Universität Bielefeld Centrum für Biotechnologie

CeBiTec Center for Biotechnology

Analyzing Metagenome Data Obtained by High-Throughput Sequencing

A. Pühler

Center for Biotechnology

Bielefeld University

International Conference: Getting Post 2010 Biodiversity Targets Right

Bragança Paulista/SP, Brazil

December 11th – 15th, 2010

Content of Talk

- Sequence analysis of the metagenome of a model microbial community
- Analysis of assembled contigs and single reads by the help of completely sequenced genomes
- The functional and taxonomic analysis of single reads using the software programs MetaSAMS and CARMA
- The taxonomic analysis of a model microbial community based on 16S-rDNA sequences

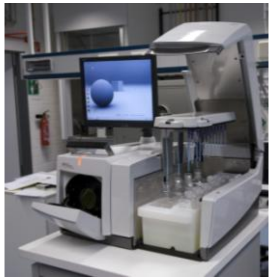
Sequence Analysis of the Metagenome of a Model Microbial Community (Part I)

- Sequencing devices at the CeBiTec of Bielefeld University
- Introduction of the model microbial community residing in an agrigultural biogas production
- Sequence analysis of the metagenome of the model microbial community

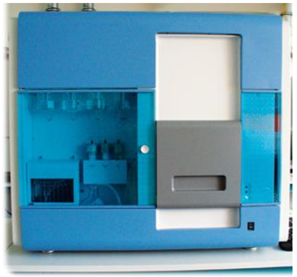
High-Throughput Sequencing Devices at the CeBiTec of Bielefeld University

Sequencing techniques

ABI 3730xl DNA Analyzer (Applied Biosystems)



Genome Sequencer GS FLX (Roche)

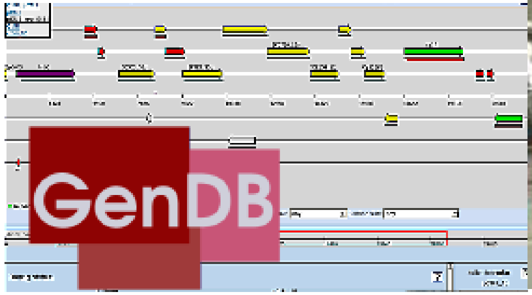


Genome Analyzer (Illumina, Inc.)

Genomics Platform

high-throughput sequencing

Bioinformatics expertise and environment



Bioinformatics Platform

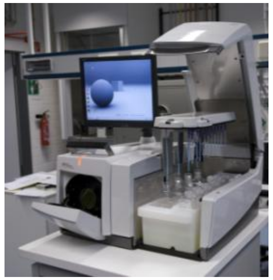
professional data evaluation

CeBiTec Center for Biotechnology

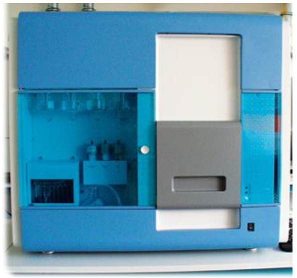
Comparison of Different Sequencing Technologies

Sequencing techniques

ABI 3730xl DNA Analyzer (Applied Biosystems)



Genome Sequencer GS FLX (Roche)



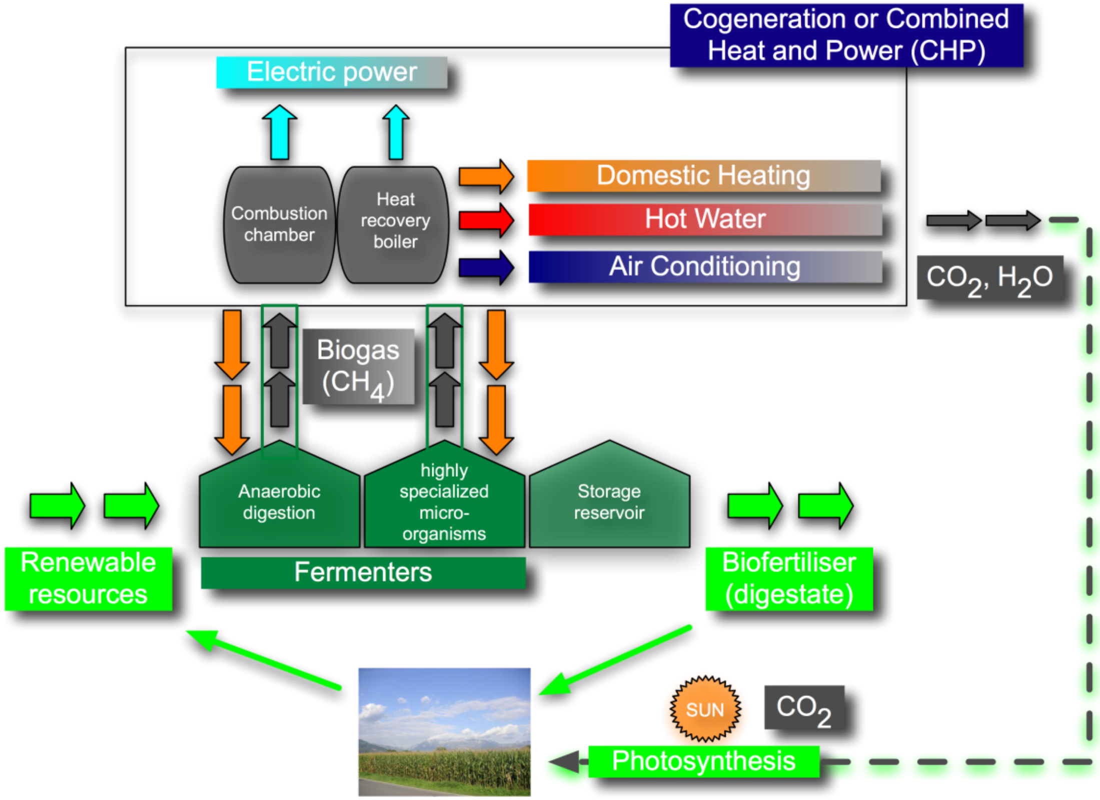
Genome Analyzer (Illumina, Inc.)

read length:	**1100 bp**	**400 bp**	**150 bp**
sequenced bases/run:	**0,1 Mb**	**500 Mb**	**45 Gb**

The GS FLX system is evidently best suited for a metagenome analysis since it offers long read length combined with an acceptable output.

Metagenome Analysis of a Model Microbial Community Residing in a Biogas Production Plant Using Ultrafast Sequencing

Biogas production from primary renewable products

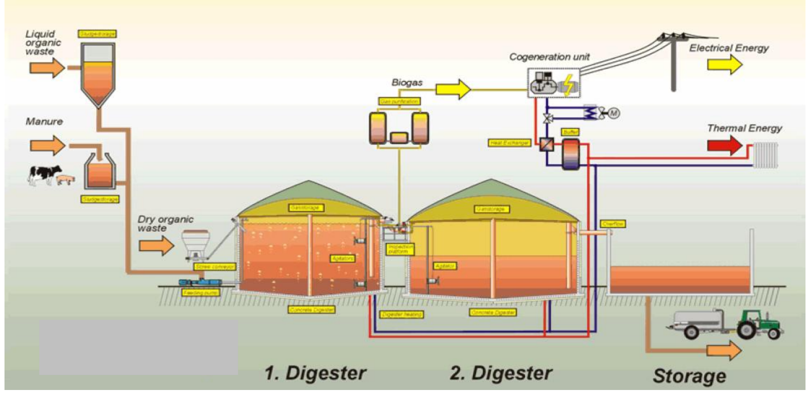


Biogas is produced during anaerobic digestion of biomass by specific microbial consortia

Characteristics of the Analyzed Biogas Plant Located Close to the City of Bielefeld

- 500 kW installed electric power
- 3 reactors (mesophilic conditions)
- 1. Fermenter (1500 m^3)
- 2. Fermenter (1700 m^3)
- 3. Storage reactor (3600 m^3)
- Substrates: Renewable primary products
  (liquid manure, maize silage, green-rye, pig and poultry manure)
- Continuous fermentation (retention period 40 – 60 days)

