Using Emergent Self-Organizing Maps to Identify Marine Group II *Archaea* Genomic Fragments From Uncharacterized Microbial Metagenomic Sequences

by

Rachel A. Hillmer

B.S., Physics (2007)

University of Illinois at Urbana-Champaign

Submitted to the Department of Biological Engineering
in partial fulfillment of the requirements for the degree of
Master of Science in Biological Engineering

at the

Massachusetts Institute of Technology

February 2010

Signature of Author ....
Department of Biological Engineering
January 15, 2010

Certified by ....
Edward F. DeLong
Professor of Biological Engineeering and Civil and Environmental Engineering
Thesis Supervisor

Accepted by ....
Darrell J. Irvine
Eugene Bell Associate Professor of Tissue Engineering
Department of Biological Engineering
Chairman, Committee on Graduate Students

USING EMERGENT SELF-ORGANIZING MAPS TO IDENTIFY MARINE GROUP II *ARCHAEA* GENOMIC FRAGMENTS FROM UNCHARACHTERIZED METAGENOMIC SEQUENCES

by

RACHEL HILLMER

Submitted to the Department of Biological Engineering
on January 15, 2010 in partial fulfillment of the
requirements for the Degree of Master of Science in
Biological Engineering

ABSTRACT

The validity and usefulness of clustering marine group II tetranucleotide signatures using emergent self-organizing maps was investigated. Fosmids from the HF200 library were chosen for sequencing based on end-sequence tetranucleotide clustering with group II seed sequences, as well as blastx homology. Fosmids were sequenced using a single 454-titanium sequencing run, and contigs subsequently assembled in silico. A total of 99 contigs over 20kb were retrieved, at least 72 of which belong to the marine group II archaea.

The phylogenetic substructure of the marine group II archaeal clusters having more than a few representatives was investigated, by clustering tetranucleotide signatures of group II contigs over 20kb, also with an emergent self-organizing map. The distribution of these clusters in the Hawaii Ocean Time Series depth profile fosmid libraries in the DeLong lab were mapped onto depth profiles from three independent cruises.

Thesis Supervisor: Edward F. DeLong

Title: Professor of Biological Engineeering and of Civil and Environmental Engineering

Table of Contents:

Introduction

That the study of microorganisms is vitally important to science goes without saying: by biomass alone, microorganisms dominate the biosphere; microorganisms are necessary for and drive a substantial part of the biogeochemical cycles on earth (ie. carbon and nitrogen cycles). Microorganisms also noticeably impact the health and function of the human body—bacterial and archaeal cells outnumber human cells 10 to 1 in the human body, and many of the notable human diseases are caused by bacterial and microbial pathogens. Microorganisms, measured by the number of parts per cell, are the simplest cellular life forms on the planet, and as such provide important clues about the evolution of cellular life; finally, microorganisms hold an owner's share on the biochemical diversity of life. How best to study these wondrous creatures is less clear.

Pure culture studies have worked well to isolated and characterize many microorganisms of medical interest, but environmental microorganisms are understood to be of equally vital importance to environmental processes since Sergei Winogradsky, who first studied microorganisms beyond the medical scope. Most of our knowledge of molecular biology has been derived from work with bacteria grown on the culture plate. Pure culture techniques continues to provide an ever-increasing palette of tests that may be run on clonal populations of microorganisms that undergo exponential growth on media. These tests, that form the arsenal of molecular biology, include screening for various biological functions, tagging and calibrating production of genes of interest, and the direct observation of response to various environmental stimuli. Yet many microorganisms have long resisted cultivation. Despite articulate arguments as to why we should take the time to find media appropriate for each strain, the sheer number of as-yet-uncultivated strains, and the trial-and-error effort needed to do so, begs the use of alternative methods. Moreover, even for those that can be cultured the culture plate is a selection step: organisms with the appropriate cocktail of genes that allow optimum selective advantage in the wild may no longer have the necessary selective advantage in culture medium. It has also recently been pointed out that cell-cell communication between non-clonal organisms may be critical in their response to various environmental changes. These are by definition difficult to probe in a clonal population. It could even be that much of the behavior of microorganisms on a culture plate is an artifact, that the "real" behavior microorganisms display in the wild is lost in culture.

The ultimate goal of microbial ecology is to understand global biogeochemical cycles—how microorganisms change and respond to their environment—on the full range of length and time scales relevant. In 1977, Carl Woese presented a classification system for microorganisms based on their 16S rRNA sequences, in stark contrast to earlier microbial classification schemes which relied on more gross measures such as the morphology, biochemical abilities of an organism, and the GC content of its genome. Because of the high resolution between species given by16S rRNA comparison, Woese's work provided for the first time an evolutionary tree for microorganisms, organisms which until the late 1970s had resisted evolutionary classification. Based on Woese's 16S classification system, Norman Pace and colleagues introduced culture-free methods of identifying microorganisms shortly thereafter, in 1985. It then soon became clear that we are not able to culture the vast majority of microorganisms in the environment; this "uncultured majority" is, by some estimates, well over 90 percent of the total extant microbial biomass on our planet. Culture-free methods are thus necessary to answer the first half of the foundational question of ecology, "Who is out there, and what do they do?"

Metagenomics is the study of microorganisms by gathering DNA directly from environmental sources, without first cultivating the organisms. DNA is sheared and either kept within E.coli plasmids—in either long inserts ranging from hundreds of kbp in Bacterial Articifical Chromosome vectors or ~36kb in fosmds, to short sequences of ~3kb in shotgun DNA sequencing libraries. Alternatively, microbial community DNA can be sequenced directly, 454 sequencing being the high-throughput next-generation sequencing of choice for uncharted microbial realms because of the longer read lengths generated (~100-400bp), compared to other technologies. Short *E.coli* plasmid inserts are amenable to shotgun sequencing, the long ones can be end-sequenced with Sanger sequencing. These end-seqeunces can then be used to choose fosmids for sequencing. While sequencing all approximately 50,000 fosmids in a given metagenomic library in the DeLong lab is not always financially feasible, targetings elect clones based on their end sequences, and sequencing the full length of each can provide an invaluable glimpse into a community, with long contiguous genome segments from uncultured, and likely entirely unexplored, organisms. —Such data are useful, since estimates based on cell counts via flow cytometry to cell counts via serial dilution and culturing of ocean waters suggest that 99.99% of all ocean microorganisms remain uncultured to date.

Recent developments also allow for the sequencing of RNA from microbial communities, through the construction of cDNA libraries, and even more recent breakthroughs are allowing for community proteomics. But finding the genomic potential of a community, stored in its DNA, is the foundation on which transcriptomic and proteomic analyses lie.

The efforts of the DeLong lab are focused on studying marine microorganisms [1,2], and their ecology. Why open ocean samples? For technical reasons, the plankton in seawater is easier to sample than soil or other solid media; length, and therefore time scales are spread out in the ocean, thus current state-of-the-art measurements give more bang for your buck, and make modeling the community theoretically more feasible.

Metagenomics is opening up the secrets of many uncultured microbial communities from those, as studied in the DeLong lab, of marine microorganisms, the human microbiome, soil-based microbial communities, and even the extinct mammoth [3,4,5]. But until single-cell sequencing comes online [6,7,8], which would allow for the complete sequencing of a microbial genome, given only a single copy of its genome—that contained in one cell—one of the greatest challenges in metagenomics that must be surmounted is the binning problem: knowing which genome fragment came from what organism so that genomes can be stitched back together, or at least partially recovered. Marking out genes, and new families of genes, "gene-finding," has its place, especially in the search for enzymes with new biochemical functions, but to truly understand a microbial community, to chart out its functioning in both space and time, and to be able to predict the effects of permutations of its natural state on that community, requires, as a first step, assembling those genes into pathways. It requires knowing *who* lives in the community, not just the names of the residents (their 16S rRNA gene sequences), but who they are – the potential of what they can do, the information in which is stored in the whole genome.

The trouble of assembling the genome of a species is compounded even more by the fact that defining a species is not trivial, or even obvious. Within-species, and with-sub-species heterogeneity, "microheterogeneity," is uncharted territory, a reality of life in complex microbial communities [9,10,11], often eliminated, or substantially lessened in the culture plate. Take *Prochlorococcus* for example, which makes up nearly a third of the microbial mass in the temperate open ocean. Among the samples taken from Venter's Sorcerer II cruise, over 42 sampling locations and up to 600,000 shotgun inserts constructed from the DNA in surface seawater gathered at each location, he estimates that no two DNA inserts came form clonally identical organisms [12]. We're a long way from the culture plate.

Thus bioinformatic methods form a critical joint in the metagenomics apparatus [13]. As shotgun sequencing sped up, nearly exponentially, the sequencing of the human genome, by allowing for computer assembly of fragments, short-circuiting the laborious process of assembling long contigs with physical overlap, so new bioinformatics methods, allow for the bypassing even requiring that segments overlap at all in order to bin them into the same organism. Binning methods involve determining sequence origin either by comparison to existing genomes in the ever-growing database of genome sequences on NCBI, or most recently, by analysis of tetranucleotide signatures of long (>20kb) genome fragments. The latter allows for database-independent analysis of genome sequence, particularly amenable to microbial communities where representatives of most species remain unsequenced.

Enter the marine group II *Archaea*. They are among the mesophilic *Archaea* found in ocean waters around the globe. Little is known about these archaea except for their phylogeny, and the genes immediately adjacent to the 16S gene, and thus on the same contig as the 16S gene, the phylogeny indicator. While marine planktonic *Crenarchaea* (known as marine group I) are found predominately in deep waters (those below the photic zone), the euryarchaea are found throughout the water column, in substantial abundance—the highest measured to date is about 13% of the microbial planktonic mass, in the Hawaii Ocean Time Series Depth column. The marine group II *Archaea* are particularly amenable to highlighting the importance of database-independent binning techniques, since no marine euryarchaea (an entire division of one of the three domains of life) have any cultured representatives; the closest sequenced relatives of the marine group II archaea are the thermoplasmas, which reside in sulfurous hot springs.

In this thesis, emergent self-organizing maps are used to choose fosmids from the Hawaii Ocean Time-Series (HOT) site microbial genome fragment libraries in the DeLong lab for sequencing, from the end-sequences available. The results of sequencing all the fosmids together in a single 454-titanium run and assembling the contigs using the accessory assembly software from Roche are processed to choose those sequences that are certainly marine group II. The phylogenetic substructure of the full set of marine group II sequences is then analyzed, and the group II presence in the HOT depth profiles is then mapped out.

References

[1] DeLong EF Microbial community genomics in the ocean. Microbiology Nature Reviews 2005; 3: 459-468.

[2] DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU et al. Community genomics among stratified microbial assemblages in the ocean's interior. Science 2006; 311: 496-503.

[3] Handelsman J. Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiology and Molec. Biol. Reviews 2004; 669-685.

[4] Whitaker RJ, Banfield JF. Population genomics in natural microbial communities. Microbial ecology 2006; 21:509-516.

[5] Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 2004; 428: 37-43.

[6] Lasken RS. Single-cell genomic sequencing using Multiple Displacement Amplification. Current Opinion in Microbiology. 2007; 10:510-516.

[7] Ottesen EA, Hong JW, Quake SR, Leadbetter, JR. Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. Science 2006; 314:1464-1467.

[8] Zhang K, Martiny AC, Nikos BR, Barry KW, Malek J, Chisholm SW et al. Sequencing genomes from single cells by polymerase cloning. Nature Biotechnology 2006; 24:680-686.

[9] Ward DM, Cohan FM, Bhaya D, Heidelberf JF, Kuhl M, Grossman A. Genomics, environmental genomics and the issue of microbial species. Heredity 2008; 100:207-219.

[10] Denef VJ, Shah MV, VerBerkmoes NC, Hettich, RL, Banfield JF. Implications of strain- and species-level sequence divergence for community and isolate shotgun proteomic analysis. Journal of Proteome Research 2007; 6:3152-3161.

[11] Merring C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen, LJ et al. Quantitative phylogenetic assessment of microbial communities in diverse environments. Science 2007; 315: 1126-1130.

[12] Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S. The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biology 2007; 5: 398-431.

[13] Raes J, Foerstner KU, Bork P. Get the most out of your metagenome: computational analysis of environmental sequence data. Current Opinion in Microbiology 2007; 10:492-492.

Chapter 1

Clustering Metagenomic Sequences with Emergent Self-Organizing Maps

Introduction

It was recognized through the efforts of systematic bacteriologists in the mid-1900s that GC content, while reasonably well conserved across a genome, is usually not a sufficiently narrow signature to resolve closely related bacterial species, or genomes in a heterogeneous population. Thus until the advent of whole-genome sequencing, DNA-DNA or DNA-RNA hybridization, between regions that are now known as marker genes, was used to identify species.

In the last decade, as numerous microbes have been fully sequenced (1337 bacteria and 84 archaea to date), mathematical extensions of calculating GC content have been explored as a method to computationally distinguish species. Instead of habitating 2D space—there are two numbers associated with GC content: percent pyrimidine and percent purine—methods considering more than one nucleotide at a time live in far higher-dimensional spaces. Di-nucleotide signatures (12-dimensional space), tri-nucleotide signatures (64-dimensional space), tetranucleotide signatures (136-dimensional space) and others have been widely investigated in the bioinformatics literature [1,2,3,4]. Even higher-dimensional signatures, including octanucleotide and 15-mer signatures have received some attention [5,6,7].

As a signature to resolve between species, the tetranucleotide signature has received special attention: The variation in the signature between species is noticeably larger than the variation within species, and the signature is conserved, even for relatively short genome fragments (3kb-20kb, 0.3% - 8% of the genome length, assuming an average prokaryotic microbial genome size of 2.5Mb) [8,9]. Signatures of longer oligonucleotides require longer genome fragments to recover a statistically significant signature measurement, and shorter signatures are not always effective at resolving between closely related species.

Why the tetranucleotide signature is conserved so well across known genomes is still a matter for debate, although genome repair and mutation mechanisms driven by slightly different enzymes in different organisms has been suggested as the source. It is known that, in general, the signature is not driven by a few over-represented oligos, but rather by the frequencies of the bulk of the oligonucleotides. Not all the oligos appear necessary, some even make the resolution between species less clear (personal observation, unpublished), but which oligos should be removed to improve clustering seems to vary between species—it's not obvious, not knowing the origin of a fragment, which oligos should be removed.

Because the tetranucleotide signature is conserved across the genome and can resolve between species (in most cases), it is a natural target for metagenomic analyses [10,11,12, 13, 15]. Standard procedure in many metagenomic analysis pipelines involves categorizing genome fragments by their closest BLAST hits. In many cases, however, the majority of species in a given environment are not represented in the online NCBI database—current microbial databases are *heavily* weighted towards species with cultured representatives. In the case of the open ocean, where an estimated 99.99% of all species present remain to be cultured, it is highly desirable to have a database-independent method of binning sequences, just as our collection methods in metagenomic studies are themselves culture-independent.

