Microbial Dark Matter: Unusual intervening sequences in 16S rRNA genes of candidate phyla from the deep subsurface

Jessica Jarett[1*], Ramunas Stepanauskas[2], Thomas Kieft[3], Tullis Onstott[4], Tanja Woyke[1],

[1] LBNL - Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA USA

[2] Bigelow Laboratory for Ocean Sciences, East Boothbay, ME USA

[3] New Mexico Tech, Soccorro, NM USA

[4] Princeton University, Princeton, NJ USA

**To whom correspondence should be addressed*: Email: jkjarett@lbl.gov

March 21, 2014

ACKNOWLEDGMENTS:

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. We thank Jan Amend and Lily Momper of the University of Southern California for providing 16S rRNA sequences from Homestake Mine for continued testing of this method. JJ wishes to thank Chris Rinke, Jim Tripp, Esther Singer, Scott Clingenpeel, and Nanichi Ramos-Roldan for assistance with data analysis and helpful discussions.

DISCLAIMER:

LBNL: 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

Microbial Dark Matter: Unusual intervening sequences in 16S rRNA genes of candidate phyla from the deep subsurface

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.

Microbial Dark Matter: Unusual intervening sequences in 16S rRNA genes of candidate phyla from the deep subsurface

Jessica Jarett[1], Ramunas Stepanauskas[2], Thomas L. Kieft[3], Tullis Onstott[4], Tanja Woyke[1]

[1]DOE Joint Genome Institue, Walnut Creek, CA; [2]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME; [3]New Mexico Tech, Soccoro, NM

ABSTRACT

The Microbial Dark Matter project has sequenced genomes from over 200 single cells from candidate phyla, greatly expanding our knowledge of the ecology, inferred metabolism, and evolution of these widely distributed, yet poorly understood lineages. The second phase of this project aims to sequence an additional 800 single cells from known as well as potentially novel candidate phyla derived from a variety of environments. In order to identify whole genome amplified single cells, screening based on phylogenetic placement of 16S rRNA gene sequences is being conducted. Briefly, derived 16S rRNA gene sequences are aligned to a custom version of the Greengenes reference database and added to a reference tree in ARB using parsimony. In multiple samples from deep subsurface habitats but not from other habitats, a large number of sequences proved difficult to align and therefore to place in the tree. Based on comparisons to reference sequences and structural alignments using SSU-ALIGN, many of these “difficult” sequences appear to originate from candidate phyla, and contain intervening sequences (IVSs) within the 16S rRNA genes. These IVSs are short (39 - 79 nt) and do not appear to be self-splicing or to contain open reading frames. IVSs were found in the loop regions of stem-loop structures in several different taxonomic groups. Phylogenetic placement of sequences is strongly affected by IVSs; two out of three groups investigated were classified as different phyla after their removal. Based on data from samples screened in this project, IVSs appear to be more common in microbes occurring in deep subsurface habitats, although the reasons for this remain elusive.

OBJECTIVES

- Develop a procedure for identifying and removing IVSs from 16S rRNA sequences
- Investigate the effects of IVSs on phylogenetic placement and classification
- Determine if IVSs should be routinely removed as part of the classification process

MATERIALS AND METHODS

Single cells were sorted from samples taken from a deep subsurface habitat in South Africa, their genomes amplified with MDA, and 16S rRNA genes were PCR amplified and Sanger sequenced. Sequences were aligned to a customized version of the Greengenes reference database (October 2012 release, with additional sequences from candidate phyla), using PyNAST within QIIME. Aligned sequences were imported into ARB and added to the reference tree using the quick-add ARB parsimony method to determine a probable phylum. For sequences placed on unusually long branches, alignments were visually inspected and selected sequences were compared to the NCBI nr nucleotide database using BLAST. A consensus identity at the phylum level was drawn for each sequence by taking into account the placement on the tree, classification by the SILVA and Greengenes online classifiers, and the similarity of each sequence to its closest match in each database. Where classification methods did not agree and the similarity to reference sequences was low (< 80%), sequences were grouped into “unknown” categories based on their placement in the tree. Several of these unknown groups were investigated further using structural alignments with SSU-ALIGN. The ssu-align and ssu-mask commands were used to identify the boundaries of the intervening sequences, defined as regions of at least 5 nt that were present in query sequences but not in the reference model. IVSs were located within the secondary structure based on diagrams from the ssu-draw command. IVSs were removed from parent sequences and the alignment and treeing procedure was repeated to determine a new phylogenetic placement for each unknown group.