Biogas plant consisting of three fermenters

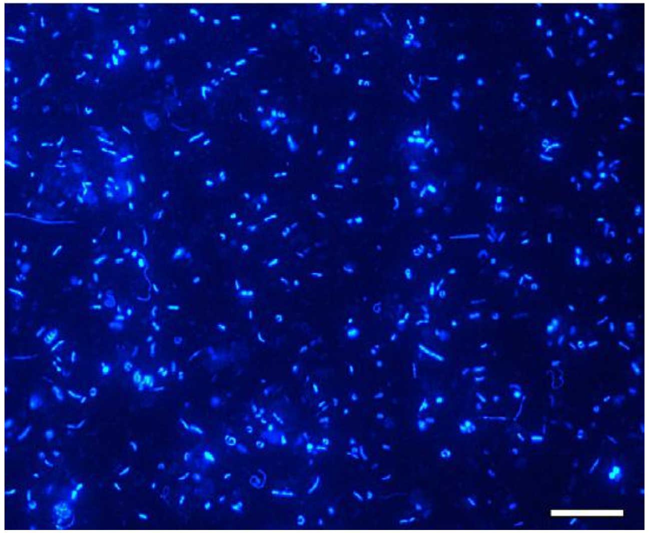


Schematic view of the biogas plant

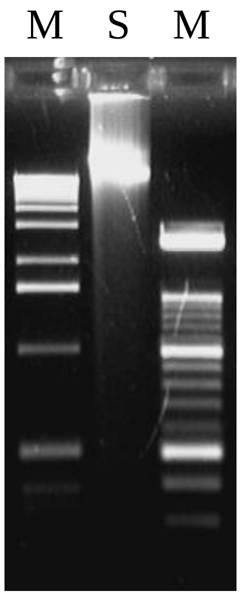
CeBiTec
Center for Biotechnology

Isolation and Sequencing of Total Community DNA Isolated From the Model Microbial Community

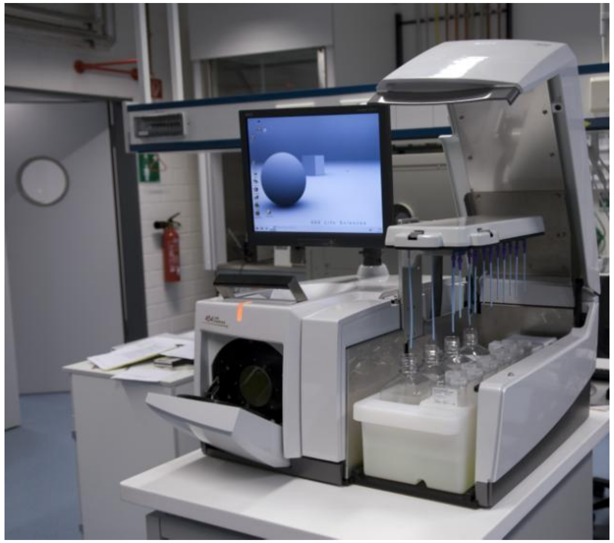
- **High molecular weight and pure total community DNA was prepared from the fermentation sample taken from the biogas plant (CTAB-based method).**



Biogas-producing Microbial Community



Total Community DNA



Genome Sequencer FLX

Analysis of assembled Contigs and Single Reads by the Help of Completely Sequenced Microbial Genomes (Part II)

- **Sequence analysis of total DNA**
- **Mapping of assembled contig reads to completely sequenced microbial genomes**
- **Mapping of metagenome sequence reads to the *Methanoculleus marisnigri* JR1 genome**
- **Coverage of the *M. marisnigri* methanogenesis gene region by metagenome sequence reads**

Sequence Analysis of Total Community DNA and Assembly of Sequence Reads

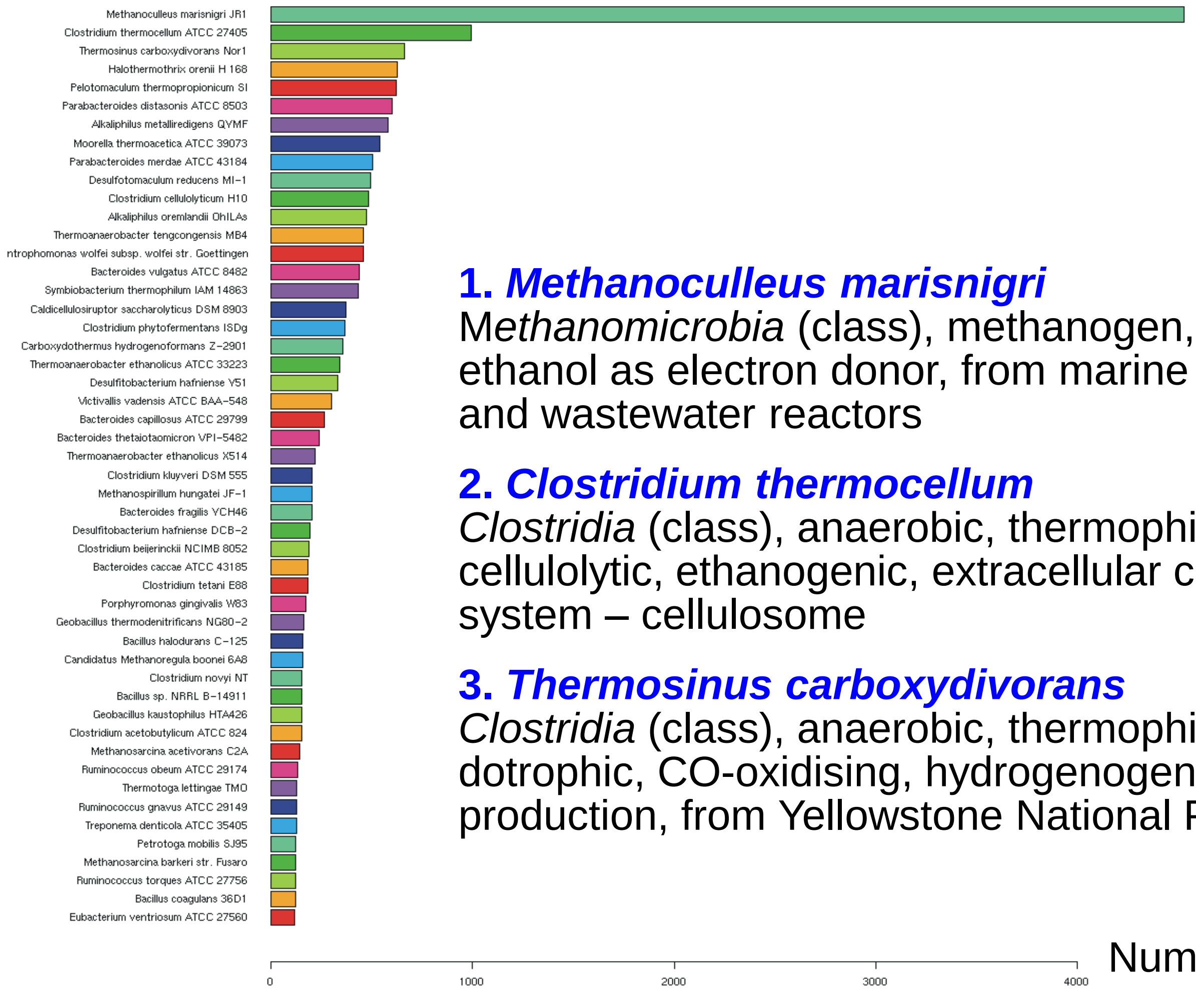
- **Total community DNA was sequenced with the Genome Sequencer FLX**

System	GS FLX	GS FLX Titanium	Factor
Number of reads	616,072	1,347,644	2.2
Number of bases	141,685,079 bases	495,506,659 bases	3.5
Average read length	230 bases	368 bases	1.6

- **Individual reads of the GS FLX run were assembled using the Newbler Assembler**

System	GS FLX	GS FLX Titanium	Factor
Number of contigs	8,752	37,645	4.3
Number of bases in contigs	11,797,906 bases	45,874,670 bases	3.9
Average contig size	1,348 bases	1,380 bases	1.0

Mapping of Assembled Contig Sequences to Completely Sequenced Microbial Genomes



1. *Methanoculleus marisnigri*
M*ethanomicrobia* (class), methanogen, use of ethanol as electron donor, from marine sediments and wastewater reactors

2. *Clostridium thermocellum*
Clostridia (class), anaerobic, thermophilic, cellulolytic, ethanogenic, extracellular cellulase system – cellulosome