Most efforts using tetranucleotides to bin metagenomic sequences have been done with simulated metagenomic datasets, although a few notable exceptions have investigated actual experimental sequences [14,16].

Methods

All tetranucleotide signatures were calculated and normalized with an in-house script (Script 1, Appendix). Clustering was performed and visualized with an Emergent self-organizing map (ESOM), freely available from Databionics here: http://databionic-esom.sourceforge.net/ with online documentation.

Individual colonies for each of the fosmids to be sequenced were hand-picked and re-grown, in separate wells of 2 two 96-well plates, to log phase; some clones grow faster than others; this ensures the fosmid-containing *E.coli* are not competing against each other. The plasmids, usually present in low-copy number were induced to high copy by induction of the pBAD promoter immediately before cell lysis. Plasmid DNA was extracted, pooled together from all the individual colonies, and run through a CsCl density gradient to separate *E.coli* chromosomal DNA from plasmid DNA. All the DNA was combined as the input to a single full plate 454 titanium run. The assembler gsAssembler was used to assemble the raw sequences; default parameters were used.

Results and Discussion

Step 1. Calibration of the ESOM

G.J. Dick et al (2009) show [16] that the ESOM from Databionics can be used to cluster fragments longer than 2kb, with an average fragment size of 5kb, when the genome fragments are taken from two acidophilic biofilm communities. These communities, however, are known for their very low amount of species diversity; the species themselves rarely experience genome rearrangment. The diversity is so low that several whole genomes have been reconstructed using only metagenomic sequence, without a culturing step (quite a feat) [17]. The general relation between fragment size and the ability to resolve species using a tetranucleotide approach remained unclear to me from literature surveys. Knowing that my goal in this project is to resolve and cluster mesophilic euryarchaea, I chose evenly spaced, nonoverlapping fragments from methanogens as my calibration sequences. The methanogens seemed a good choice since, like the marine group II *Archaea*, they are all *Euryarchaea*, and several are mesophilic (rare for *Archaea*, but universally true for the marine group II *Archaea*).

Fragment size was reduced stepwise from 20kb—the generally agreed-upon length for which clustering, by any technique, of genomic fragments is reliable, to 3kb, roughly the lower limit agreed upon in the literature for which clustering using artificial intelligence techniques (ie. self-organizing maps) is reliable.

Figures 1.1 and 1.2 show the clustering of 20kb fragments; Figure 1.1 uses all of the fragments generated when the methanogen genomes are cut into fragments of equal size. Figure 1.2 uses only 1 out of every 10, since such an undersampled situation is a more accurate representation of what may be expected in a metagenomic library. When clustering fragments from environmental sequences a perfect covering of each species is not always available. Thus it is important to verify that the species-specific clustering seen in Figure 1.1 is maintained when only a sparse selection of the fragments is chosen, as shown in Figure 1.2. As expected from the literature predictions, both Figure 1.1 and Figure 1.2 show a species-specific clustering in most cases.

While reliable clustering could be performed for 20kb fragments, as reported for most genomes investigated in the literature, in practice the genome fragments available from metagenomic datasets are substantially shorter. While clear statistical assignment of a fragment to a species usually requires at least 20kb of sequence [11], artificial intelligence (AI)-based clustering techniques, among them the self-organizing maps (SOMs), report reliable clustering for most fragments over 3kb. This is verified for the case of the cultured methanogens in Figure 1.3. Figure 1.3 depicts the clustering of 3kb fragments, using only 1 of every 3 fragments, to reduce the computational load. (Using all of the fragments would have required a clustering time of several days on our datarig server).

The clustering of 3kb methanogenic sequences is not perfect. There are a few sequences that stray over into other regions, or live far up in the mountains by themselves. These are either cases of horizontal gene transfer (HGT) or of highly repetitive sequences, which are heavily biased towards a narrow range of the available tetranucleotides, something that is uncharacteristic of most areas of prokaryotic genomes.

When the variation of the tetranucleotide signature for fragments within a species, which will increase as fragment size decreases, becomes comparable to the variation between species, the ability to resolve fragments into species bins becomes less optimal. The question, therefore, is; what is this limit?

The results of Figure 1.3 suggest that it may indeed be possible to cluster environmental sequences as small as 3kb into near-species bins. The length that is relevant to shotgun sequence datasets, however, is the limitation set by shotgun sequencing, which generally produces about 700-900 bp of usable sequence. Figure 1.4 depicts the clustering of 1 in every 10 700bp fragments from the cultured methanogens. The visual elevations marking off boundaries between clusters of sequences, so apparent in Figures 1.1-1.3, are no longer visible.

However, upon closer viewing (second image in Figure 1.4), although boundaries may not naturally arise, as happened in the maps previously shown, sequences that appeared clustered together in the previous maps do appear to, for the most part, cluster together. From Figure 1.4, it follows that although all of the sequences belonging to a naturally arising cluster may not be readily extracted from an ESOM, it nonetheless could be possible, given seed sequences, whose origin is known, to extract sequences that cluster within that boundary drawn around the seed sequences. Several sequences within that boundary may be incorrectly clustered, but on the whole the extraction will be reliable if the seed sequences are reliable.

However, it should be noted that if the seed sequences represent only a piece of the area belonging to the genome(s) of interest, then only *some* of the sequences belonging to those species will, in general, be clustered. This is in marked difference to the clustering shown above for sequences 20kb, and 3kb in length. In those cases, a marker sequence is needed to select which valleys are of interest, but the boundaries of those valleys are self-emergent, in a given collection of sequences. In the case of 700bp sequences, however, seed sequences are needed to define the boundaries *themselves* of the regions of interest.

Step 2: Selection of fosmids to sequence on a single 454 run.

A previous study had been made using 16S rDNA sequences found within fosmid libraries constructed from the water column during the "HF" cruise by hybridizing fosmid macroarrays with probes targeting archaeal rRNA. Figure S6 of [18] shows the distribution of marine group II archaeal 16S rRNA genes from that cruise, in the form of a tree, with topology calculated from multiple sequence alignments of the 1200+bp 16S rRNA sequences. The tightest, most abundant clade from a single depth is found among sequences from the 200m library. I therefore decided to choose fosmids to sequence from the HF200 library.

The HF200 library consists of several thousand fosmid sequences (~36kb long plasmid inserts), which had been end-sequenced using shotgun sequencing. Thus, for each fosmid, on average, approximately 700bp of sequence are available for each end of the fosmid.

I received 39 fosmid sequences from the HF130m library that the DeLong lab had had sequenced at an earlier date, which were labeled, putatively, as group II *Archaea*. To add to this, I collected all the full-length fosmid sequences on NCBI labeled as group II because they contained a 16S rRNA gene. Additionally, I gathered other sequences on NCBI showing BLAST homology to those 39 HF130 fosmid sequences, or clustering with them in an SOM. All of the unlabeled fosmids (with respect to phylogeny) retrieved from NCBI are listed below. NCBI fosmids labeled as group II are included and listed in Chapter 2.

NCBI Fosmids that show strong blast homology to putative group II HF130 contigs:

EBAC37F11

APKG2H5

APKG8D23

HF4000_ANIW141L21

HF4000_ANIW137P11

HF4000_ANIW141C7

HF4000_ANIW133F6

APKG1C9

APKG7N23

HF4000_001N02

NCBI Fosmids that do not show blast homology to any group II contigs but whose tetranucleotide signatures make them possible candidates for being group II:

HF4000_ANIW137G21

APKG2M17 (??)

HF4000_137B17

HF4000_ANIW141A21 (??)

HF4000_[384]001N02

HF70_19B12

The 39 fosmids over 20kb sequenced earlier from HF130 belong to a collection of putative group II archaeal fosmids sequenced because the end sequences showed homology to euryarchaeal archaeal marker genes. Only 4 groups of *Archaea* are known to live among the plankton in open ocean waters: marine groups I, II, III, and IV. Group I are *Crenarchaea*, the other three groups are *Euryarchaea*. According to 16S rRNA surveys from the HF cruise, most of the *Archaea* were from the group I and group II planktonic archaeal types. Thus any fosmids showing homology to euryarchaeal marker genes were likely group II. Fosmids were hand selected in this manner for sequencing by JGI. Returned fosmid sequences (assembled by phrap from shotgun sequences) were then clustered by tetranucleotide signatures to verify that they were group II *Archaea*.

The aforementioned clustering is shown in Figure 1.6. Three sequences clearly clustered away from the other sequences. When clustering these three sequences with other full-length fosmids on NCBI (both group II *Archaea*, non-group II *Archaea*, and sequences whose phylogeny is unclear), the three sequences in Figure 1.6 that cluster away from the remaining sequences consistently cluster closer to group III and group I sequences than to known group II *Archaea* sequences. They were therefore removed from among the seed sequences.

Since SOM clustering of HF200 fosmid end sequences requires that seed sequences be shorter than 2kb, I investigated all HF130 putative fosmid contigs longer than 2kb. Although only 39 contigs over 20kb were returned to the DeLong lab by JGI, several shorter contigs were also included. Clustering of all contigs over 2 kb is shown in Figure 1.7. Note that these sequences vary in length. Again, sequences that clustered far away from the bulk of the sequences were removed from the pool of group II archaeal seed sequences. The re-clustering, without outliers, is shown in Figure 1.8.

Using seed sequences from 3 sources—NCBI group II fosmids, NCBI putative group II fosmids, and HF130 contigs selected above—all cut into 1kb pieces, I clustered the HF200 fosmid end sequences, using 3-letter codes to identify the fosmid sequences. Ideally I would have trimmed off the poor-quality portions of the sequences, to reduce noise in clustering, but quality score files were not available for any of the HF fosmid library end-sequences, so I just used the whole sequences. Clustering of the HF200 library end sequences with background seed sequences is shown in Figure 1.9. Putative group II archaeal fosmid end-sequences were selected by hand by encircling the portions of the map most densely populated with seed sequences.

These fosmid end-sequences were then run through a blast filter: those sequences for which both ends had strong homology to archaeal genes, or to existing group II archaeal fosmids, were selected for sequencing. The top blast hit(s) of the sequenced fosmid ends are listed in List 1.1 (see Appendix). Assembled contigs returned from the gsAssembler after fosmid sequencing with a single 454 titanium run (see Methods) were covered to varying degrees by raw sequence (Figure 1.15).

By BLASTing the assembled contigs longer than 25kb against the end sequences of those fosmids input into the 454 sequencer, contigs were assigned to original fosmid names (List 1.2, Appendix). Chapter 2 deals with the investigation and analysis of the assembled sequences from the 454 run.

Step 3: Additional investigation into minimum fragment length generating emergent clustering of sequences.

While individual shotgun sequences are about 700 bp long, fosmid end sequences come in pairs. It follows that combining the tetranucleotide signatures from the two ends into a single point may be approximately equivalent to clustering 1.5kb fragments, even though the two fosmid ends are not directly adjacent.

Clustering of 1.5kb fragments from the methanogens is depicted in Figure 1.5. This is rather fortuitous: whereas no boundaries were visible for the 700 bp fragments, several boundaries in the above graph can be seen. It would be ideal if two end sequences were sufficient for drawing community composition maps, independent of seed sequence-produced boundaries.

Unfortunately, 1.5kb does not seem to be sufficient to produce emergent clusters within the HF200 library: clustering of combined fosend tetranucleotide signatures from HF200 is shown in Figure 1.12. To more clearly see the topology, clear outliers evident from Figure 1.10 were removed. Reclustering of 700 bp fragments without the outliers selected from Figure 1.10 and listed in the text of Figure 1.11 is shown in Figure 1.11.

Since quality files were not available for any of the HF libraries, it seemed plausible that perhaps the combined fosmid end sequence clustering of the HF200 library was less successful than hoped because of noise from the "junk" portions of each sequence. Cleaned sequences (where portions containing quality scores lower than 20 were removed) are available for a more recent cruise (H179), which, from BLAST analysis to existing group II archaeal seed sequences, seems to contain a considerable portion of group II archaeal seed sequences. Though the H179 depth libraries are substantially larger than the HF200m library, after cleaning, only several thousand end sequences exceed 700 bp. Figures 1.13 and 1.14 show clustering of H179 125m libraries for combined sequences exceeding 1.5 kb and 1.4 kb, respectively. Boundaries here are still not nearly as clear as those seen for the methanogens.

Conclusion

While the effectiveness of using an ESOM to cluster 1.5 kb fragments from fully sequenced, cultivated methanogen genomes has been clearly demonstrated, the effectiveness of using an ESOM in choosing sequences to cluster from the DeLong lab HF libraries was inconclusive. Clustering of unknown end sequences to known group II archaeal genome fragment sequences using the ESOM yielded a list of 334 fosmid ends, of which only 129 were chosen for sequencing, based on a second filter: BLAST homology to euryarchaeal housekeeping genes, or to genes from existing putative group II archaeal contigs. This second filter was used because a single sequencing run could only handle about 130 fosmids. How many fosmids would belong to an intended group, chosen for sequencing *only* by their ESOM clustering with given seed sequences, is still not known. The ESOM shows great potential, however, for accurately clustering longer environmental sequences. The application of ESOM methods to investigate long (>20kb) fragments is detailed in Chapter 2.

References:

[1] Noble PA, Citek RW, Ogunseitan OA. Tetranucleotide frequencies in microbial genomes. Electrophoresis, 1998; 19:528-535.

[2] Mrazek J. Phylogenetic Signals in DNA Composition: Limitations and Prospects. Mol. Biol. Evol, 2009; 26: 1163-1169.

[3] Wang YW, Hill K, Singh S, Kari L. The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene, 2005; 346:173-185.

[4] Sandberg R, Branden CI, Ernberg I, Coster J. Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. Gene, 2003:35-42.

[5] Davenport CF, Wiehlmann L, Reva ON, Tummler B. Visualization of *Pseudomonas* genomic structure by abundant 8-14mer oligonucleotides. Environmental Microbiology, 2005; 11:1092-1104.

[6] Takahashi M, Kryukov K, Saitou N. Estimation of bacterial species phylogeny through oligonucleotide frequency distances. Genomics, 2009; 93:525-533.

[7] Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B. Metagenome Fragment Classification Using N-Mer Frequency Profiles. Advances in Bioinformatics, 2008; 205969.

[8] Bohlin J, Skjerve E, Ussery DW. Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. BMC Genomics, 2008; 9:104.

[9] Deschavanne P, Giron A, Vilain J, Dufraigne C, Fertil B. Genomic Signature is Preserved in Short DNA Fragments. Proc. of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering, 2000; 161-167.

[10] Rosen GL, Sokhansanj BA, Polikar R, Bruns MA, Russell J, Garbarine E. Signal Processing for Metagenomics: Extracting Information from the Soup. Current Genomics, 2009; 10:493-510.

[11] Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics, 2004; 5:163.

