

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

## 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

Diversity

### Ecology

Functional roles of candidate phyla?  
Are phyla functionally homogeneous?

Function

### Phylogenetic distribution of functions

Photosynthesis? CO<sub>2</sub> fixation?

### DOE relevance:

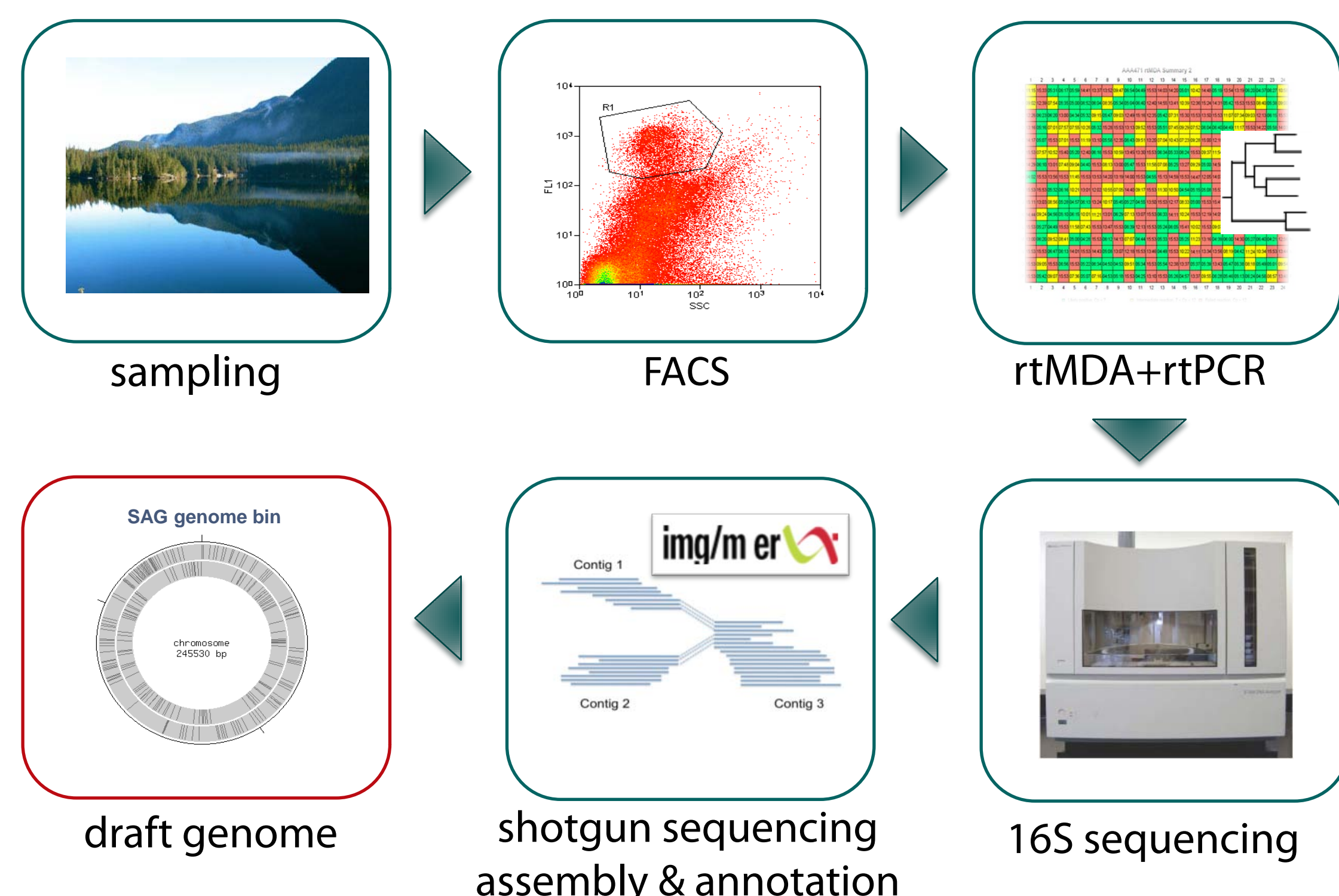
Functions of DOE interest in novel lineages of single cell genomes?

