Expansion of the Genomic Encyclopedia of Bacteria and Archaea

Christian Rinke[1], Alex Sczyrba[1], Stephanie Malfatti[1],Janey Lee[1], Jan-Fang Cheng[1], Ramunas Stepanauskas[2], Jonathan A. Eisen[1,3], Steven Hallam[4]; William P. Inskeep[5]; Brian P. Hedlund[6]; Stefan M. Sievert[7]; Wen-Tso Liu[8]; George Tsiamis[9]; Philip Hugenholtz[10]; Tanja Woyke[1]

[1] DOE Joint Genome Institute, **[2]**Bigelow Laboratory for Ocean Sciences; **[3]**Department of Evolution and Ecology, University of California Davis, **[4]**Department of Microbiology and Immunology, University of British Columbia; **[5]**Department of Land Resources and Environmental Sciences, Montana State University, **[6]**School of Life Sciences, University of Nevada, Las Vegas; **[7]**Biology Department, Woods Hole Oceanographic Institution; **[8]**Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign; **[9]**Department of Environmental and Natural Resources Management, University of Ioannina, **[10]**Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland

March 2011

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

Expansion of the Genomic Encyclopedia of Bacteria and Archaea

Christian Rinke[1], Alex Sczyrba[1], Stephanie Malfatti[1],Janey Lee[1], Jan-Fang Cheng[1], Ramunas Stepanauskas[2], Jonathan A. Eisen[1,3], Steven Hallam[4]; William P. Inskeep[5]; Brian P. Hedlund[6]; Stefan M. Sievert[7]; Wen-Tso Liu[8]; George Tsiamis[9]; Philip Hugenholtz[10]; Tanja Woyke[1]

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Abstract

To date the vast majority of bacterial and archaeal genomes sequenced are of rather limited phylogenetic diversity as they were chosen based on their physiology and/ or medical importance. The Genomic Encyclopedia of Bacteria and Archaea (GEBA) project (Wu et al. 2009) is aimed at systematically filling the gaps of the tree of life with phylogenetically diverse reference genomes. However more than 99% of microorganisms elude current culturing attempts, severely limiting the ability to recover complete or even partial genomes of these largely mysterious species. These limitations gave rise to the GEBA uncultured project. Here we propose to use single cell genomics to massively expand the Genomic Encyclopedia of Bacteria and Archaea by targeting 80 single cell representatives of uncultured candidate phyla which have no or very few cultured representatives. Generating these reference genomes of uncultured microbes will dramatically increase the discovery rate of novel protein families and biological functions, shed light on the numerous underrepresented phyla that likely play important roles in the environment, and will assist in improving the reconstruction of the evolutionary history of *Bacteria* and *Archaea*. Moreover, these data will improve our ability to interpret metagenomics sequence data from diverse environments, which will be of tremendous value for microbial ecology and evolutionary studies to come.

Scientific Questions/ Goals

Diversity
Discovery of novel genes, protein families, pathways

Ecology
Functional roles of candidate phyla?
Are phyla functionally homogeneous?

Phylogenetic distribution of functions
Photosynthesis? CO_2 fixation?

DOE relevance:
Functions of DOE interest in novel lineages of single cell genomes?

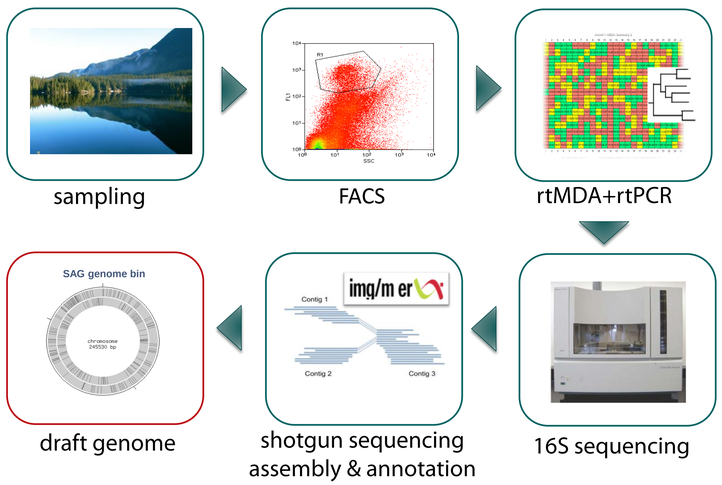
Evolution
Improved understanding on evolutionary diversification.
How did the bacterial domain evolve?