[12] Chan CK, Hsu AL, Tang S, Halgamuge SK. Using Growing Self-Organizing Maps to Improve the Binning Process in Environmental Whole-Genome Shotgun Sequencing. Journal of Biomedicine and Biotechnology, 2008; 513701.

[13] Nasser S, Breland A, Harris FC, Nicolescu M. A Fuzzy Classifier to Taxonomically Group DNA Fragments within a Metagenome. IEEE, 2008; 277-282.

[14] McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. Nature Methods, 2007; 4:63-72.

[15] Chan CK, Hsu AL, Halgamuge SK, Tang S. Binning sequences using very sparse labels within a metagenome. BMC Bioninformatics, 2008; 9:215.

[16] Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, et al. Community-wide analysis of microbial genome sequence signatures. Genome Biology, 2009; 10:R85.

[17] Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature, 2004; 428:37-43

[18] DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU et al. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. Science, 2006; 311:496-503.

Chapter 2.

Using Emergent Self-Organizing Maps to Investigate the Community Structure of the Marine Group II Archaea and their Provenance in the North Pacific Subtropical Gyre Water Column.

Introduction

Marine group II *Archaea* were discovered in phylogenetic surveys in ocean waters about a decade ago. To date, little is known about either their function or their provenance, with the following exceptions. Several marine group II *Archaea* in the Hawaii Ocean Time-Series (HOT) waters have acquired a proteorhodopsin [1], most likely by horizontal gene transfer, since the proteorhodopsin bears closer homology to rhodopsins from the proteobacteria than to any known archaeal rhodopsin. Marine group II *Archaea* have been found in ocean waters, both open and coastal throughout the world. [2, NCBI 16S libraries].

Little is known about the distribution of the marine group II *Archaea* within those oceans, except that this group is not limited to one depth – they are found throughout the water column [3], from the surface to the deep, unlike the marine group I *Crenarchaea* which dominate and are found exclusively in the deep oceans.

About the genome itself, little is known except that the 16S and 23S ribosomal RNA operons are no linked in all the group II archaeal fragments examined to date, something usually characteristic of eukaryotes, but found in some other *Archaea* as well. The closest relatives of the group II *Euryarchaea* are the *Thermoplasma*, found in sulfur hot springs. How these two, inhabit such disparate environments yet are each other's closest relatives is not understood.

This chapter investigates the emergent clusters of those organisms within the group II *Archaea*, and the distribution of those clusters within the water column near station ALOHA at the Hawaii Ocean Time-series (HOT) site. The data used for this chapter came from 4 sources [4]:

1. The assembled contigs from the HF200 fosmids sequenced, as detailed in chapter 1.

2. The end sequences of the HF (station ALOHA, HOT site) cruise fosmid libraries. Seawater collection filtration dates:

10m – 500m: Oct. 6-7 2002

770m, 4000m: Dec.21 2003

3. The end sequences of the HOT179 fosmid library. Cruise date: 08 March – 12 March 2006

4. The end sequences of the HOT186 fosmid library. Cruise date: HOT186: 18 Oct – 24 Oct 2006.

Methods

The same bioinformatics methods used in Chapter 1 are used again in Chapter 2. See methods section of Chapter 1 for details.

Results and Discussion

A tree of 134 16S rRNA sequences, collected during the HF cruise, is available in Figure S6 in DeLong et al (2006) [3]. It would be ideal if this phylogenetic information could be paired with functional information. This chapter takes a step in that direction by looking at the naturally emerging groups among long (>20kb) genomic fragments, and mapping their distribution in the HOT water column during the three cruises listed above. Some of the emergent clusters can be mapped directly to phylogenetic clusters. For others, additional work will be required to make such assignments.

In order to investigate the genomic content of the marine group II *Archaea*, it was first necessary to determine which fosmids were, and which were not marine group II *Archaea*. Based on the clear reliability of ESOM clustering for 20kb fragments, both in chapter 1, and in the bioinformatics literature, I chose to limit my clustering to those I could cluster with certainty. In chapter 1, it was sufficient to look for sequences that were very likely group II *Archaea*, and for this, shorter sequences were sufficient. But when looking for metabolic pathways or for the distribution of phylogenetic clusters, sequences incorrectly assigned to the marine group II *Archaea* would act as noise and could result in erroneous conclusions. So I chose to eliminate noise, at the cost of eliminating some available signal—short sequences that are group II *Archaea*, but whose identity as group II *Archaea* cannot be verified.

Figure 2.1 shows the clustering of all assembled contigs over 20kb, assembled from the 454 sequence run of fosmids pulled from the HF200 library, as described in Chapter 1. Of the 99 contigs over 20 kb retrieved, 23 cluster far away from the other sequences. A total of 22 of these 23 sequences do not show end homology to *any* of the sequences in the HF200 library. It is important to understand where these contigs that are clearly *not* of group II archaeal origin came from, in order to measure the reliability of the method used to choose fosmids for sequencing. In the list of fosmid names, whose end sequences show homology to any of the assembled contigs over 10 kb (List 1.2, Appendix), there is a statistically significant, nonrandom gap of 26 fosmids all from the same physical library location. These fosmids are assigend the set of names from ASNG3084 to ASNG3403, inclusive. These fosmids represent all of and the only ones picked from 96 well microtiter plates 90,91,92, and 93. The most likely explanation for the error seen is either that the colonies containing these fosmids were picked incorrectly, or that there is an error in the perl script converting the fomid names beginning with "ASNG" to physical plates or coordinates. These fosmids were not re-sequenced before the 454 sequencing run. Generally it is advisable to end-sequence fosmids chosen for sequencing, to ensure that the fosmids sequenced correspond to the correct plate and well coordinates that were identified from the fosmid end sequences.

Once the 23 clear outliers were removed from the pool of putative marine group II archaeal sequences, all the tetranucleotide signatures of the marine group II *Archaea* fosmid sequences from NCBI (both those containing a 16S rRNA operon, and those with either high blast homology to HF130 contigs, or reasonable proximity to known group II archaeal sequences when clustered with an ESOM) and putative marine group II archaeal sequences over 20kb (from HF130 and HF200) were re-clustered (Figure 2.2). An additional 6 outliers were identified, 2 unclassified marine microorganism sequences from NCBI, and 4 contigs from HF200.

The simplest, and most standard, bioinformatic analysis of genomic sequence is to look at its GC content. The results of this clustering are shown in Figure 2.3. Two clusters emerge clearly: a high-GC cluster, centered at about 50% GC, and a low GC cluster, centered at about 58% GC. Of the remaining sequences, the one furthest from the others contained a 16S group II archaeal rRNA operon, strongly suggesting that all the remaining sequences are indeed marine group II *Euryarchaea.*

Blast analysis largely agreed with the assignment of the remaining sequences to the *Euryarchaea.* Looking at the assignment of HF200 contigs, all but 29 contigs (47 of 76 contigs) could be assigned by BLAST results alone. Sequences of these 47 contigs had highest similarity to euryarchaeal ribosomal proteins, DNA and/or RNA polymerases, transcription factors or translation initiation factors, helicases, and DNA primase. In addition to these 47 sequences, another 6 had blastx homology to euryarchaeal tRNA synthetases. Yet the validity of such a classification is not definitive, since some tRNA synthetases, whose closest homologs were *not* euryarchaeal were found on contigs containing ribosomal proteins that clearly marked those contigs as euryarchaeal.

Contigs *not* clearly assigned to the marine *Euryarchaea* via BLAST, but clearly clustering on the SOM with those that *were* assigned to the marine euryarchaea were: 16, 17, 64, 66, 76, 105, 110, 113, 131, 151, 164, 291, 297, 341, 369, 397, 415, 418, 424, 451, 465, 480, 507, 531, 533, 597, 598 and 625. Each one of these contigs contained at least one gene (ORF calling determined with MetaGene [5]) showing closest blastx homology to a euryarchaeal gene, but the nature of those genes—often one hypothetical protein flanked by genes with highest sequence similarity to those from bacteria —was insufficient for unambiguously calling the contigs as euryarchaeal, based on blast information alone. Inspection of Figure 2.4 shows that these sequences clearly cluster with other sequences – they do not sort off by themselves. A clear case for using an ESOM to collect long sequence fragments belonging to a given phylogenetic group is thus presented: the SOM can cluster far more sequence fragments than a clustering by BLAST alone can achieve.

Not only are more sequences given a phylogenetic identity, the phylogenetic structure within the broad group of the marine group II *Archaea* can be visually seen. Figures 2.4, 2.5 and 2.6 show the 6 clusters with substantial representation among the sequences to date. Figure 2.6 shows the cluster made up only of sequences from HOT 4000m or deep Antarctic waters. While clusters 1-4 fall into either the high or low GC cluster—that is the ESOM clustering produces a higher-resolution clustering, confirming the GC clustering, but adding additional sub-structure, Cluster 5 (by far the cluster with the largest total representation) contains a few stray high-GC sequences. Re-clustering just Cluster 5 sequences (Figure 2.5), reveals 7 sequences that do not cluster tightly with the remaining Cluster 5 contigs. Of interest here is that all nearly all of the high GC fosmids in Cluster 5 (all except 4003694_fasta.screen.Contig4) do not belong to the bulk of the Cluster 5 sequences. With these 7 removed, only one high-GC sequence falls into the Cluster 5 core, confirming that clusters produced by the ESOM fall into either one GC cluster or the other.

This conclusion is supported by most all the data, even though there is a stray sequence in Cluster 5. Note that the boundary of 55% GC is not an exact boundary: the peaks appear Gaussian, the GC values fall in a distribution around the mean, not exactly at the mean. Thus *most* of the sequences less than 55% GC belong to the low GC clade, but there may be a few sequences whose values are slightly over the boundary. The same argument applies to the high GC clade.

How do these putative group II *Archaea* cluster, identified by SOM methods overlay onto the phylogenetic tree in Figure S6 of [3]? Only 5 HF130 and HF200 contigs over 20kb contain either a 16S or 23S rDNA operon:

23S: 4003651_Contig2, 4003696_Contig2, contig00526
16S: 4003658_Contig1, contig597

This makes for 5 ribosomal genes, in addition to the existing 19 16S rRNA gene-containing fosmids found on NCBI, for a total of 24 rRNA containing contigs/fosmids out of 134 total putative group II fosmids/contigs to date. The sampling rate of 5 out of 120 is roughly what one would expect when sampling the group II archaeal metagenome randomly: the 16S rRNA gene is about 1500bp long; the 23S gene about 3kb. Together their length comes to about 4.5kb. 120 contigs at a little over 30kb each together fill about 4Mbp of sequence. Assuming a genome size of about 2Mbp, the group II archaeal metagenome has been sampled about twice. If the 16S and 23S operons are joined together, about 2% of the contigs should have rRNA genes, but since for the marine group II *Archaea* the operons are split, 4% of the contigs are expected to have rRNA content—which is about what is seen (5/120 = 4.2%). The 16S rRNA operon on contig 597 is right at the end of the contig; as a result only about half of the contig is present. Since no other 23S rDNA operons have been sequenced for the marine group II *Archaea*, the 23S rRNA sequences will no doubt prove useful for future phylogentic studies. 4003658_Contig1 shows the closest homology to the 16S rRNA sequence HF130_40B02. Contig597 shows the closest homology to HF200_63E02. Both of the contigs are in Cluster 5. Thus, Cluster 5 most likely fits with the topmost cluster in Figure S6 of [3]. The phylogeny of the remaining clusters is less clear.

An interesting aside is that the 6 clusters formed by ESOM clustering fall under either one or the other of the two GC % bins in Figure 2.2. While some phylogenetic structure can be resolved by looking at GC content, clustering tetranucleotide signatures gives a much higher-resolution glimpse into the structure of a microbial community. List 2.2 (Appendix), summarized in Table 2.1, shows the GC content of each fosmid in each of the group II rRNA clusters. Sequence names in blue fall below the GC cutoff of 55% GC; sequence names in red fall into the high-GC cluster.

Six high-level clusters (well above the level of species) for the marine group II *Archaea* are now available. The natural questions are then, how are these clusters distributed within the water column, and what do they do. The remainder of this thesis discusses the distribution the group II archaeal sequence clusters within the water column of the Centroal North Pacific Gyre. Interesting hints are available as to the function of at least some of the group II archaea: several genes in the 3-hydroxypropionate/4-hydroxybutyrate cycle, a carbon fixation cycle in archaea are present. Whether the whole cycle, or some variant thereof, is present in the group II *Archaea* will require additional sequencing.

In addition to the fosmid libraries for the HF cruise, much larger fosmid libraries are available for both the HOT179 and HOT186 cruises are end sequenced and stored in the DeLong lab. Presence of the marine group II *Archaea* within those libraries, as a function depth is shown in Figure 2.7. While both the HF and HOT179 datasets contain a group II archaeal bloom in the photic zone (>= 200m) waters, *no* end sequences in the HOT186 surface fosmid libraries show substantial blastn homology to any group II archaeal sequences. This is very odd, indeed, but by all measures used to look for group II archaeal sequence types, including looking for blastx homology to any existing group II archaeal fosmid, group II archaea appear all but absent from the photic zone at the time the HOT186 samples were taken. They are of course likely present in very low abundance, but below the detection limits of our survey and sequencing efforts.

Figure 2.8 shows the distribution of the 6 group II archaeal clusters within each fosmid library. It is interesting to note that while Cluster 5 is present in both HF130 and HF200 libraries, HF130 shows a strong preference for those sequences within Cluster 5 that originally came from HF130, and likewise for HF200. That is, there is a depth dependent organization of the species within clusters. It should be noted that the current pool of group II archaeal sequences is heavily biased towards the 130m and 200m depths from the HF library; it thus follows that the actual fold change in group II population numbers from the very shallow depths (<100m) to the deeper photic zone waters is certainly exaggerated. Moreover, clusters for the marine group II were only defined if sufficiently many sequences fell into a clear ESOM cluster. It seems likely therefore that there may be several clusters confined to very shallow or very deep waters that were not defined, because they currently lack sufficient representation. Of interest to note as well is that when group II *Archaea* were observed in high abundance at depth, corresponding 500 m sequences consisted entirely of cluster 4 sequences, In contrast, when the bloom was absent (in HOT186), there was considerably more diversity in the sequences at 500 m.

For future work, additional sequencing might most naturally begin with all those fosmids whose end-sequences show homology to Cluster 5. That cluster appears to dominate in the water column, and is present throughout the surface waters.

References

[1] Frigaard N, Martinez A, Mincer TJ, DeLong EF. Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. Nature, 2006; 439: 847-850.

[2] Martin-Cuadrado A, Rodrigues-Valera F, Moreira D, Alba JC, Ivars-Martinez E, Henn M. et al. Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. ISME Journal, 2008; 2:865-886.

[3] DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N et al. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. Science, 2006; 311: 496-503.

[4] HOT website: http://hahana.soest.hawaii.edu/hot/hot_jgofs.html

[5] Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Research, 2006; 34:5623-5630.