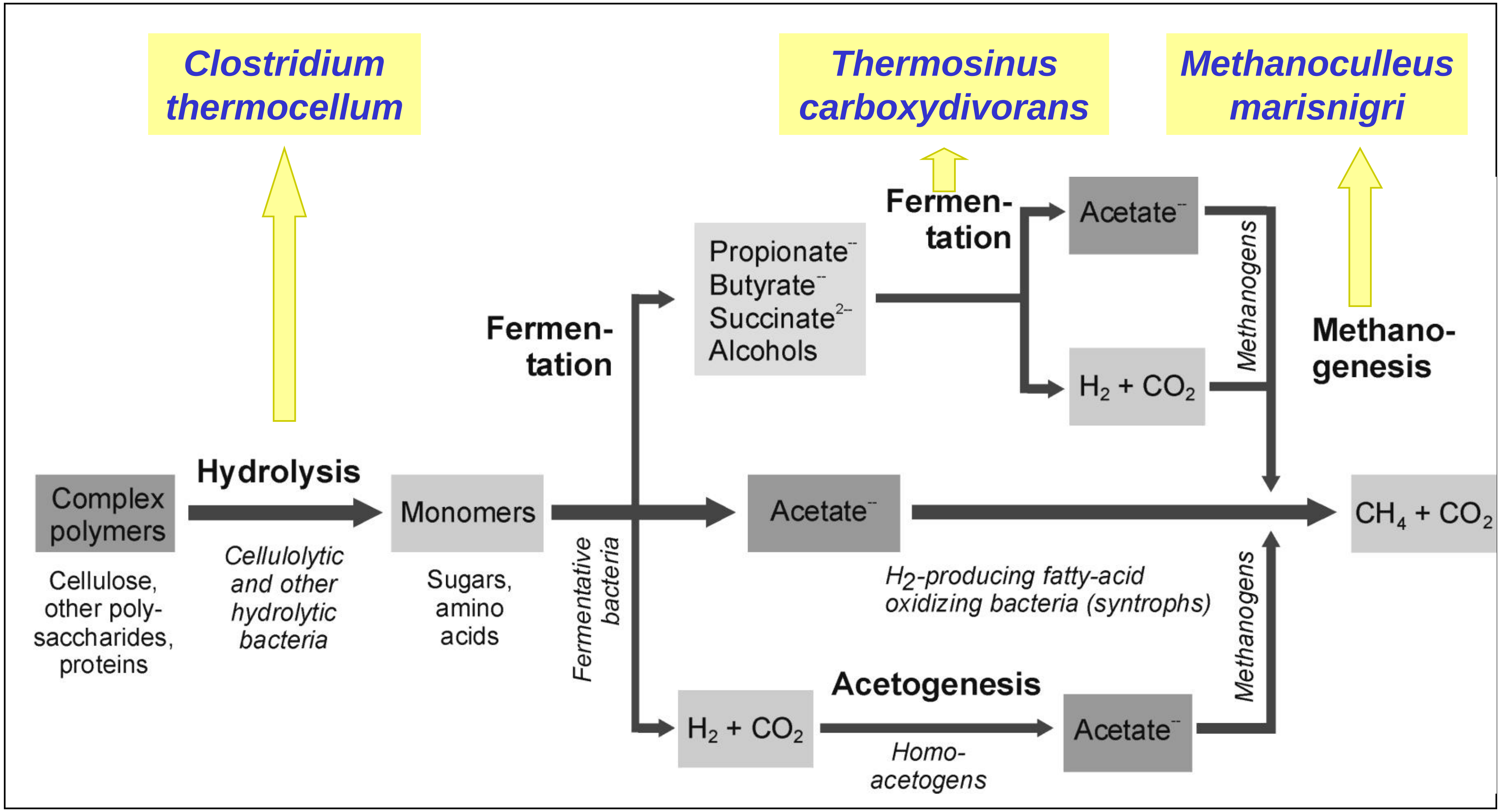
3. *Thermosinus carboxydivorans*
Clostridia (class), anaerobic, thermophilic, carboxy-dotrophic, CO-oxidising, hydrogenogenic, acetate production, from Yellowstone National Park

Number of contig matches

Reference: Schlüter et al., J.Biotechnology 136: 77-90 (2008)

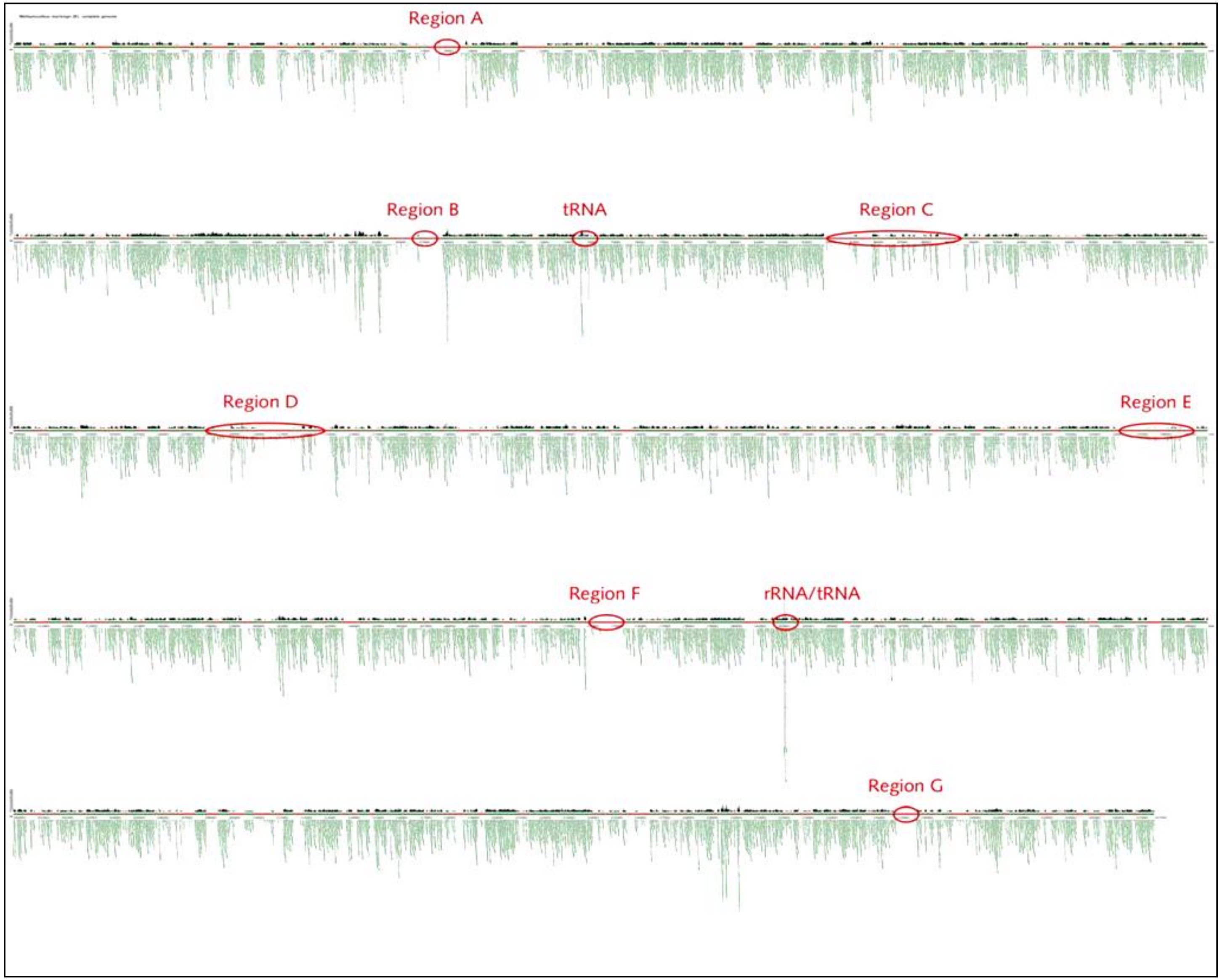
CeBiTec
Center for Biotechnology

Biochemical Processes Taking Place in a Biogas Fermenter



Anoxic decomposition. Shown is the overall process of anoxic decomposition, in which various groups of fermentative anaerobes cooperate in the conversion of complex organic materials ultimately to methane (CH_4) and CO_2.