Diversity

Function

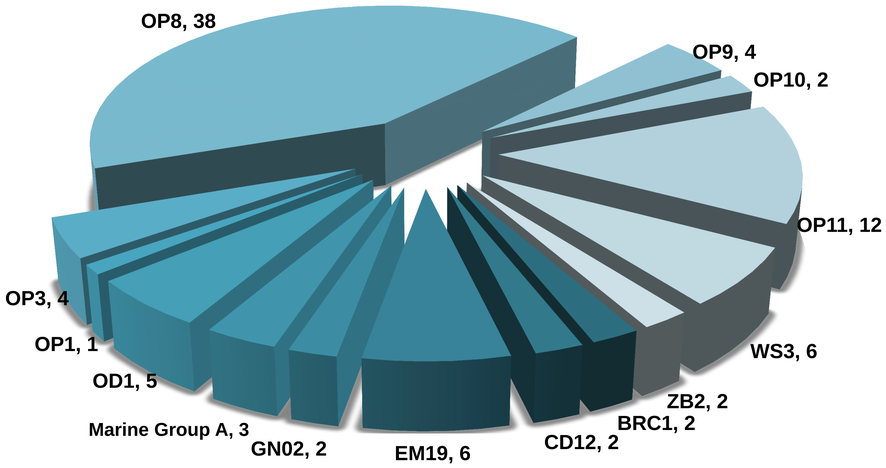
Evolution

Single Cell Genomics Pipeline

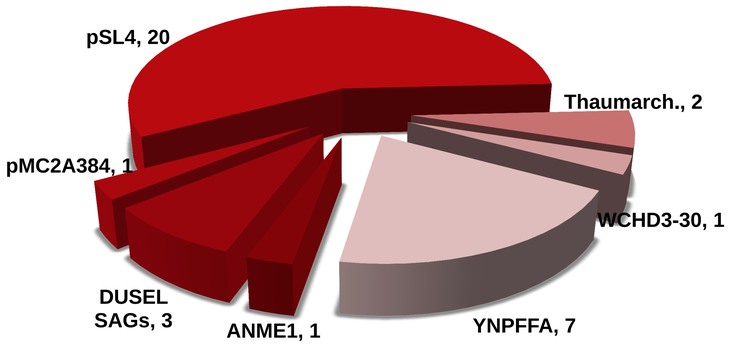


Results - Single Cell 16S Sequences

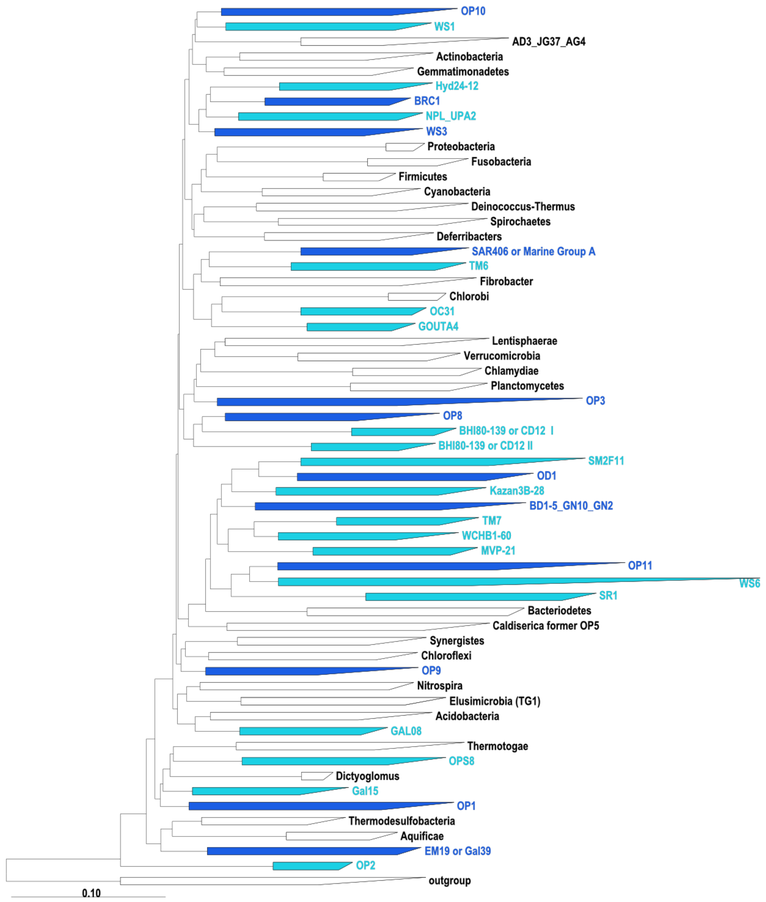
Novel Bacterial Lineages (89 total)