Conclusion

The utility of clustering tetranucleotide signatures via an emergent self-organizing map for mining fosend libraries for specific genome fragments of interest was investigate, targeting in particular marine group II archaeal genome fragments. While calibration using whole-genome sequences from cultured methanogens indicated that 1.5kb of sequence—approximately that available when both fosmid end sequences are taken together—was sufficient for finding species-level emergent groups within datasets, even without the use of seed sequences, similar topographical boundaries were not seen after clustering HF200 fosmid library tetranucleotide signatures of fosmid end sequences.

Clustering methanogen tetranucleotide signatures with lengths typical of those for single fosmid end-sequence (~700bp fragments) show that sequences from the same species cluster together, though the boundaries are not self-emergent: seed sequences are required to draw statistically significant boundaries between species. This lends credence to the hypothesis that using seed sequences to cluster either single end-sequence tetranucleotide signatures or joined end-sequence tetranucleotide signatures of fosmids from the Hawaii Ocean Time-Series fosmid libraries can identify fosmids belonging to groups of interest. This method was used to select 334 possibly group II archaeal sequences from the HF200 library. A single 454 sequencing run could only accomodate approximately 130 individual fosmids, however, so this the end-sequences on this list were sent through a second filter: blast homology to existing euryarchaeal housekeeping genes, or to genes on existing group II archaeal contigs—those containing a 16S rRNA gene, or directly contiguous with such a fragment. Among those contigs assembled from 454 sequence, ignoring the clear outliers that seem to be an error in picking or converting sequence names to plate names, only 4 of 76 seem to not be group II *Archaea* : a success rate of about 95% for the procedure detailed in Chapter 1. If all 334 fosmids had been sequenced, therefore, the success rate could not have been worse (after subtracting error not related to the bioinformatics selection procedure) than 22% (72/334 = 22%). But whether the efficiency of ESOM clustering, used alone as a selection method, would yield a success rate closer to 22% or closer to 95% has yet to be investigated.

What was confirmed, in agreement with the bioinformatics literature, was the validity and usefulness of ESOM clustering to cluster sequences over 20kb reliably, and possibly even to the species level. Among those contigs assembled from 454 sequence, and those previously sequenced by traditional transposon insertion methods from HF130, definite community structure within the marine group II archaeal sequences was visible, a structure much more rich than that accessible by clustering GC content, and capturing more approximately 50% more sequences than blast homology alone could assign to the marine group II *Archaea*, substantially augmenting the pool of interesting genes waiting to be examined, in order to tease out the marine group II archaeal functional properties and capabilities. This is among the clear next steps to extend the work in this thesis.

Perhaps the most interesting product of this thesis was the discovery of a very cost-effective way to sequence a large (~100) group of fosmids. It has been shown that a single 454 titanium run with sequence prepared from 129 fosmids as input, can produce an output of sequences from which a large number of contigs can be re-assembled – 99 contigs over 20kb in this case. A majority of these contigs can be assigned to original end-sequences, allowing for deeper sequencing into the same library, while nearly eliminating the concern of re-sequencing the same fosmids.

It has been shown that clustering tetranucleotide signatures of long genomic fragments—greater than about 20 kb—is both an effective and important way of assigning fragments to novel prokaryotic groups, specifically the marine group II *Archaea*. The information accessed from such clusterings can resolve between species, in the full genomes sequenced to date, and is an important tool in metagenomics for pooling long fragments into phylogenetic bins, especially in those environments where few cultured representatives are available.

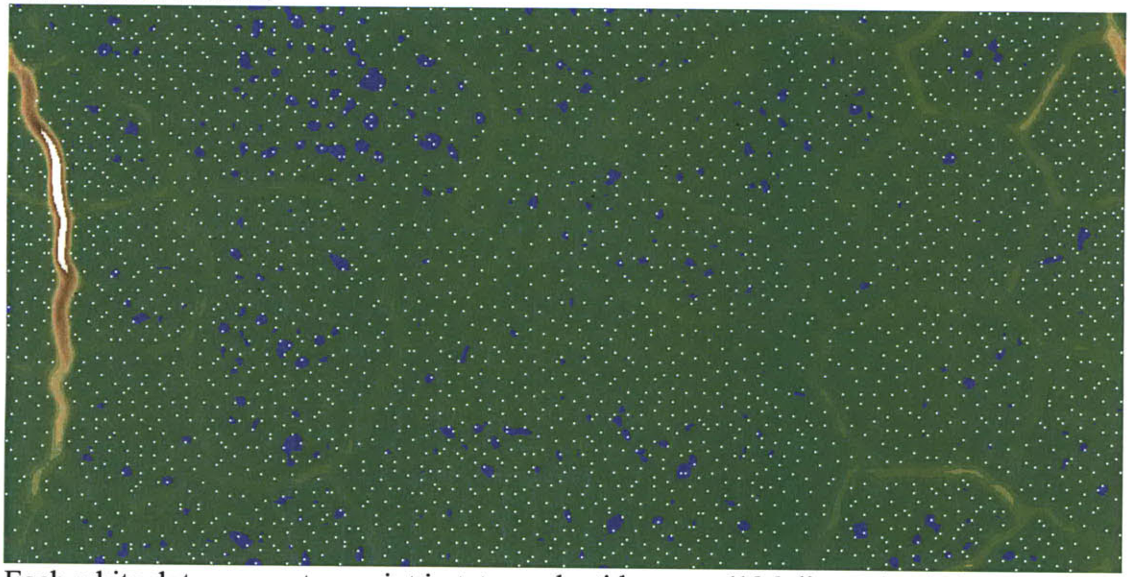
The work in this thesis was based on one of the oldest microbial habitats known—the open ocean. It is a place where sampling may be simpler than, for example, in soil, but the communities themselves in the ocean are far from simple. Yet even in such a truly complex environment—a tough test—clustering tetranucleotide signatures with emergent self-organizing maps has been demonstrated a valuable tool, worth continued use in the context of marine metagenomics.

Acknowledgements

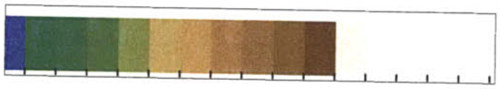
The author would like to acknowledge the work of A. Martinez and T. Palden in performing all of the wetlab work detailed in Chapter 1—picking colonies, regrowing them, extracting fosmid DNA, purifying it and preparing it for sequencing. Many thanks to J. Eppley for computational assistance, and to Y. Shi and J. McCarren and A. Martinez for helpful discussions. And special thanks to my advisor, EF DeLong, for his guidance, encouragement and support.

Figures

Figure 1.1. Clustering of 20kb fragments from NCBI methanogens on the Databionics ESOM.

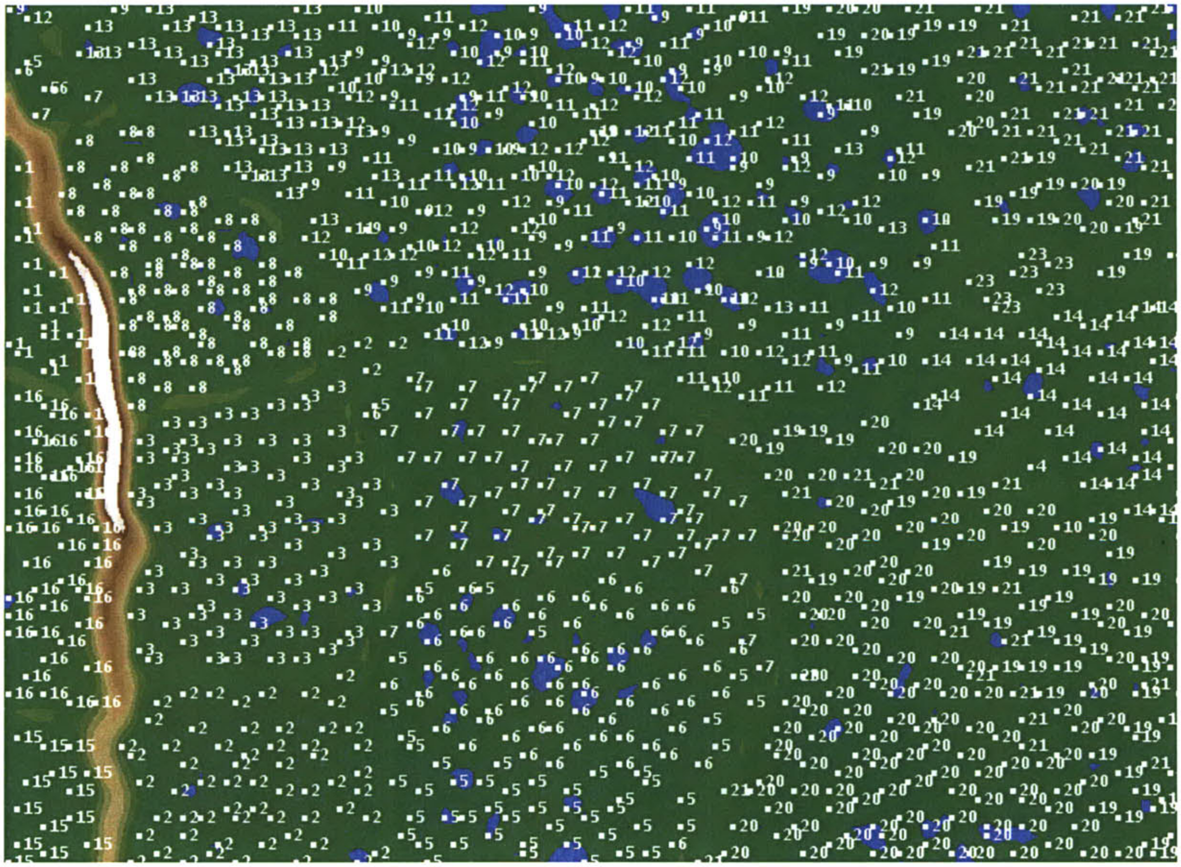


Each white dot represents a point in tetranucleotide space (136 dimensions) for 20 kb fragments of cultured, sequenced methanogenic genomes, downloaded from NCBI. The fragments are non-overlapping, sequentially taken fragments, without gaps. The map reads as a topographic map: blue means very little distance in tetranucleotide space, to white being a large distance between points. Clustering can be performed by eye: sequences all in the same valley, with a hill-like, or mountain-like ridge separating those points from the rest of the map indicates a cluster.



Key: *Small Distance* *Large Distance*

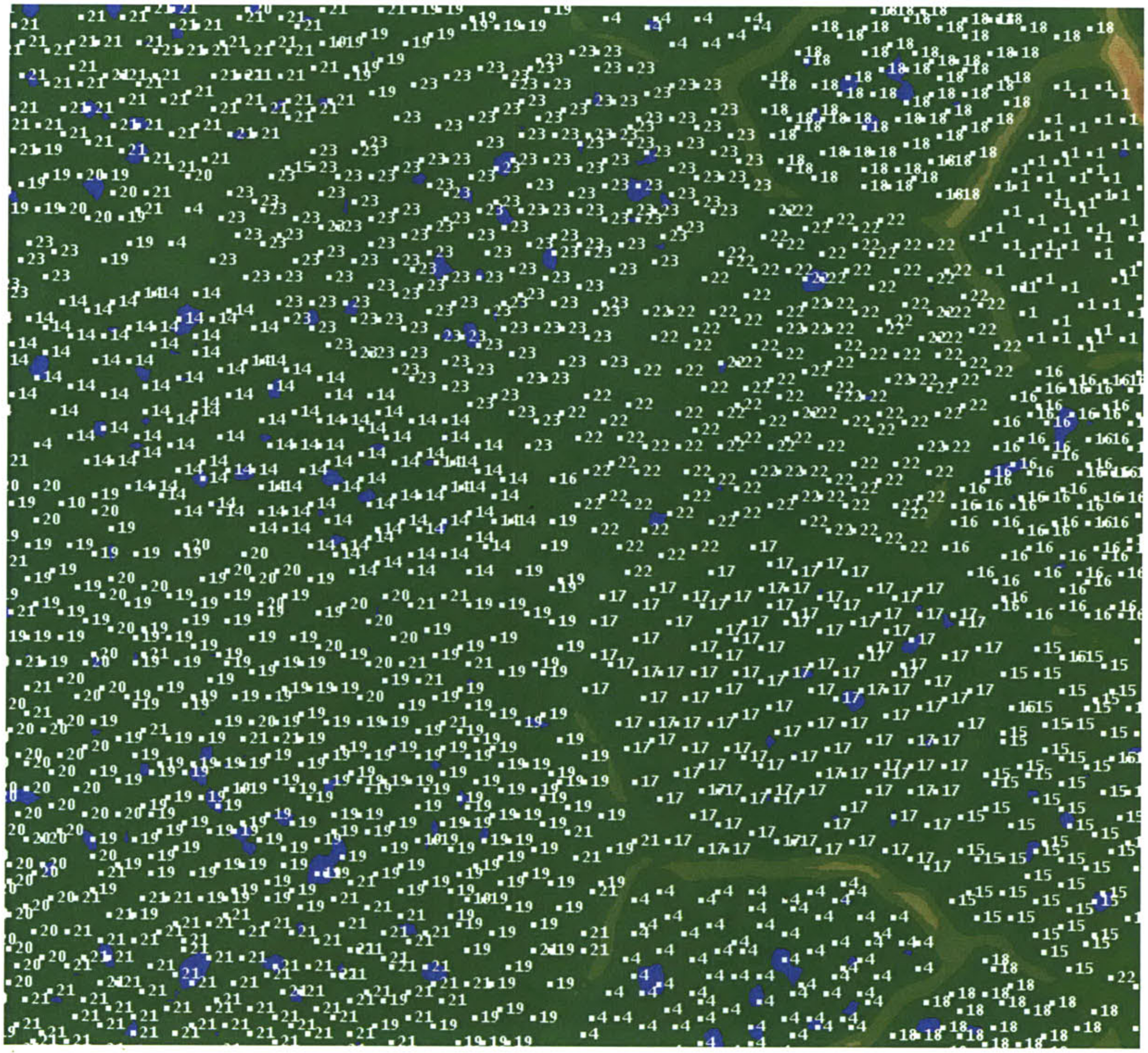
The map is automatically scaled so that the largest distance between two points appears white, the smallest blue. A word about the topology of the map: the ESOM is a borderless map, meaning that the left and right sides are the same line (traveling left to right, just as you pass the right border, you find yourself just inside the left edge). Similarly, the top and bottom edges are the same, and as a corollary of the above two facts, the four corners are all the same point. Topologically, a borderless 2D map exists on the surface of a donut. To visualize this, imagine constructing this donut, roll the map around the horizontal axis, connecting the top and bottom surfaces to form a tube. The cross section on the right end was the right hand side of the image; similarly with the left end cross-section. Now take these two circular ends and connect them to form a donut. This is the natural topology of a borderless 2D map.