Mapping of Metagenome Sequence Reads to the *Methanoculleus marisnigri* JR1 Genome



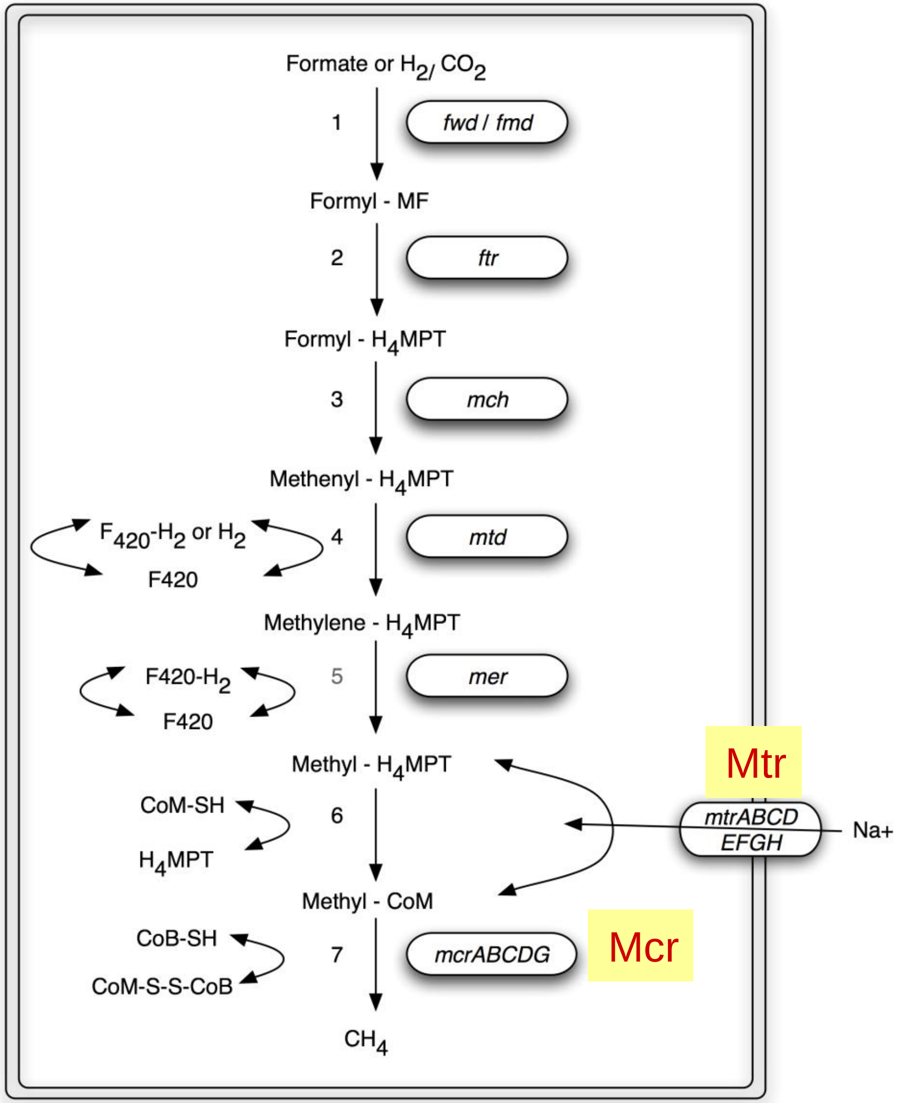
Metagenome sequence reads obtained from the biogas fermentation sample were aligned to the complete *M. marisnigri* genome sequence. Length of vertical bars indicates the local coverage at a given genome position. *M. marisnigri* JR1 genome subregions covered by metagenome reads are coloured in green; non-covered region are visualised in red. Aligned reads are highlighted as green bars in the lower part of the plot.

Coverage of the Methanoculleus marisnigri Reference Genome by Metagenome Reads

Data set	Coverage [%]
GS FLX	39.8
Titanium	41.7
combined	45.4

- Approx., 45.4% of the *M. marisnigri* genome are coveredby metagenome reads.

Methanogenesis Pathway Using CO_2 and H_2



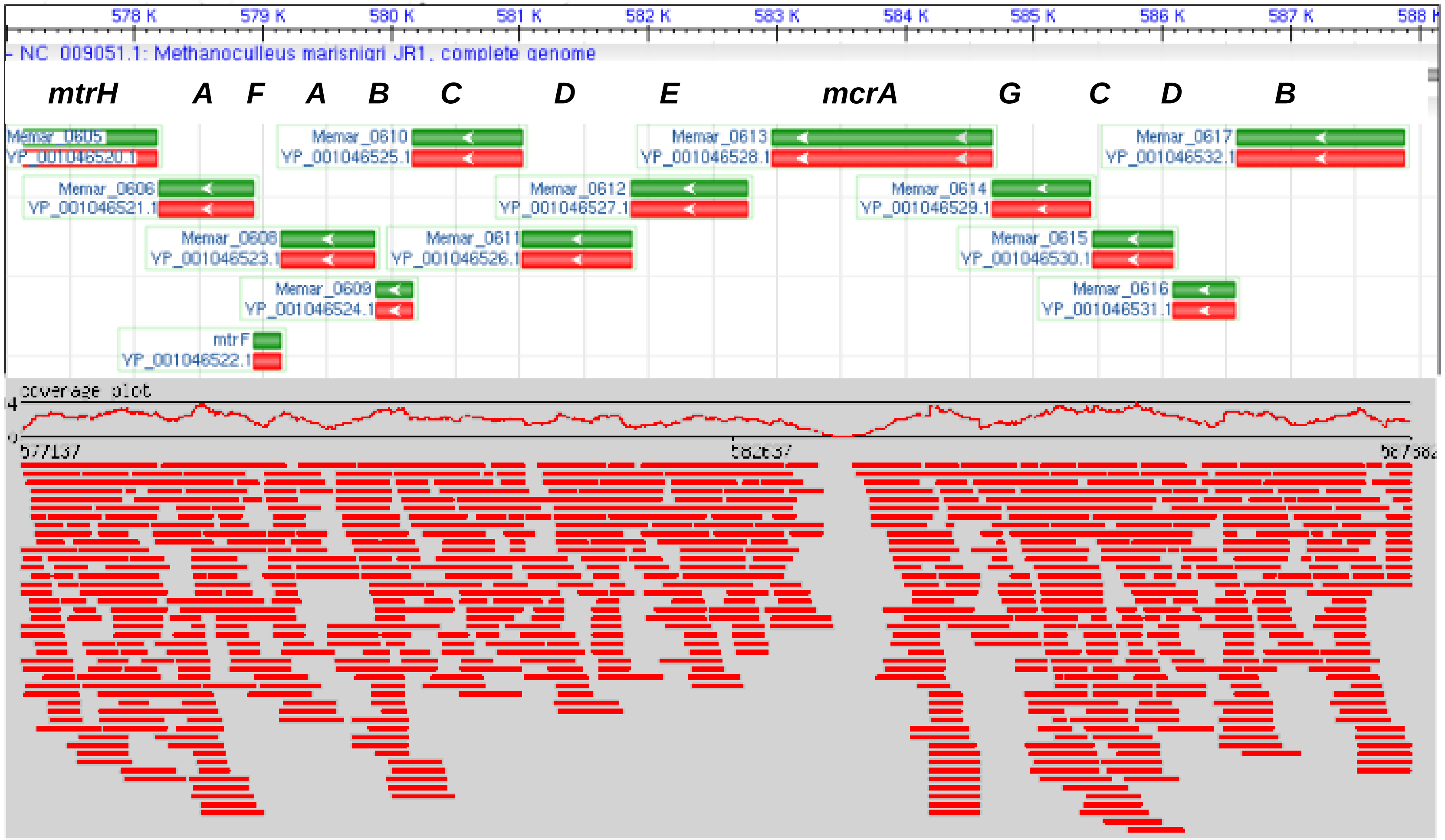
MF – Methanofuran, H_4MPT - Tetrahydromethanopterin

Hydrogen dependent pathway of CO_2 reduction to methane (CH_4):

- Transfer of the C1 unit to methanofuran (MF), (step 1).
- Transfer of the formyl group to tetrahydromethanopterin (H_4MPT), (step 2).
- Finally, reduction of the methyl group bound to coenzyme M to CH_4 (step 7) catalysed by methyl-coenzyme M reductase (Mcr).

