud based systems
- allow annotation from multiple sources to be combined
- reduce wasted cycles associated with data refreshes.



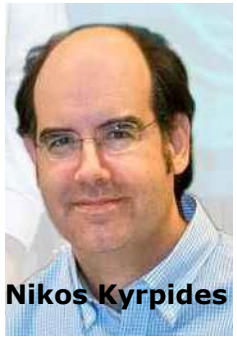
DSMZ GEBA Team
Braunschweig

Birte **Abt**
Evelyne **Brambilla**
Markus **Göker**
Sabine **Gronow**
Elke **Lang**
Rüdiger **Pukall**
Johannes **Sikorski**
Stefan **Spring**
Brian **Tindall**

JGI GEBA Team
Walnut Creek, CA

Jim **Bristow**
Jan-Fang **Chen**
Jonathan A. **Eisen**
Lynne **Goodwin**
Philip **Hugenholtz**
Nikos C. **Kyrpides**
Alla **Lapidus**
Victor **Markowitz**
Tanja **Woyke**


Jonathan Eisen


Nikos Kyrpides


Philip Hugenholtz