Evolution

### Evolution

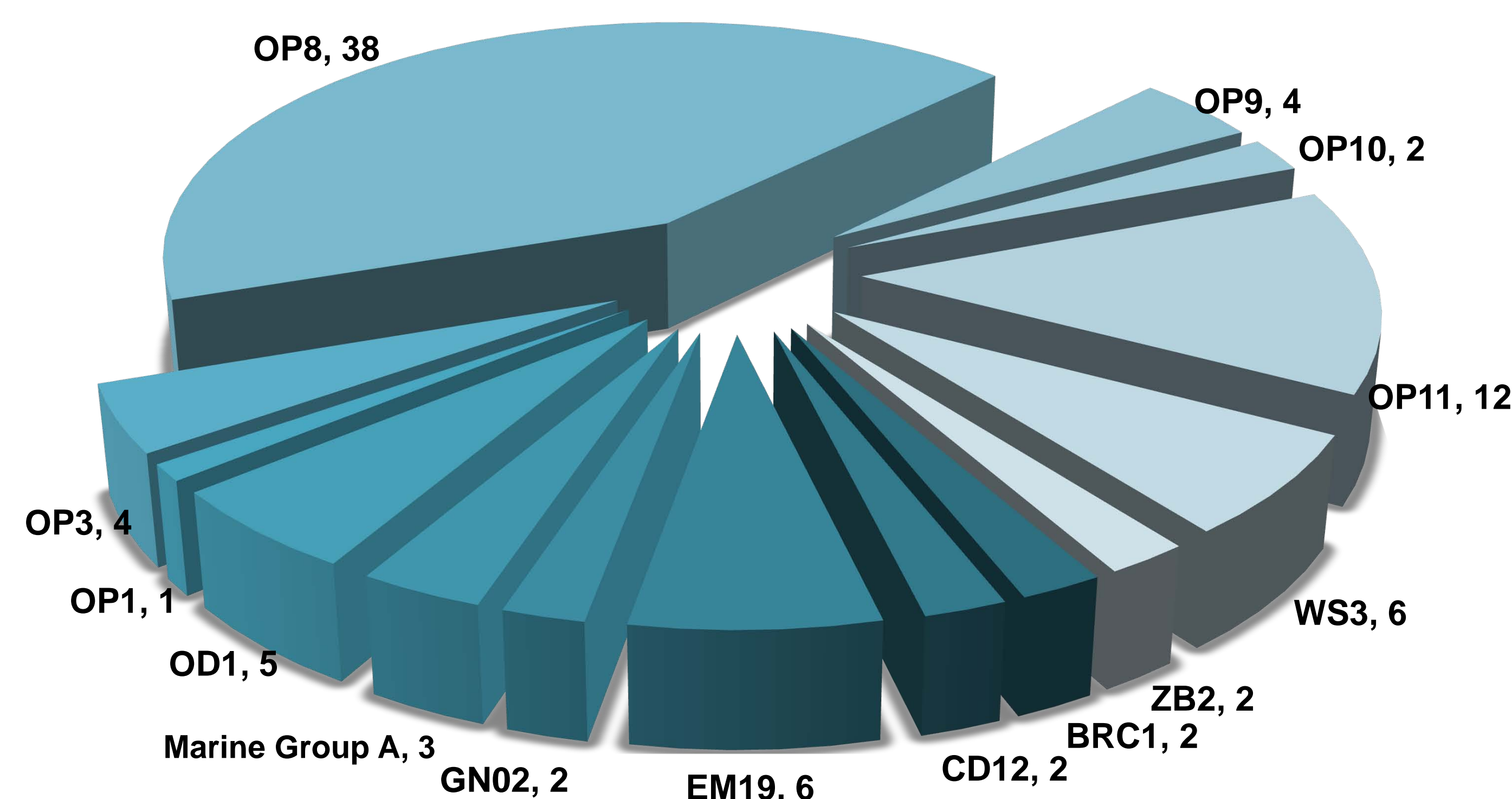
Improved understanding on evolutionary diversification.  
How did the bacterial domain evolve?