S. Shima *et al*. (2002), J. Bioscience and Bioengineering

Coverage of the Central *M. marisnigri* Methanogenesis Gene Region by Metagenome Reads



mtrHAFABCDE - tetrahydromethanopterin S-methyltransferase
mcrAGCDB - methyl-coenzyme M reductase

The Functional and Taxonomic Analysis of Single Reads Using the Software Programs MetaSAMS and CARMA (Part III)

- Introduction to the software programs MetaSAMS and CARMA
- The taxonomic profile of the model microbial community established with the CARMA program

Relevant questions:

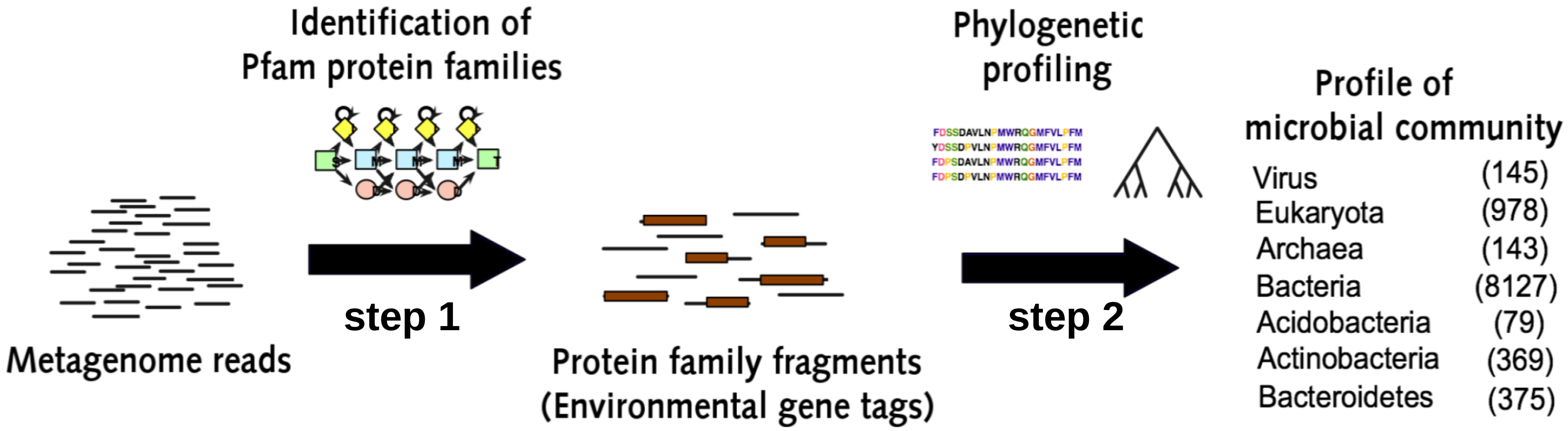
- Is it possible to use short sequencing reads for the determination of gene functions?
- Is it possible to establish the taxonomic composition of the microbial community from short sequencing reads?

Introduction to the Software Programs MetaSAMS and CARMA

- **MetaSAMS is a relational database which enables the processing of single metagenome sequence reads.**

- **CARMA is a program which enables the functional and taxonomic analysis of single sequence reads.**
  - **In a first step, the functional analysis of single sequence reads is carried out by asigning the reads to Pfam protein families.**
  - **In a second step, sequence reads are analyzed. The best filling member within a Pfam protein family is used as a taxonomic marker.**

The Environmental Gene Tag (EGT) Analysis by Means of CARMA



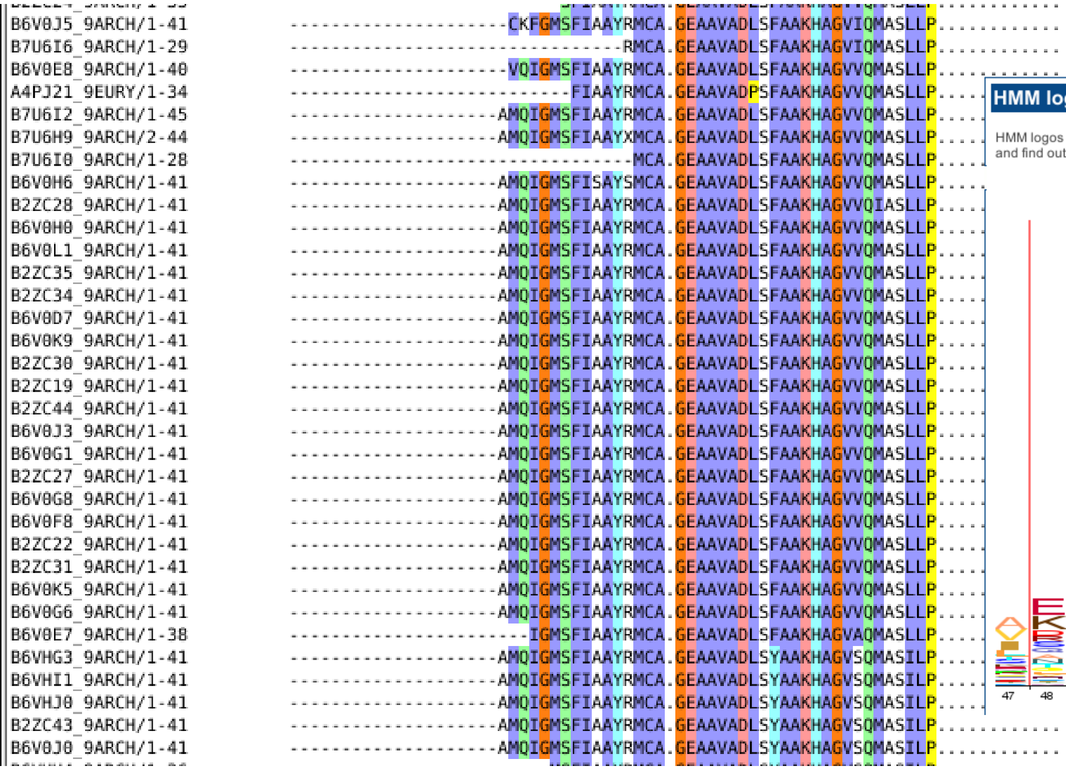
Virus	(145)
Eukaryota	(978)
Archaea	(143)
Bacteria	(8127)
Acidobacteria	(79)
Actinobacteria	(369)
Bacteroidetes	(375)

- CARMA enables the functional and taxonomic characterization of single metagenome reads.
- Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families.

L. Krause *et al.* (2008), J. Biotechnol.; W. Gerlach *et al*. (2009), BMC Bioinformatics

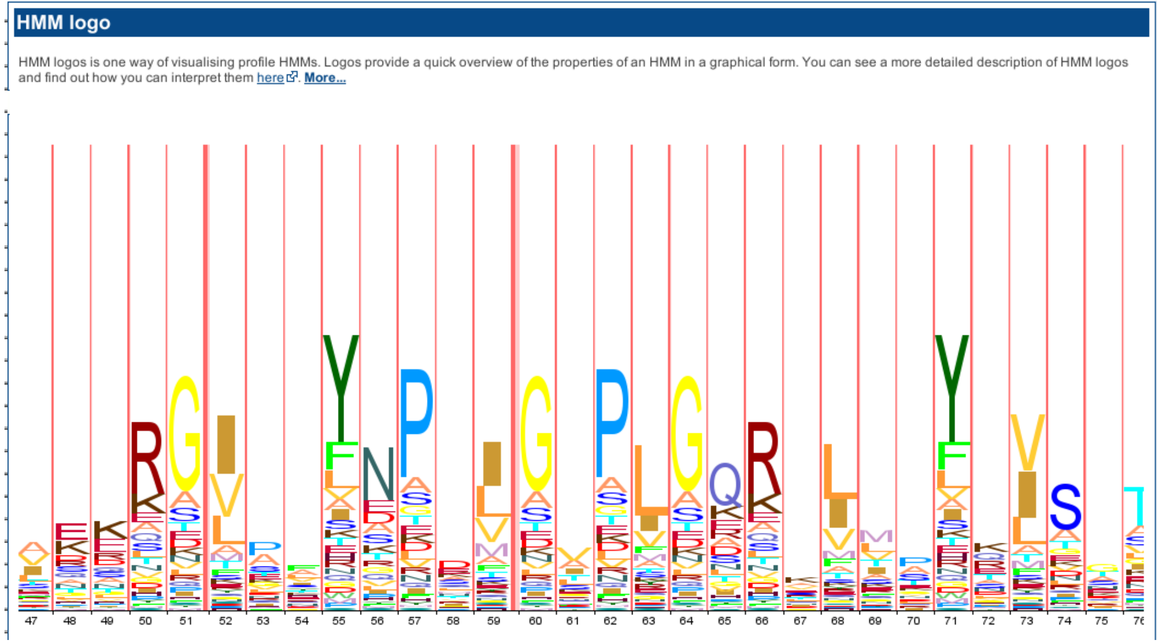
The Pfam Entry for Methyl coenzyme M reductase A (McrA), the key enzyme for methane formation