The above image is the left hand side of the above map, zoomed in, with labels. Fragments from each methanogen genome are labeled as follows:

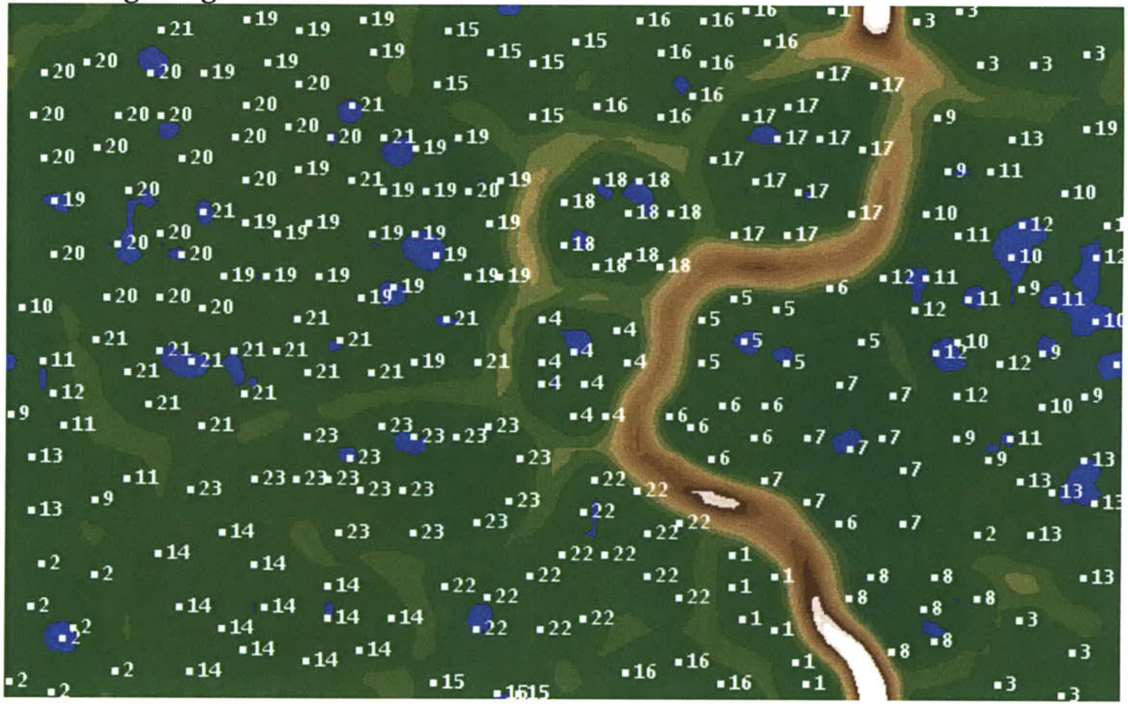
1 gi|20093440|ref|NC_003551.1| Methanopyrus kandleri AV19

2 gi|148642060|ref|NC_009515.1| Methanobrevibacter smithii ATCC 35061
3 gi|84488831|ref|NC_007681.1| Methanosphaera stadtmanae DSM 3091
4 gi|6626257|gb|AE000666.1| Methanothermobacter thermautotrophicus str. Delta H
5 gi|256809973|ref|NC_013156.1| Methanocaldococcus fervens AG86
6 gi|15668172|ref|NC_000909.1| Methanocaldococcus jannaschii DSM 2661
7 gi|261402131|ref|NC_013407.1| Methanocaldococcus vulcanius M7
8 gi|150400439|ref|NC_009635.1| Methanococcus aeolicus Nankai-3
9 gi|134045046|ref|NC_009135.1| Methanococcus maripaludis C5
10 gi|159904396|ref|NC_009975.1| Methanococcus maripaludis C6
11 gi|150401930|ref|NC_009637.1| Methanococcus maripaludis C7
12 gi|45357563|ref|NC_005791.1| Methanococcus maripaludis S2
13 gi|150398760|ref|NC_009634.1| Methanococcus vannielii SB
14 gi|91772082|ref|NC_007955.1| Methanococcoides burtonii DSM 6242
15 gi|124484829|ref|NC_008942.1| Methanocorpusculum labreanum Z
16 gi|126177952|ref|NC_009051.1| Methanoculleus marisnigri JR1
17 gi|154149549|ref|NC_009712.1| Candidatus Methanoregula boonei 6A8
18 gi|116753325|ref|NC_008553.1| Methanosaeta thermophila PT
19 gi|20088899|ref|NC_003552.1| Methanosarcina acetivorans C2A
20 gi|73667559|ref|NC_007355.1| Methanosarcina barkeri str. fusaro chromosome 1
21 gi|21226102|ref|NC_003901.1| Methanosarcina mazei Go1
22 gi|219850687|ref|NC_011832.1| Candidatus Methanosphaerula palustris E1-9c
23 gi|88601322|ref|NC_007796.1| Methanospirillum hungatei



Above is the right hand side of the original map, zoomed in, with labels. Note that in many cases, most of the fragments from the same genome all fall within the same valley, with either a lighter-green or more brown-looking convex polygon boundary separating those sequences from the remainder of the sequences on the map. Note that for the four different strains of *Methanococcus maripaludis*, sequences from all four genomes lie jumbled together in the same valley, as expected—tetranucleotide clustering in the literature is reported as a species-specific signature. Since these four genomes represent the same species, it is not expected that they could be resolved from each other using a tetranucleotide method.

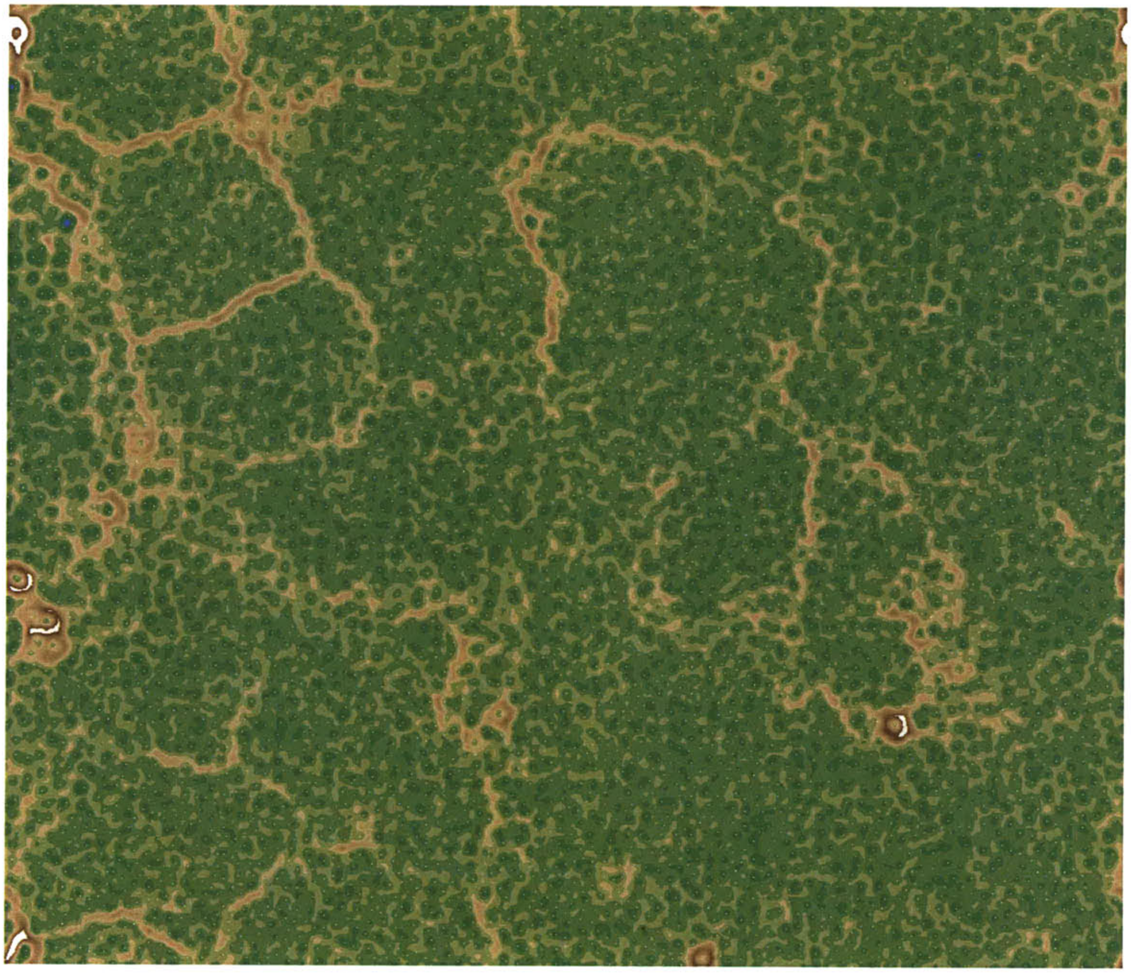
Figure 1.2. Clustering of a sparse selection of fragments from cultured methanogenic genomes.



1 out of every 10 fragments was chosen in a linear manner (choose the first encountered along the genome, skip 9, choose the next, etc.) for each genome; clustering on the ESOM was then performed.

Note that the patterns of clustering observed in Figure 1 hold true in the above image.

Figure 1.3. ESOM clustering of 3kb fragments of cultured methanogic genomes.

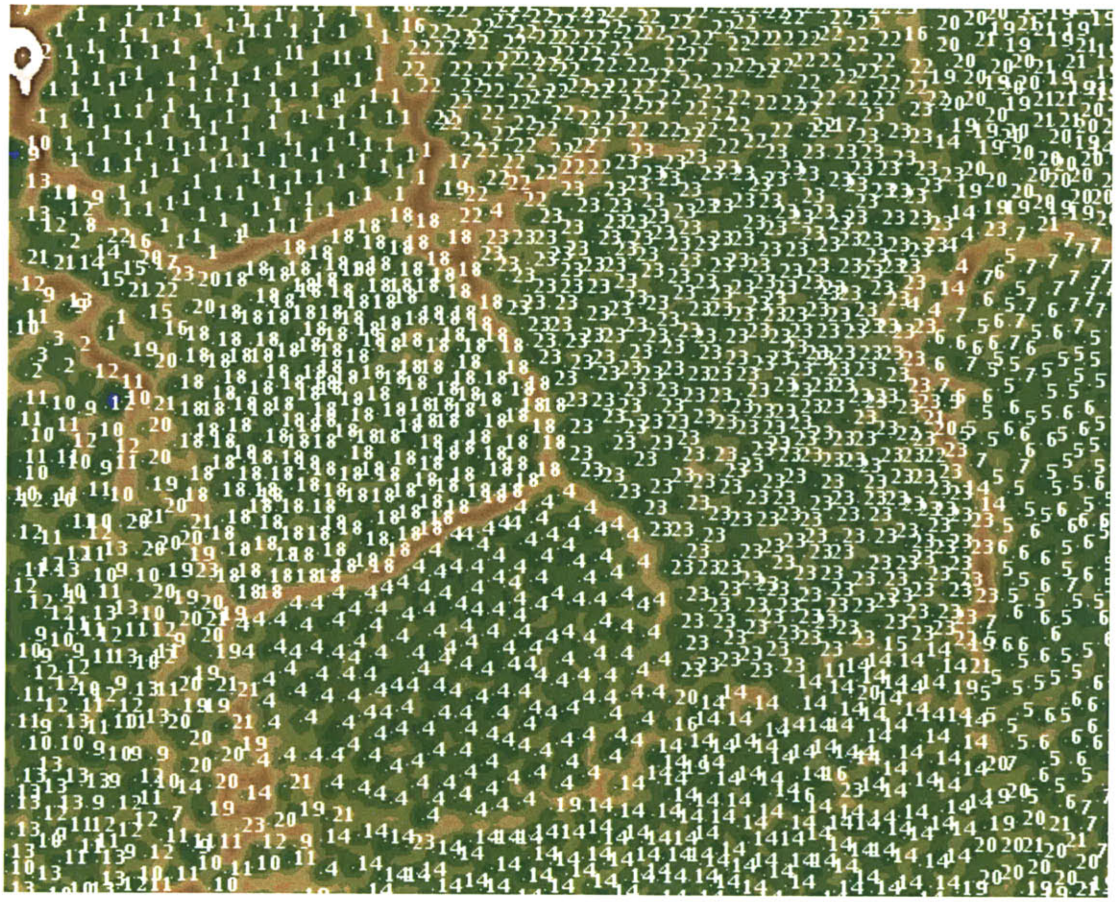


Clustering of 3 kb fragments. Below are the zoomed-in regions, with labels:

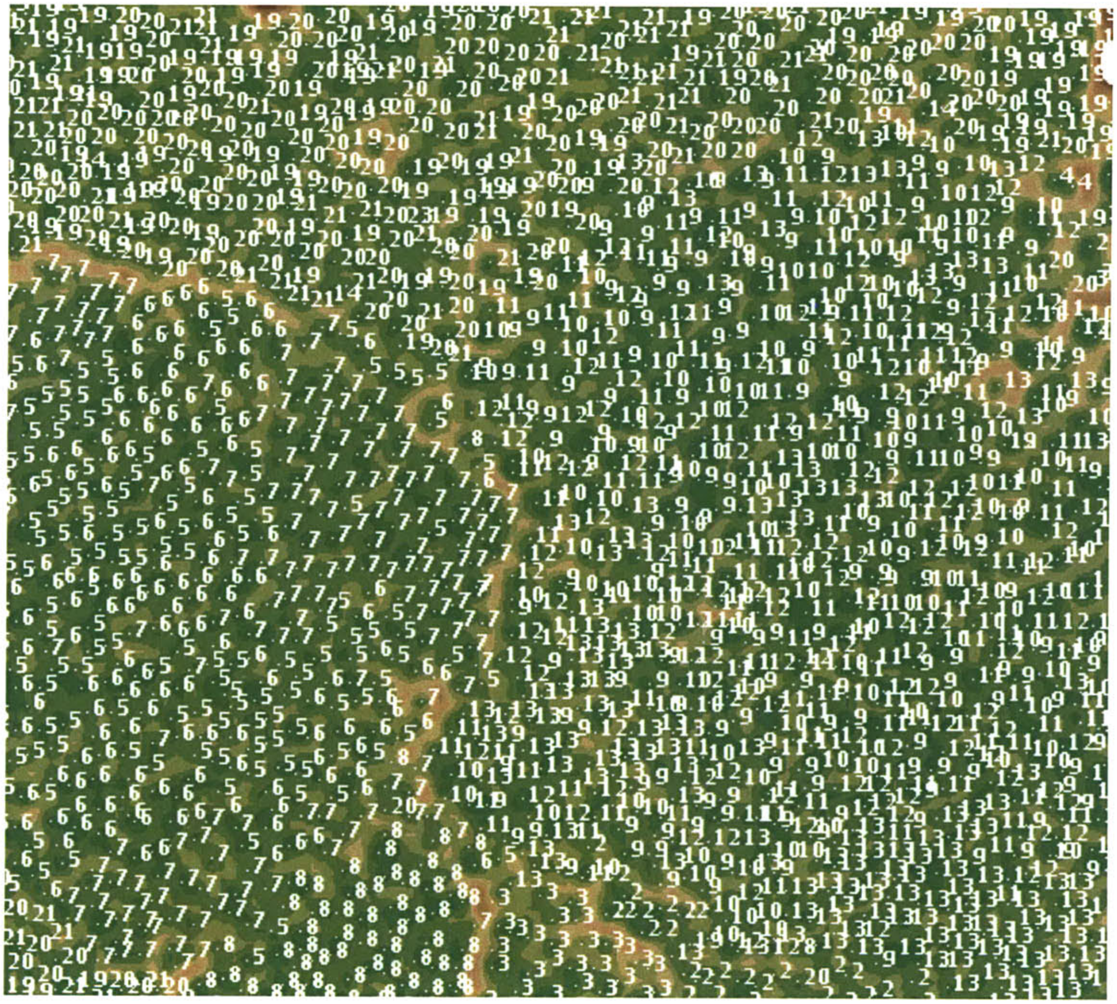
A. Upper left
B. Upper right
C. Lower right
D. Lower left

To reduce the computational load, only 1 out of every 3 fragments was selected for this clustering. N.B.: the clustering seen above still holds in most cases. More sequences are found off in the mountains by themselves, or broken away from the other sequences taken from the same genomes, but in most cases, the sequences still cluster reliably and well.