## Single Cell Genomics Pipeline

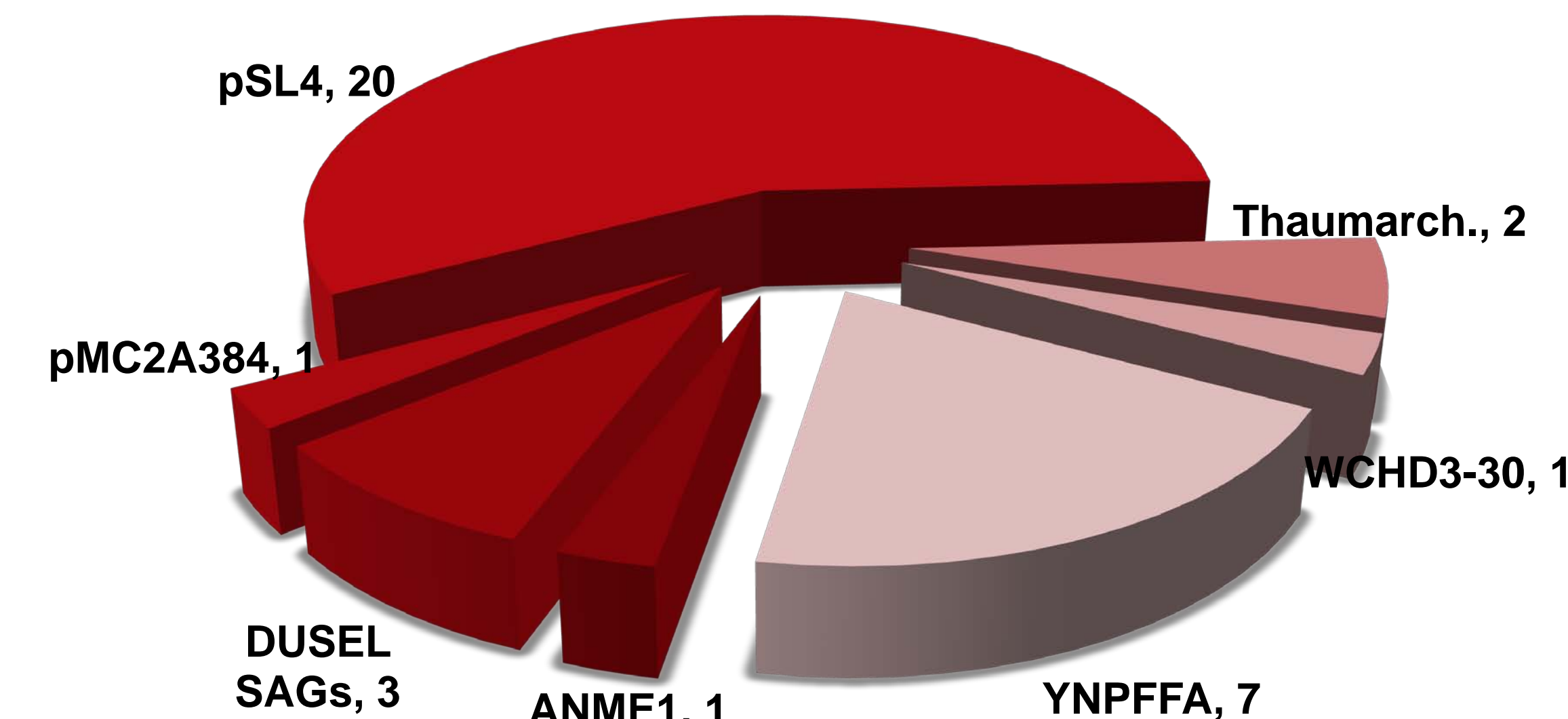


## Results - Single Cell 