Pfam alignment for McrA (PF02745)



```
B6V0J5_9ARCH/1-41      ---------------------CKFGMSFIAAYRMCA.GEAAVADLSFAAKHAGVIQMASLLP............
B7U6I6_9ARCH/1-29      --------------------------------RMCA.GEAAVADLSFAAKHAGVIQMASLLP............
B6V0E8_9ARCH/1-40      ---------------------VQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP............
A4PJ21_9EURY/1-34      ---------------------------FIAAYRMCA.GEAAVADPSFAAKHAGVVQMASLLP....
B7U6I2_9ARCH/1-45      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B7U6H9_9ARCH/2-44      --------------------AMQIGMSFIAAYXMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B7U6I0_9ARCH/1-28      ---------------------------------MCA.GEAAVADLSFAAKHAGVVQMASLLP....
B6V0H6_9ARCH/1-41      --------------------AMQIGMSFISAYSMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B2ZC28_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQIASLLP....
B6V0H0_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B6V0L1_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B2ZC35_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B2ZC34_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B6V0D7_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B6V0K9_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B2ZC30_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B2ZC19_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B2ZC44_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B6V0J3_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B6V0G1_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B2ZC27_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B6V0G8_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B6V0F8_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B2ZC22_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B2ZC31_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B6V0K5_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B6V0G6_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSFAAKHAGVVQMASLLP....
B6V0E7_9ARCH/1-38      -----------------------IGMSFIAAYRMCA.GEAAVADLSFAAKHAGVAQMASLLP....
B6VHG3_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSYAAKHAGVSQMASILP....
B6VHI1_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSYAAKHAGVSQMASILP....
B6VHJ0_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSYAAKHAGVSQMASILP....
B2ZC43_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSYAAKHAGVSQMASILP............
B6V0J0_9ARCH/1-41      --------------------AMQIGMSFIAAYRMCA.GEAAVADLSYAAKHAGVSQMASILP............
```

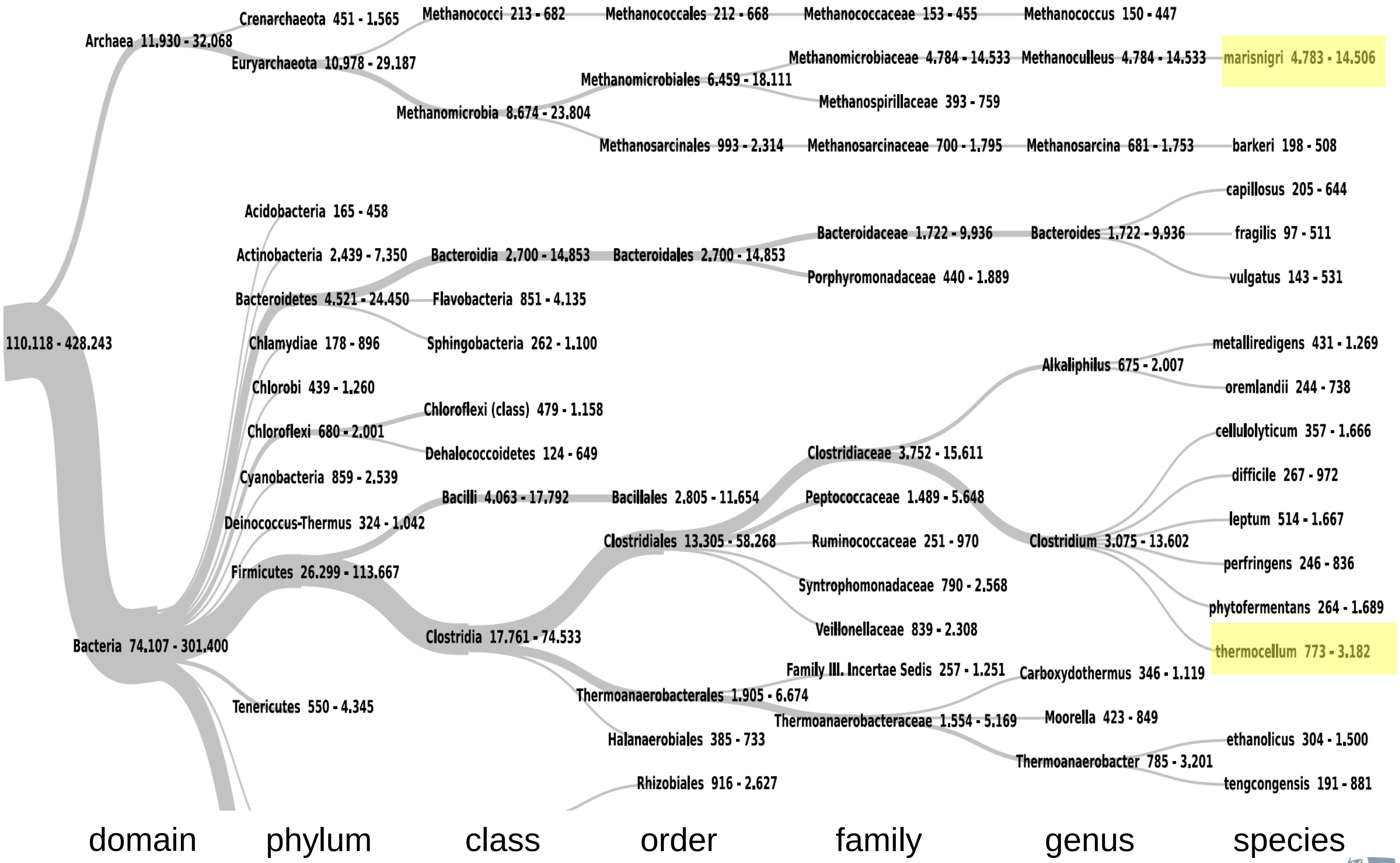
Pfam HMM (hidden Markov model) logo for McrA (PF02745)



HMM logo

HMM logos is one way of visualising profile HMMs. Logos provide a quick overview of the properties of an HMM in a graphical form. You can see a more detailed description of HMM logos and find out how you can interpret them here. More...