Novel Archaeal Lineages (35 total)

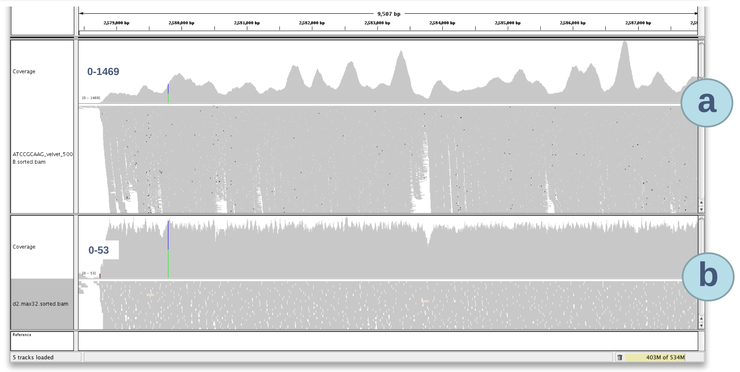


A total of 124 short 16S rRNA gene sequences of Single Amplified Genomes (SAGs) are available and will undergo low level shotgun sequencing in order to evaluate SAGS for deeper sequencing.

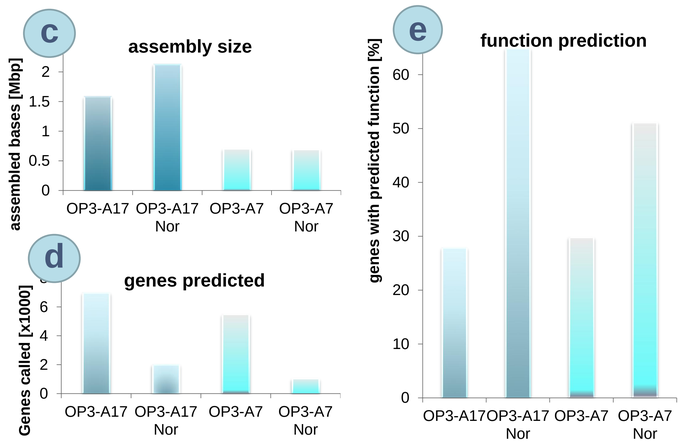


16S rRNA gene phylogenetic neighbor-joining tree of *Bacteria* based on the SILVA ARB database (www.arb-silva.de). Novel lineages of uncultivated representatives known only from environmental sequences are in blue: dark blue lineages are covered by the GEBA uncultured project, light blue lineages are not yet covered.

Results – Sequencing/Assembly



(a) MDA bias causes uneven genome coverage as shown by mapping the reads back to reference (velvet assembly). (b) Computational normalization based on k-mer frequency scales down coverage. (c) Assembly size of normalized reads (Nor) is equal to non- normalized reads. (d) Number of predicted genes is too high for non normalized assemblies. (e) Percentage of genes with a predicted function is higher for normalized assemblies.



Conclusion

Our single cell genomics pipeline allows amplification of genomes from microbial single cells directly from environmental samples, and thus the exploration of novel uncultured lineages.

References:
Wu D, et al. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462(7276):1056-60.

Author affiliations
[1] DOE Joint Genome Institute, Walnut Creek, California, United States of America; [2] Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, Maine, United States of America; [3] Department of Evolution and Ecology, University of California Davis, Davis, United States of America, [4] Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada; [5] Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, United States of America; [6] School of Life Sciences, University of Nevada, Las Vegas, Nevada, United States of America; [7] Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, United States of America; [8] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America; [9] Department of Environmental and Natural Resources Management, University of Ioannina, Agrinio, Greece; [10] Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, Australia.

Acknowledgements:
We thank Eric Tang for library creation and the JGI production team for sequencing effort.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.