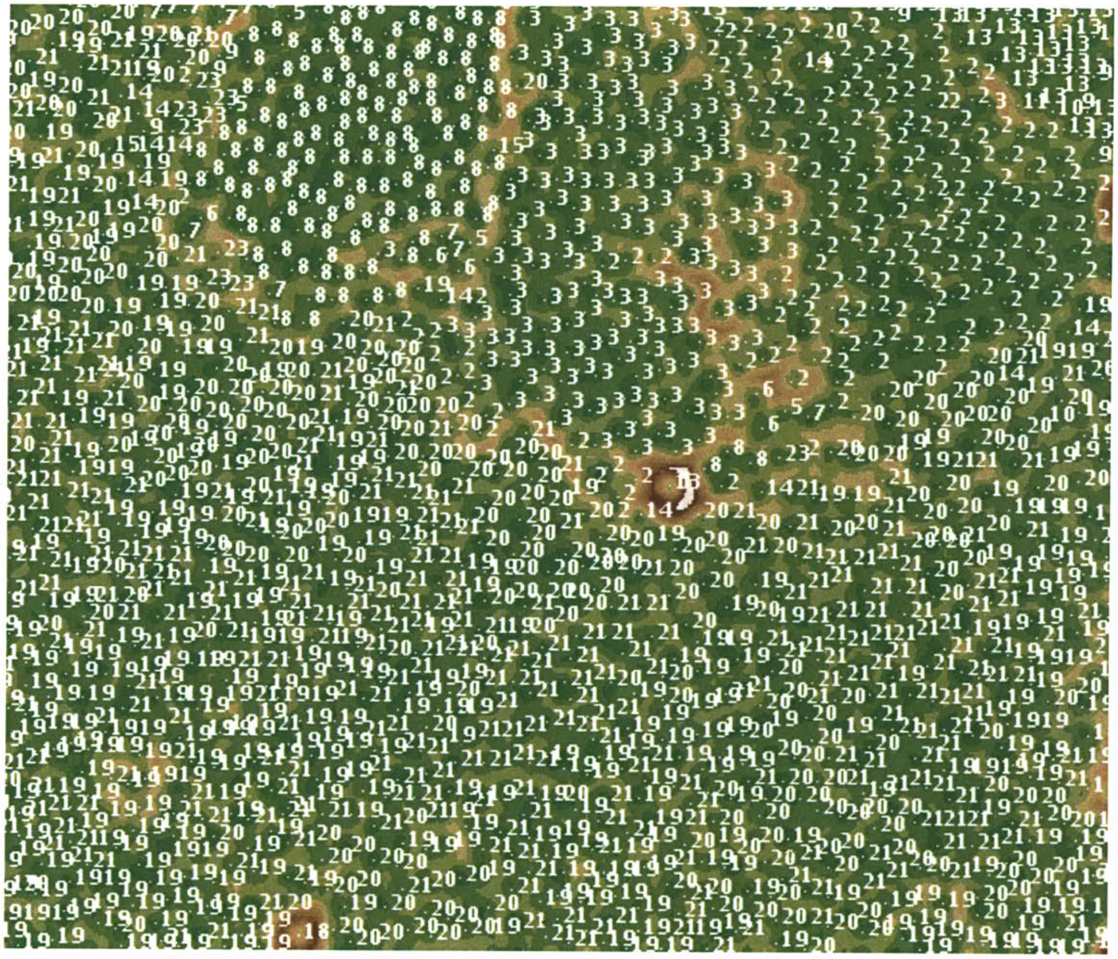
A. Upper left with labels



B. Upper right with labels



C. Lower right with labels



D. Lower left with labels

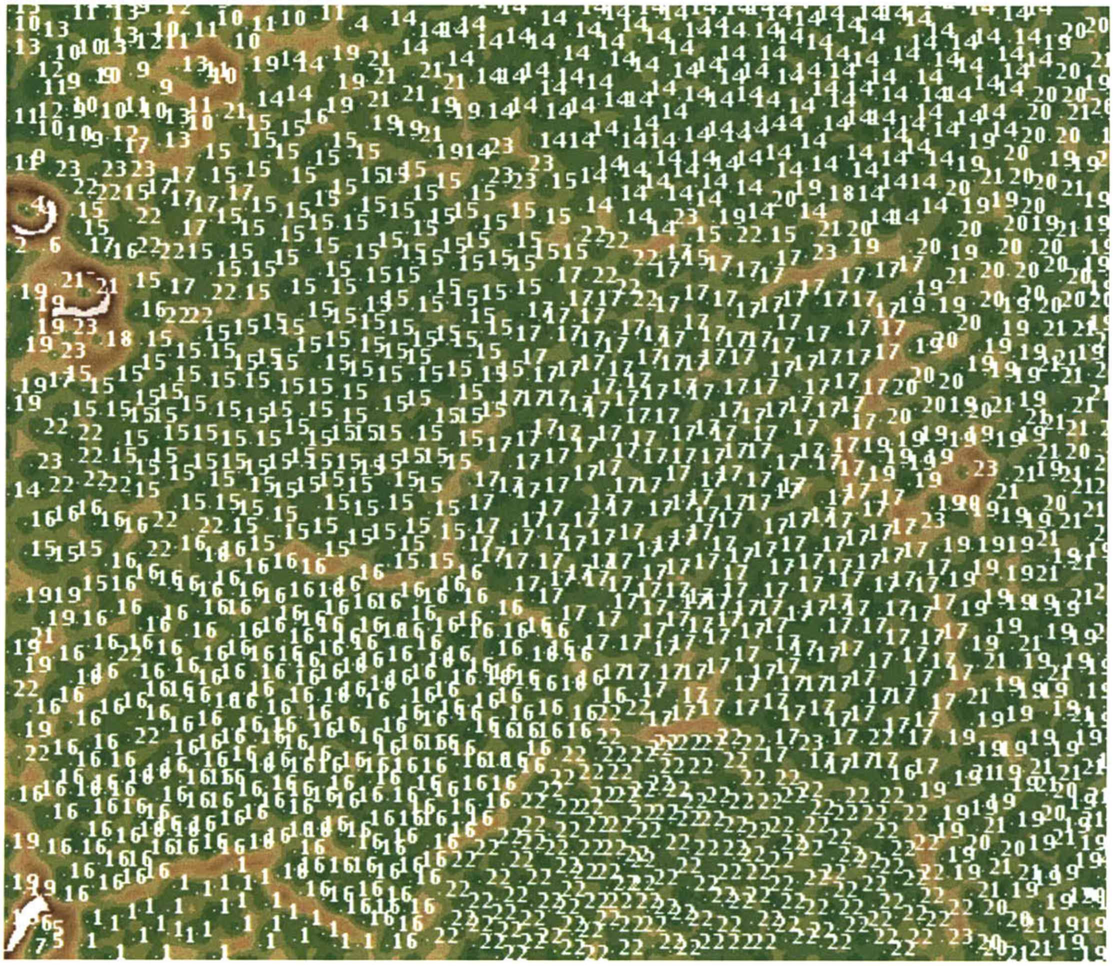
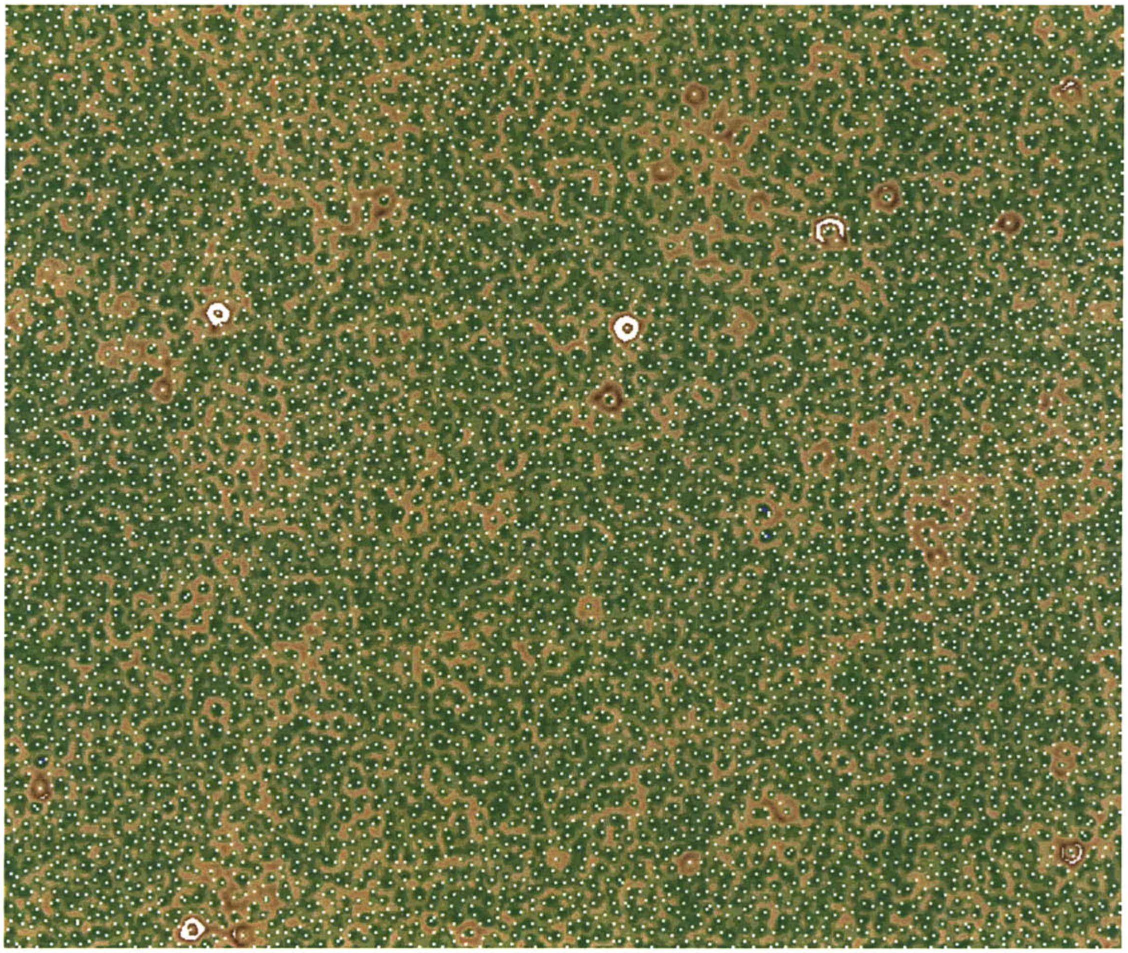
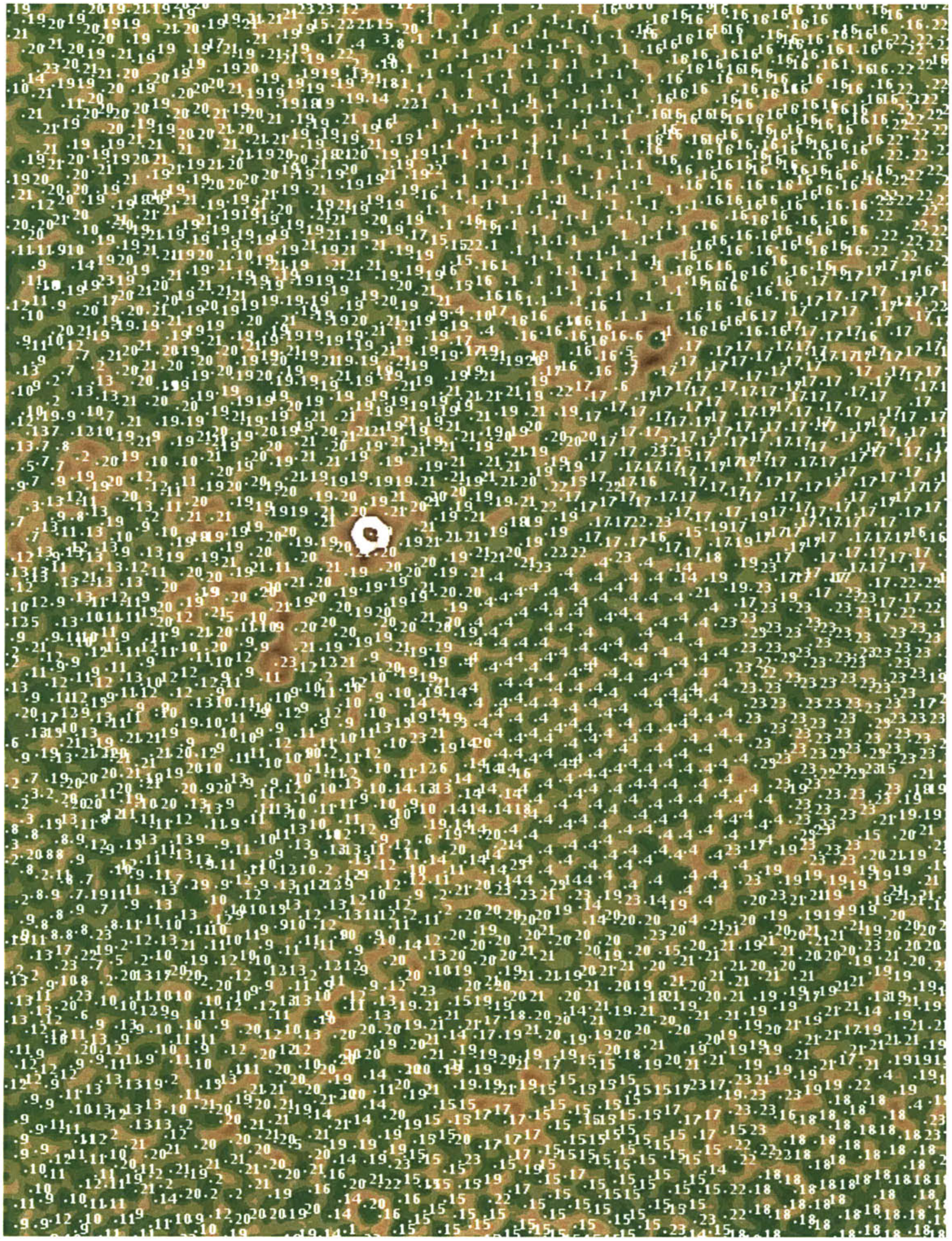


Figure 1.4. Clustering 700bp fragments from cultured methanogenic genomes.