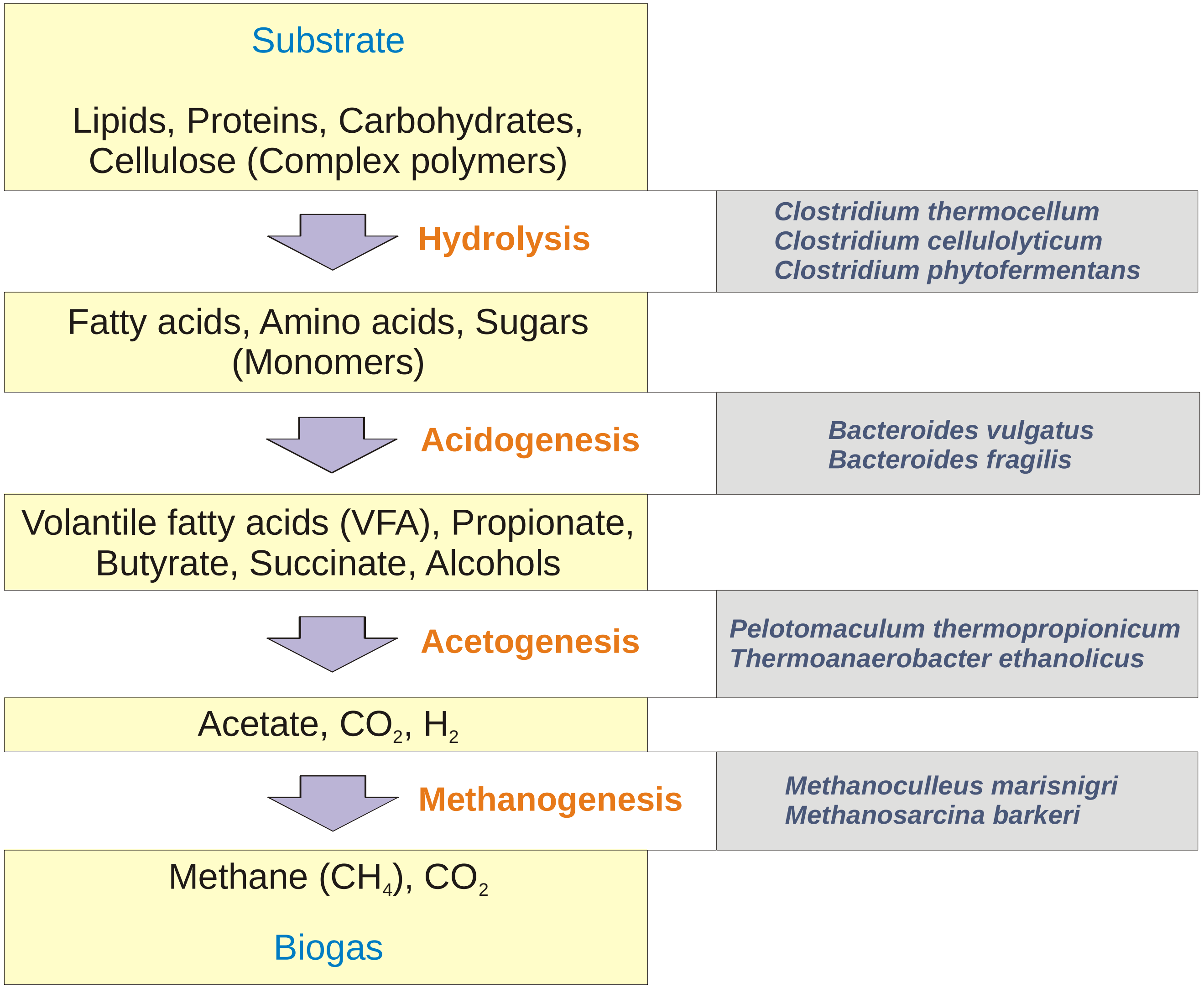
Taxonomic Profile Based on CARMA



110.118 - 428.243
Archaea 11.930 - 32.068
Crenarchaeota 451 - 1.565
Euryarchaeota 10.978 - 29.187
Methanococci 213 - 682
Methanococcales 212 - 668
Methanococcaceae 153 - 455
Methanococcus 150 - 447
Methanomicrobia 8.674 - 23.804
Methanomicrobiales 6.459 - 18.111
Methanomicrobiaceae 4.784 - 14.533
Methanoculleus 4.784 - 14.533
marisnigri 4.783 - 14.506
Methanospirillaceae 393 - 759
Methanosarcinales 993 - 2.314
Methanosarcinaceae 700 - 1.795
Methanosarcina 681 - 1.753
barkeri 198 - 508
Bacteria 74.107 - 301.400
Acidobacteria 165 - 458
Actinobacteria 2.439 - 7.350
Bacteroidetes 4.521 - 24.450
Bacteroidia 2.700 - 14.853
Bacteroidales 2.700 - 14.853
Bacteroidaceae 1.722 - 9.936
Bacteroides 1.722 - 9.936
capillosus 205 - 644
fragilis 97 - 511
vulgatus 143 - 531
Porphyromonadaceae 440 - 1.889
Flavobacteria 851 - 4.135
Sphingobacteria 262 - 1.100
Chlamydiae 178 - 896
Chlorobi 439 - 1.260
Chloroflexi 680 - 2.001
Chloroflexi (class) 479 - 1.158
Dehalococcoidetes 124 - 649
Cyanobacteria 859 - 2.539
Deinococcus-Thermus 324 - 1.042
Firmicutes 26.299 - 113.667
Bacilli 4.063 - 17.792
Bacillales 2.805 - 11.654
Clostridia 17.761 - 74.533
Clostridiales 13.305 - 58.268
Clostridiaceae 3.752 - 15.611
Alkaliphilus 675 - 2.007
metalliredigens 431 - 1.269
oremlandii 244 - 738
Clostridium 3.075 - 13.602
cellulolyticum 357 - 1.666
difficile 267 - 972
leptum 514 - 1.667
perfringens 246 - 836
phytofermentans 264 - 1.689
thermocellum 773 - 3.182
Peptococcaceae 1.489 - 5.648
Ruminococcaceae 251 - 970
Syntrophomonadaceae 790 - 2.568
Veillonellaceae 839 - 2.308
Thermoanaerobacterales 1.905 - 6.674
Family III. Incertae Sedis 257 - 1.251
Carboxydothermus 346 - 1.119
Thermoanaerobacteraceae 1.554 - 5.169
Moorella 423 - 849
Thermoanaerobacter 785 - 3.201
ethanolicus 304 - 1.500
tengcongensis 191 - 881
Halanaerobiales 385 - 733
Tenericutes 550 - 4.345
Rhizobiales 916 - 2.627
domain
phylum
class
order
family
genus
species

CeBiTec
Center for Biotechnology

Assignment of Species to Process Steps



Substrate

Lipids, Proteins, Carbohydrates, Cellulose (Complex polymers)

Hydrolysis

Clostridium thermocellum
Clostridium cellulolyticum
Clostridium phytofermentans

Fatty acids, Amino acids, Sugars (Monomers)

Acidogenesis

Bacteroides vulgatus
Bacteroides fragilis

Volantile fatty acids (VFA), Propionate, Butyrate, Succinate, Alcohols

Acetogenesis

Pelotomaculum thermopropionicum
Thermoanaerobacter ethanolicus

Acetate, CO_2, H_2

Methanogenesis

Methanoculleus marisnigri
Methanosarcina barkeri

Methane (CH_4), CO_2

Biogas

The Taxonomic Analysis of a Model Microbial Community Based on 16S-rDNA Sequences (Part IV)

- **Taxonomic profiling using 16S-rDNA sequences obtained from metagenome data sets**
- **Taxonomic profiling using 16S-rDNA sequences obtained by amplicon sequencing**

Taxonomic Profiling Using 16S-rDNA Sequences Obtained from Metagenome Sata Sets

- **Extraction of 16S-rDNA sequences from the metagenome dataset**
- **Trimming to 16S-rDNA-specific sequences**
- **Classification of 16S-rDNA sequences by means of the RDP classifier**

System	Number of 16S-rDNA fragments	Average read length	% of all reads
GS FLX	669	159	0.16
Titanium	3,516	243	0.27
Σ	4,185		0.21

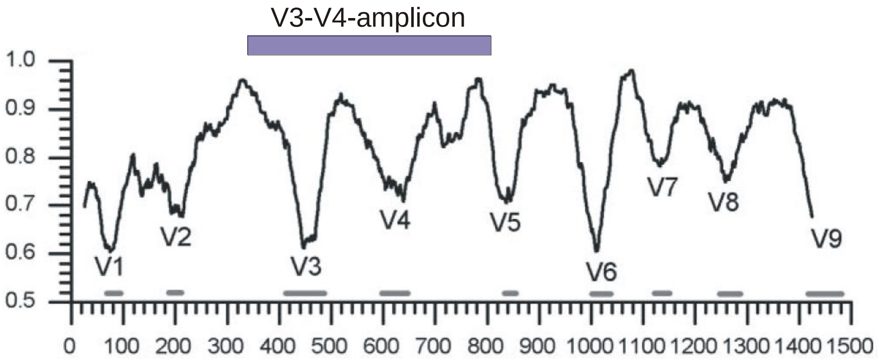
Main result:

The orders *Methanomicrobiales, Bacteroidales* and *Clostridiales* are dominant within the community.