16S Sequences

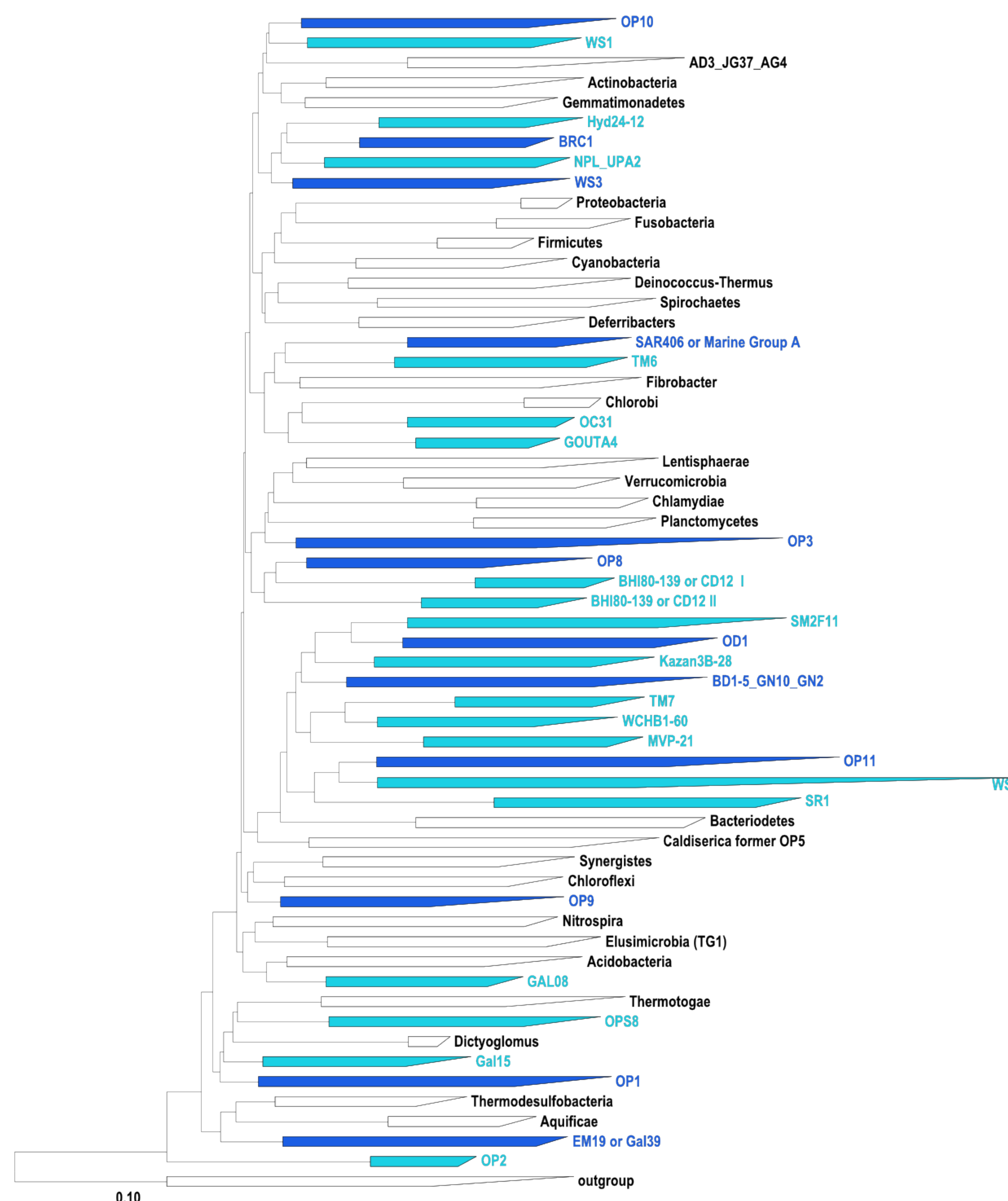
### Novel Bacterial Lineages (89 total)