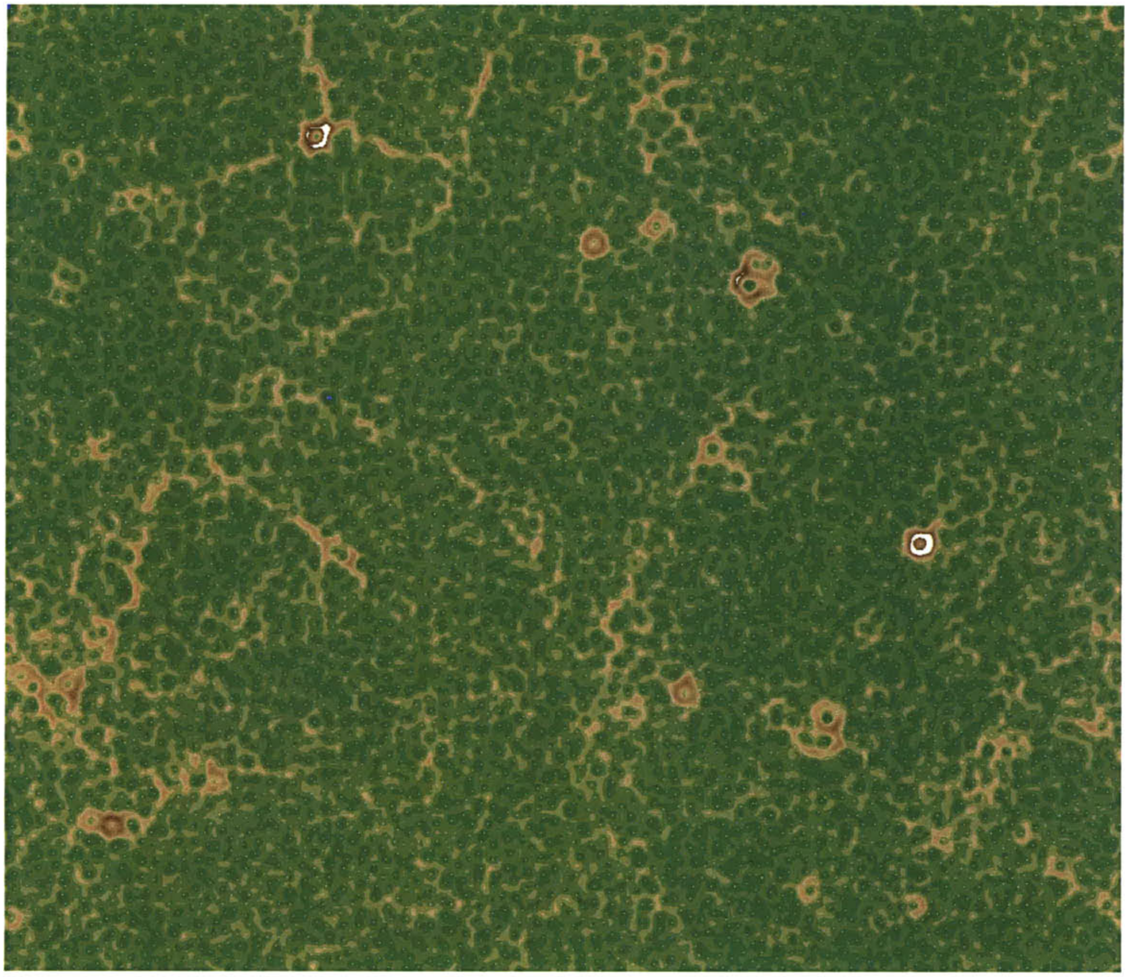


ESOM clustering of 1 in every 10 700 bp-long fragments from cultured methanogenic genomes. Clear boundaries segmenting the map into separate, closed polygons are no longer visually apparent.

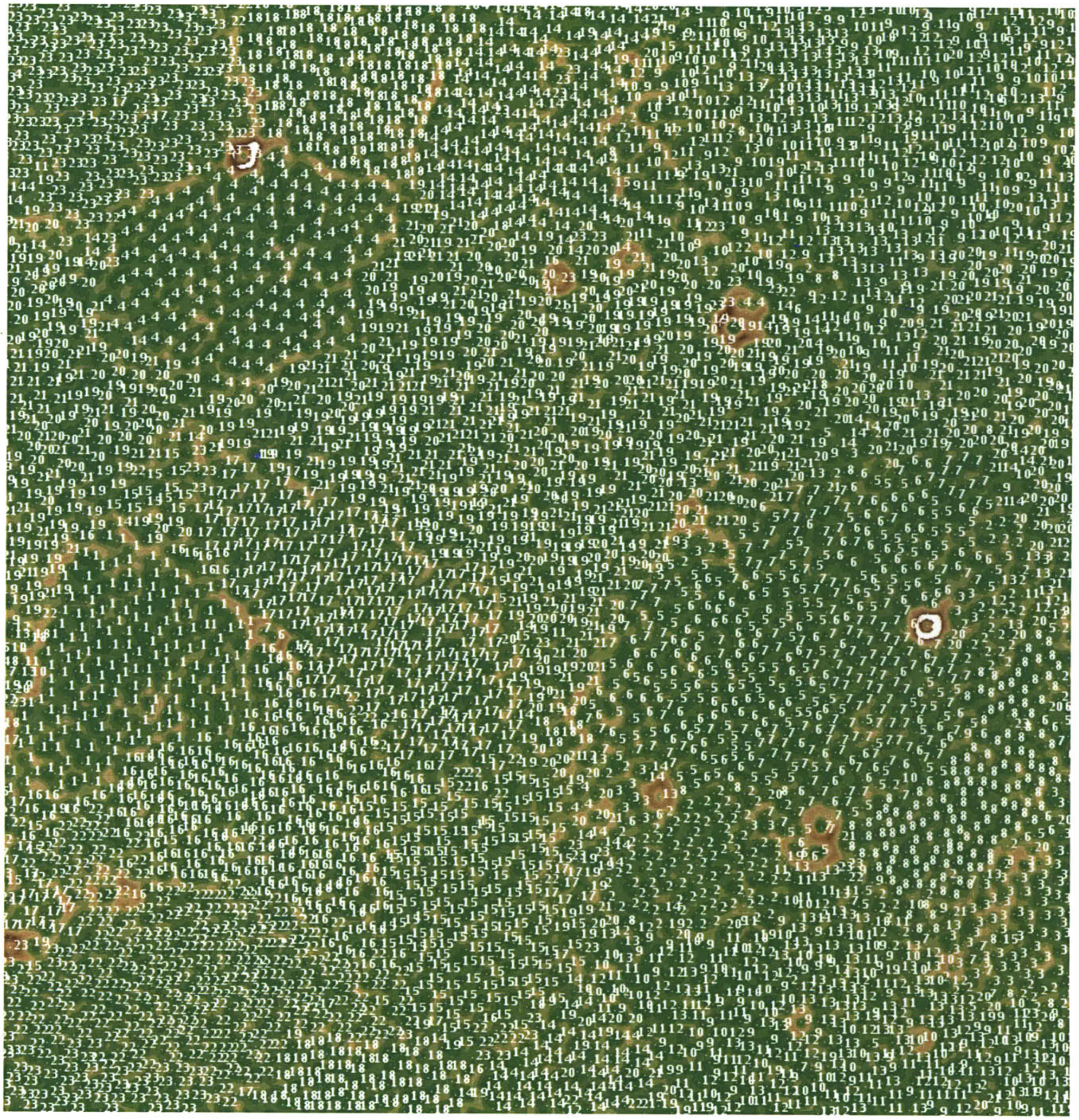


A magnificnation of the middle section of the previous map. Despite the lack of topographic boundaries separating fragments from different species, many of the sequences from genomes whose fragments clustered together in the Figures 1, 2, 3 still cluster together.

Figure 1.5. Clustering of 1.5kb fragments from cultured methanogenic genomes.

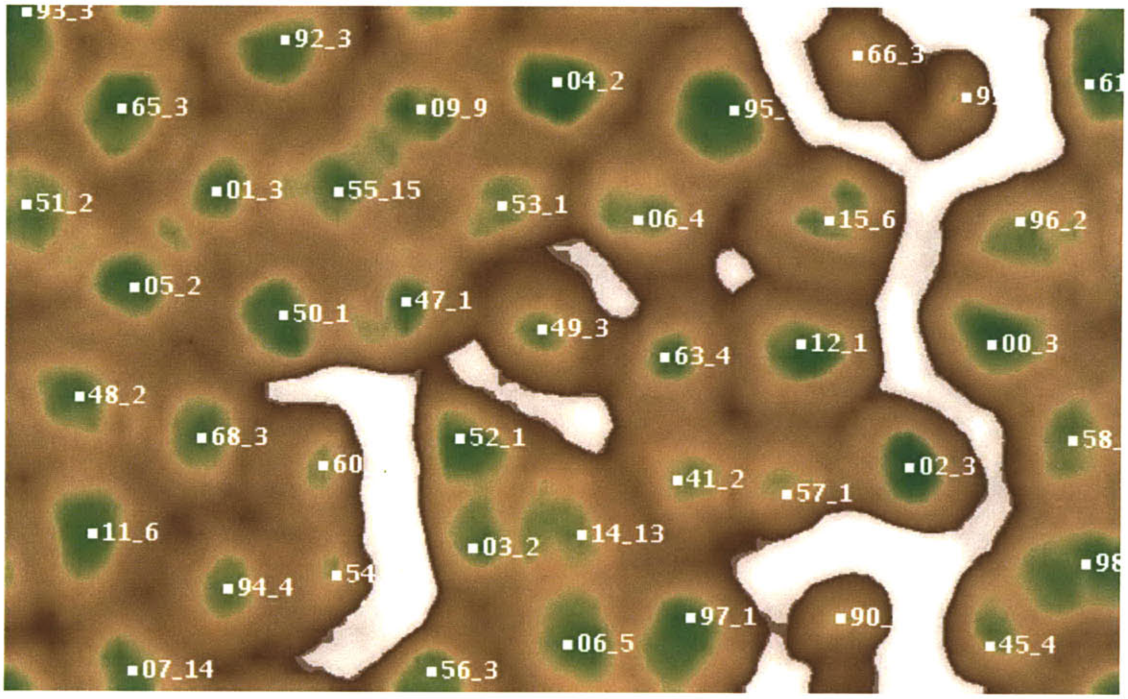


Unlike in Figure 4, here several mountainous boundaries can be seen that encircle several portions of the map. To reduce computational load to a reasonable time (several hours), only 1 of every 6 fragments was used.

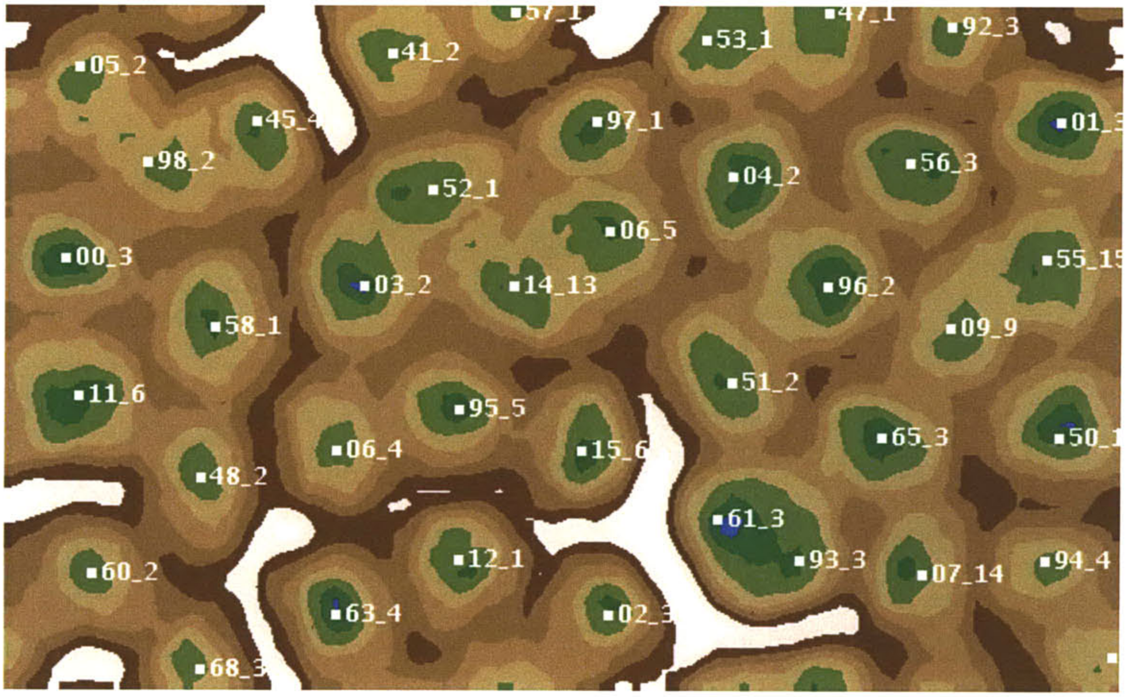


A zoomed-in, labeled version of the map on the previous page. The mountain boundaries hinted at on the previous map line up, in many cases, with fragments from the same genomes. The patterns seen here mirror those seen in the maps made with longer fragments.

Figure 1.6. Clustering of putative HF130 putative marine group II contigs longer than 20kb.