Taxonomic Profiling Using 16S-rDNA Sequences Obtained by Amplicon Sequencing

The amplicon sequencing procedure:

- **Amplification of the 16S-rDNA V3-V4-region (466 bp)**
- **Primers: PRK-341F and PRK-806R specific for prokaryotes (Yu *et al*., 2005)**
- **Uni-directional high-throughput sequencing of 16S-rDNA amplicons on the GS FLX Titanium platform**
- **18,599 sequences (average sequence length 388 bases)**

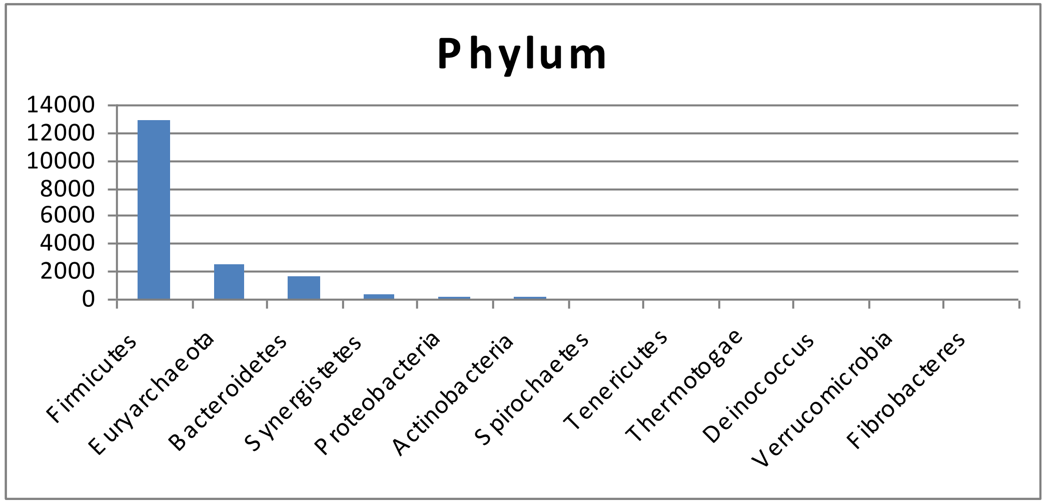


Classification of 16S-rDNA amplicon sequences

Rank	**Assigned sequences** (confidence level > 80%)
Domain*	18,581 (99.90%)
Phylum	17,259 (92.80%)
Class	15,285 (82.18%)
Order	12,074 (64.92%)
Family	6,871 (36.94%)
Genus	6,871 (36.94%)

* *Bacteria*: 87.1%; *Archaea*: 12.9%

Classification of 16S-rDNA amplicon sequences

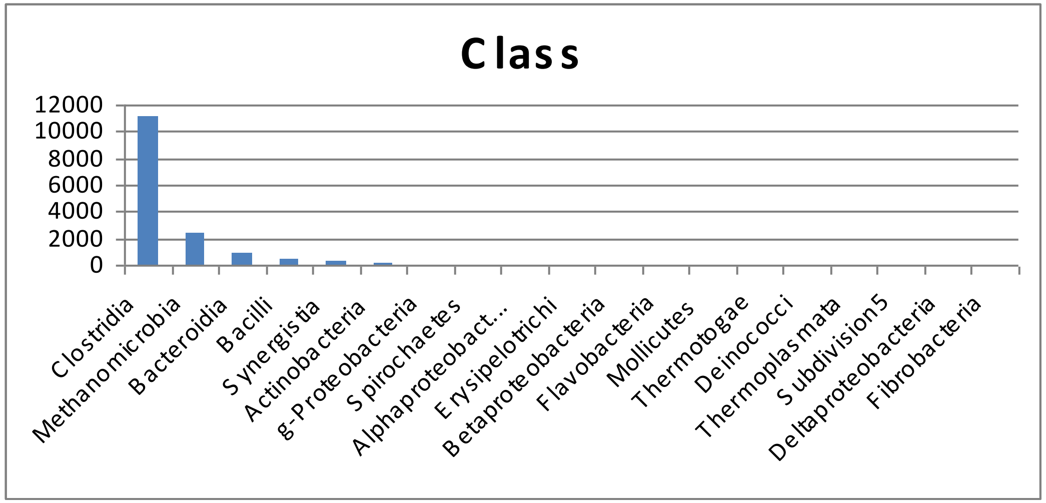


Dominant phyla:

- *Firmicutes*
- *Euryarchaeota*
- *Bacteroidetes*
- *Synergistetes*

Dominant classes:

- *Clostridia*
- *Methanomicrobia*
- *Bacteroidia*
- *Bacilli*
- *Synergistia*



CeBiTec
Center for Biotechnology

Comments to the Dominant Phyla

- ***Firmicutes*** **contain the classes** ***Clostridia*** **and** ***Bacilli*****. They hydrolyze polymers to monomers.**
- ***Bacteroidetes*** **contain the class** ***Bacteroidia*****. They ferment the monomers and produce organic acids.**
- ***Synergistetes*** **contain the class** ***Synergistia*****. They degrade amino acids and produce acetate, butyrate, hydrogen and carbon dioxide.**
- ***Euryarchaeota*** **belong to the domain** ***Archaea*** **and contain the class** ***Methanomicrobia*****, They are resonsible for methane biosynthesis.**

Comparison of Taxonomic Profiles Obtained by Amplicon Sequencing and Metagenome Sequencing

- **Taxonomic profiles based on 16S-rDNA amplicon sequences are biased:**
  - **Choice of PCR primers might "select" for certain sequences**
  - **PCR amplification might have an effect on the abundance of certain 16S-rDNA amplicons**

- **Taxonomic profiles based on the CARMA pipeline do not suffer from such a bias**

Summary

- The model microbial community residing in a biogas fermentation plant was used for a metagenome analysis.
- Microbial metagenomes were efficiently sequenced with the 454 technology.
- Mapping of single reads and contigs on completely sequenced microbial genomes resulted in a first characterization of the metagenome dataset.
- MetaSAMS and CARMA represent a valuable platform for the functional and taxonomic analysis of single reads.
- For a deeper taxonomic analysis, amplicon sequencing is an appropriate tool.

Acknowledgements

J. Kalinowski
Head of the technology platform Genomics

A. Schlüter
Leader of the Metagenomics group

R. Szczepanowski
Leader of the High throughput sequencing group

A. Goesmann
Head of the technology platform Bioinformatics

Universität Bielefeld

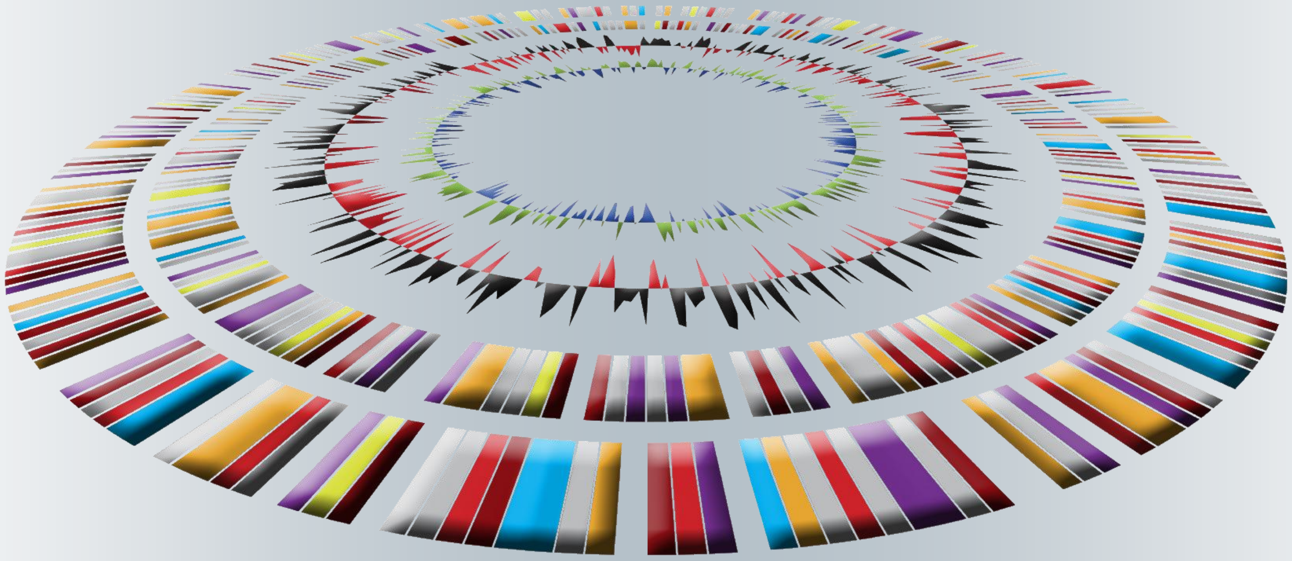
CeBiTec
Center for Biotechnology

GenoMik

Bundesministerium
für Bildung
und Forschung

CENTER FOR BIOTECHNOLOGY
BIELEFELD UNIVERSITY
GERMANY

- MANY THANKS FOR YOUR ATTENTION -



The Genomics and Bioinformatics Platform at Bielefeld University is open for collaboration in Genomics and Metagenomics