### Novel Archaeal Lineages (35 total)

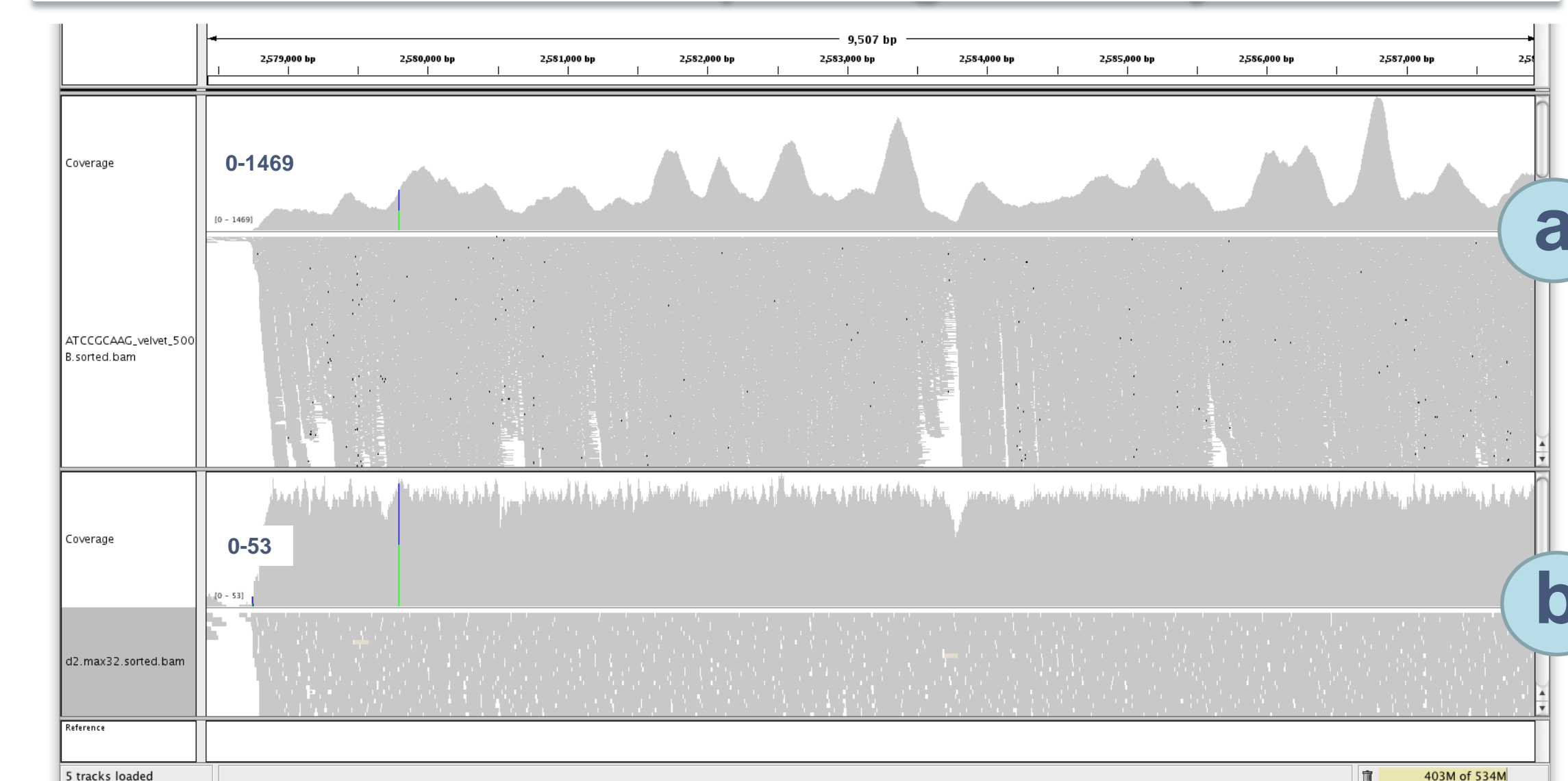


A total of 124 short 16S rRNA gene sequences of Single Amplified Genomes (SAGs) are available and will undergo low level shotgun sequencing in order to evaluate SAGs for deeper sequencing.