RESULTS

- IVSs were readily identified using SSU-ALIGN software
- Within a group, the sequences and lengths of the IVSs were identical
- In the three groups analyzed, IVSs > 5nt in length occurred in only two locations in the 16S rRNA (Fig. 1)

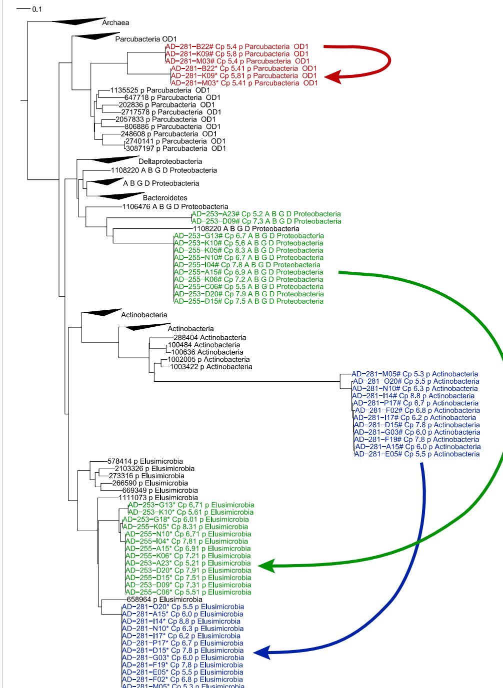

Figure 1. Approximately-maximum likelihood phylogenetic inference of the 16S rRNA gene, with sequences classified at the phylum level. Sequences shown in color were added to the tree using parsimony. Three different “unknown” groups are denoted by the colors, with arrows indicating changes in the placement of a group after IVSs were removed.

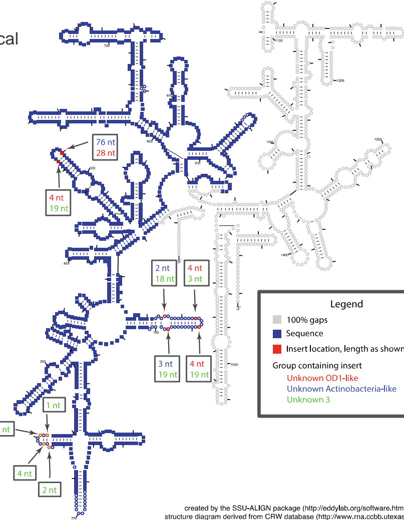

Figure 1. The location, length, and source group of intervening sequences (IVSs) within the consensus secondary structure of 16S rRNA. Note that the location of inserts is accurate but the IVSs themselves are not shown in the secondary structure.

- For the unknown Actinobacteria-like (blue) and unknown 3 (green) groups, removal of the IVSs resulted in placement in a different phylum and greatly reduced branch length.
- For the unknown OD1-like group (red), removal of the IVS resulted in placement in the same phylum, but there was no appreciable reduction in branch length.

DISCUSSION

IVSs were first discovered in Bacteria in 1994 (1) and have been found in both the 16S and 23S rRNA genes of diverse species of bacteria (2). They differ from introns in that IVSs do not contain open reading frames, are not self-splicing, and are typically short (between 100-200 nt). Many IVSs can form secondary structures (3), which may facilitate their excision from rRNA by enzymes targeting such duplex formations. (2). The potential of the IVSs from deep subsurface groups to form secondary structures at the high temperatures characteristic of these environments is being investigated.

The deep subsurface samples in this study are unusual because a large proportion of cells appear to contain IVSs (approximately 30-50%). IVSs are typically sporadically distributed taxonomically, even amongst closely related species, and this combined with phylogenetic analyses of IVS and intron sequences suggests that these elements are mobile and horizontally transferred (4, Skurnik and Toivanen). Analysis of other deep subsurface samples may reveal if this prevalence of IVSs, and perhaps of horizontal gene transfer, is a typical characteristic of these habitats.

Although the phylogenetic placement of many of these sequences changed dramatically with the removal of IVSs, it is unclear if this is the most appropriate method of classification. Planned sequencing of single-cell genomes from these groups will reveal whether IVSs serve as indicators for taxa with a unique evolutionary history or not.

CONCLUSIONS

- SSU-ALIGN provides a simple method by which to identify and thereby remove IVSs from 16S rRNA gene sequences.
- The removal of IVSs had dramatic effects on the phylogenetic placement of two out of three groups, changing their classification at the phylum level. A third group remained in its original phylum.
- Testing and refinement of this method using other taxonomic groups as well as other deep subsurface samples from the Homestake Mine* is ongoing.

JOIN THE MICROBIAL DARK MATTER PROJECT

Talk to Jessica Jarett or Tanja Woyke about being a part of Phase II of the Microbial Dark Matter Project!

The MDM project aims to shed light on the phylogeny, metabolism, and ecology of unexplored branches of the tree of life by sequencing single cell genomes from candidate phyla. Data from Phase I has been published in Nature and revealed new superphyla and novel functions in Bacteria and Archaea.For Phase II, we have approved funding for sequencing 800 single cell genomes. We are looking for samples from diverse environments with a high percentage of **candidate phyla**, or with possibly **novel phyla**. If your samples are selected, you will receive 16S rRNA itag sequences as well as assembled single-cell genomes.

For more information, email **jkjarett@lbl.gov.**

REFERENCES AND ACKNOWLEDGEMENTS

References: 1. Linton et al. (1994) Microbiology 140:847 2. Belfort et al. (1995) J. Bacteriol 177(14):3897 3. Pei et al. (2010) Appl Env Micro 76(12):3886 4. Salman et al. (2012) PNAS 109(11):4203 5. Skurnik and Toivanen (1991) Molec Micro 5(3):585

*We thank Jan Amend and Lily Momper of the University of Southern California for providing 16S rRNA sequences from Homestake Mine for continued testing of this method. JJ wishes to thank Chris Rinke, Jim Tripp, Esther Singer, Scott Clingenpeel, and Nanichi Ramos-Roldan for assistance with data analysis and helpful discussions.

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.