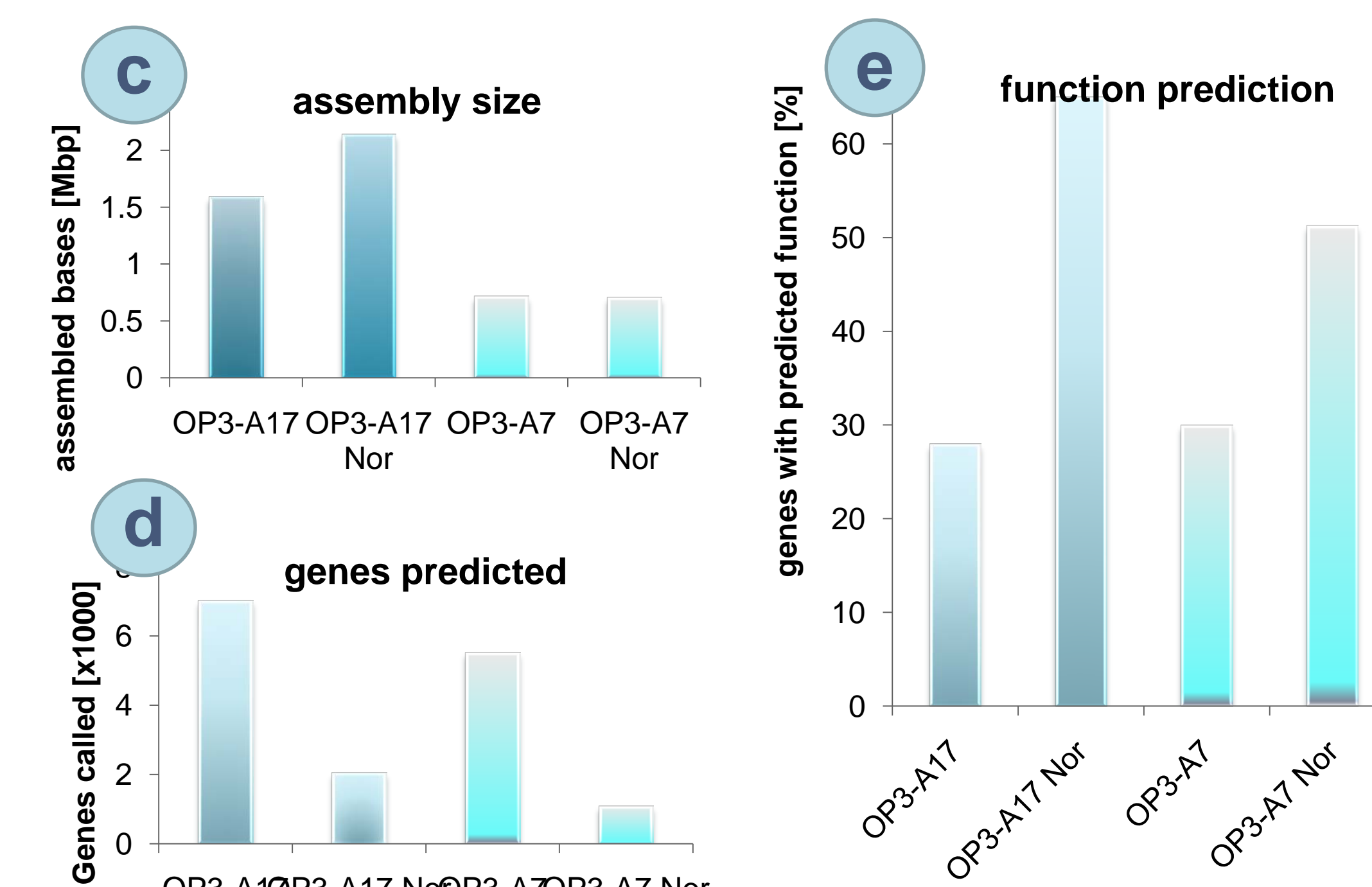


16S rRNA gene phylogenetic neighbor-joining tree of *Bacteria* based on the SILVA ARB database ([www.arb-silva.de](http://www.arb-silva.de)). Novel lineages of uncultivated representatives known only from environmental sequences are in blue: dark blue lineages are covered by the GEBA uncultured project, 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.

## **DISCLAIMER**

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.