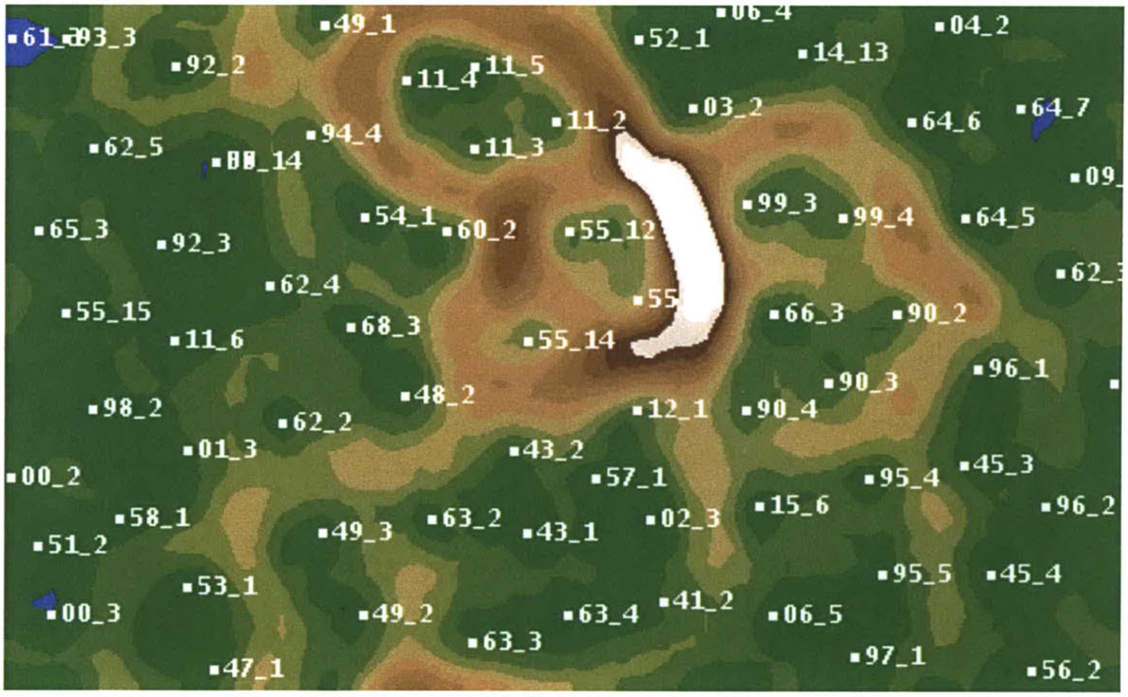


The first two digits are the JGI fosmid name, the third is the contig number. Fosmids assembled into anywhere from 1 to 15 contigs. Note that three contigs are clearly excluded: 66_3, 90_4, 99_4, and possibly 49_3. Clustering of these sequences with all the available full-length ocean fosmids shows that these sequences (the first three) cluster closer to the group III archaeal sequences than to the group II archaeal sequences.



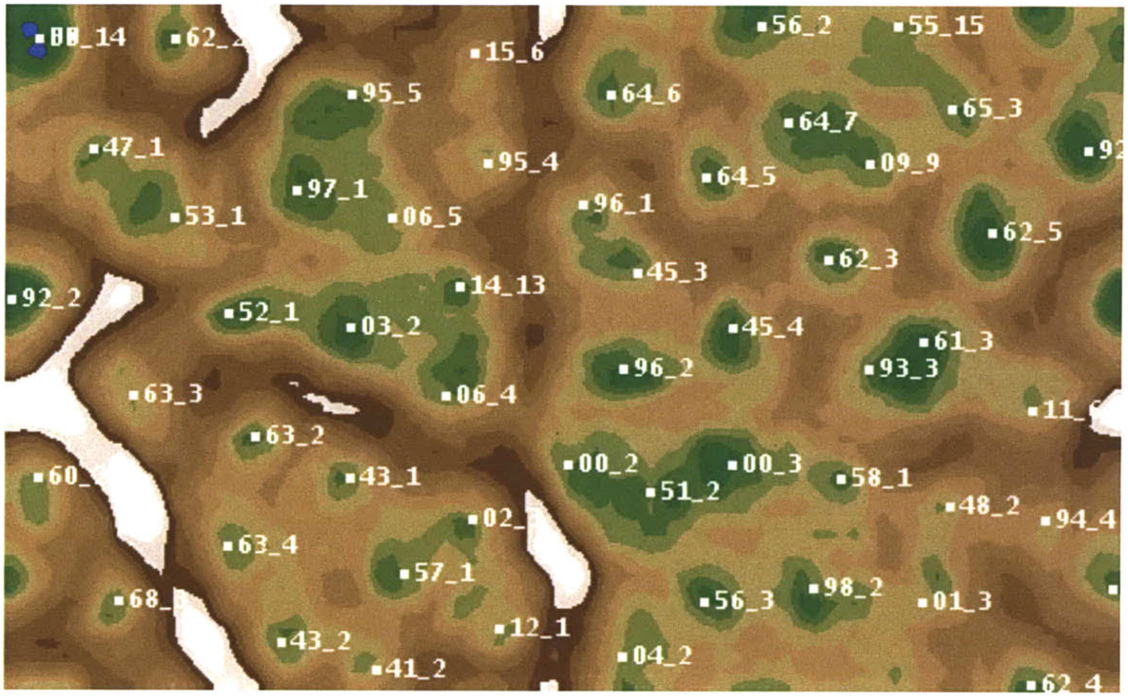
A reclustering of the previous map with the 3 outliers removed. No clear exceptions are visible any longer.

Figure 1.7. Clustering of putative HF130 putative marine group II archaeal **contigs longer than 2 kb.**



Again some fragments clearly cluster away from the other sequences, particular those from 11, 55, 99, 90, 66 and 49. Note, however, that 55_15 (middle left hand side of the image) clusters well within most of the putative group II archaeal sequences. (The sequences represented above are of varying lengths, from 2 kb to over 20 kb.) This reveals an important point: portions of group II archaeal sequences can appear bacterial (which is what the three 55 contigs up in the mountains look like), potentially (although not necessarily) due to HGT events. It may be prudent to remove these sections from seed sequences to avoid recovering sequences that look like where the HGT section came from (in this case a bacterium).

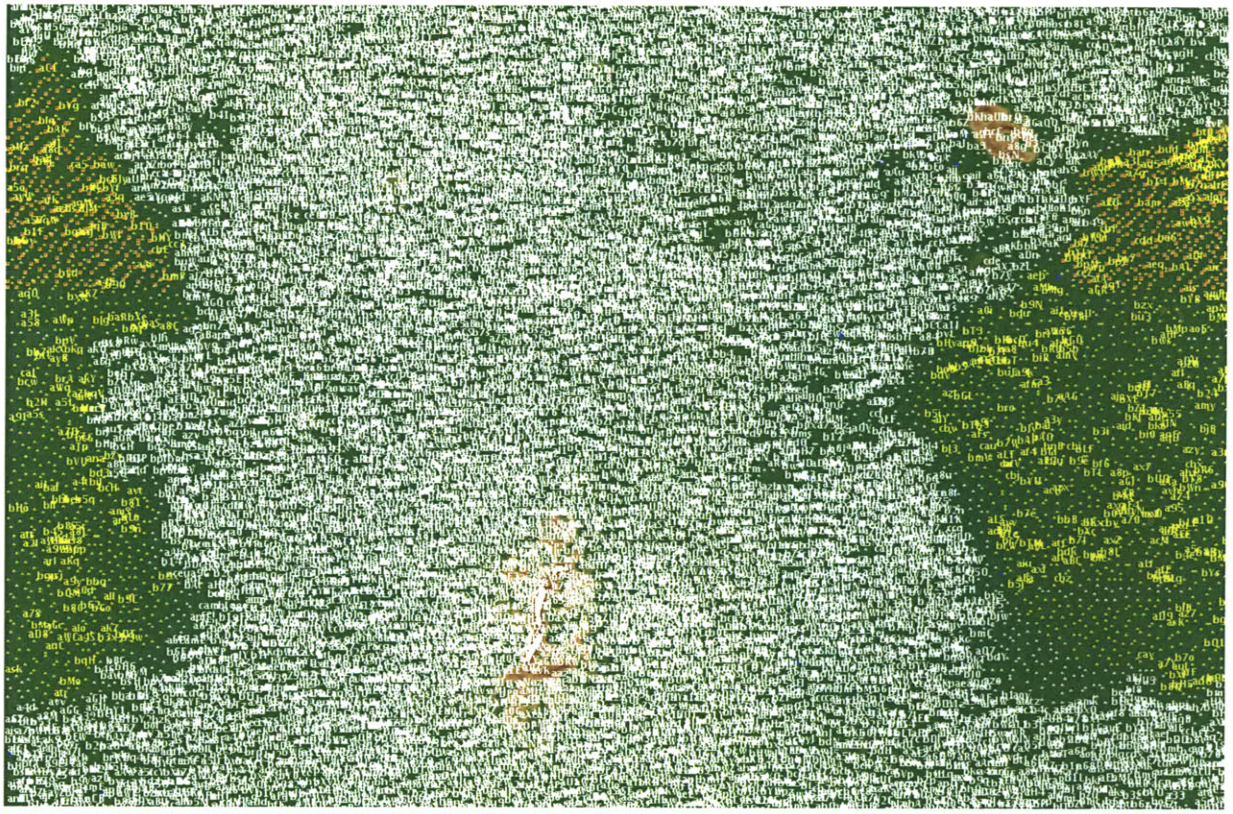
Figure 1.8. Clustering of putative HF130 putative marine group II contigs longer than 2kb; outliers in Figure 7 have been removed.



No more clusters are sectioned off far away from the other sequences.

Figure 1.9. Clustering of all HF200 fosend sequences longer than 700bp with putative marine group II sequences cut into 1kb segments.

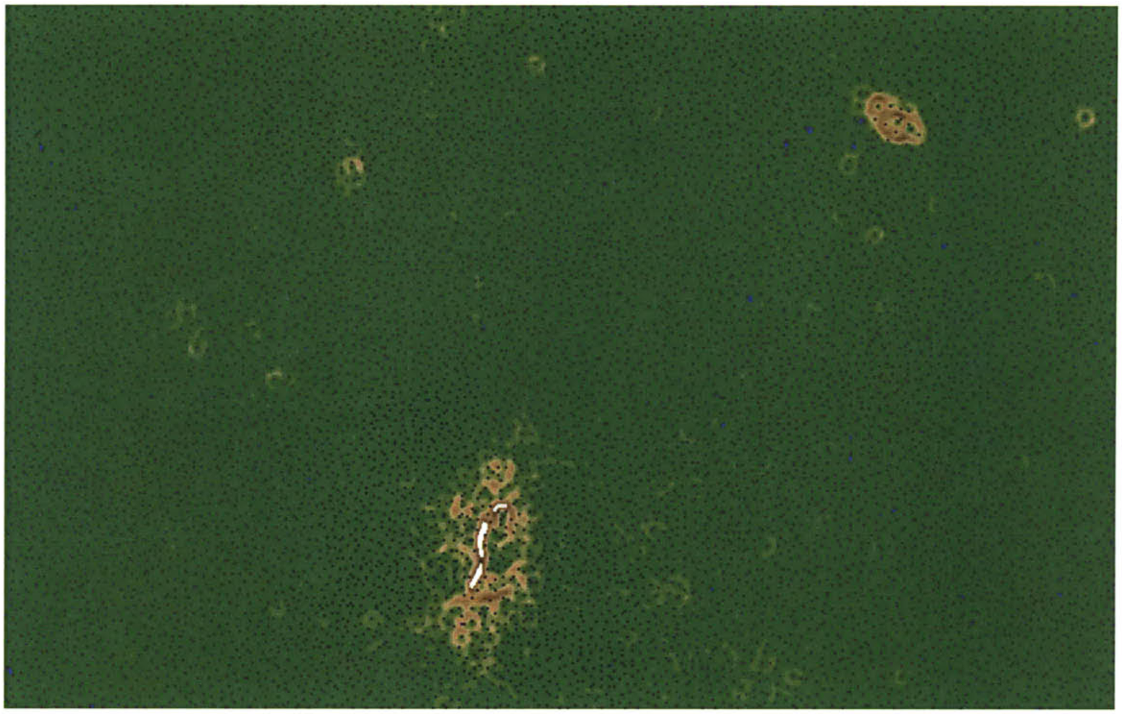
Putative marine group II archaeal sequences came from three sources: contigs assembled from the sequencing of HF130 fosmids, NCBI group II archaeal fosmids, and putative group II archaeal NCBI deposited fosmid sequences, based on homology to HF130 contigs. Each fosmid end sequence is marked with a three-letter code; group II archaeal seed sequences are labeled with an asterisk (*).



Clustering of putative group II archaeal fosmid ends was done by hand, collecting (collected sequences are highlighted above in yellow) all sequences that fell within the bulk of those marked with an asterisk.

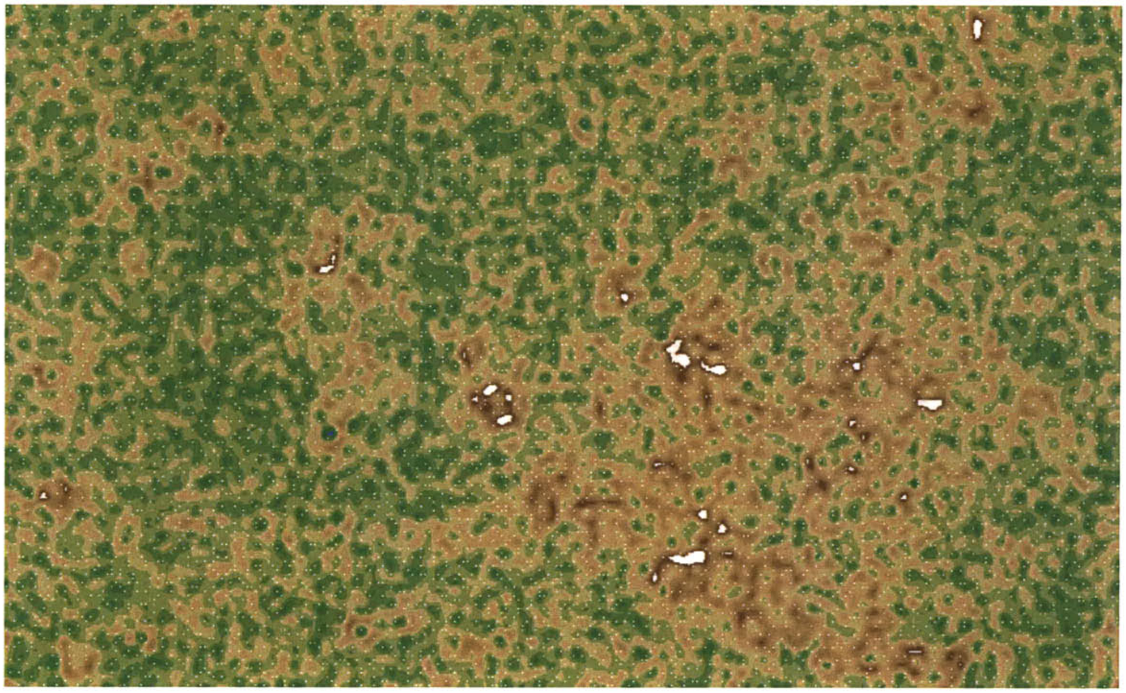
Above is a zoomed-in version of the upper left hand side of the previous graph.

Figure 1.10. Clustering of all HF200 fosend sequences longer than 700bp with putative group II sequences included, as background; labels removed.



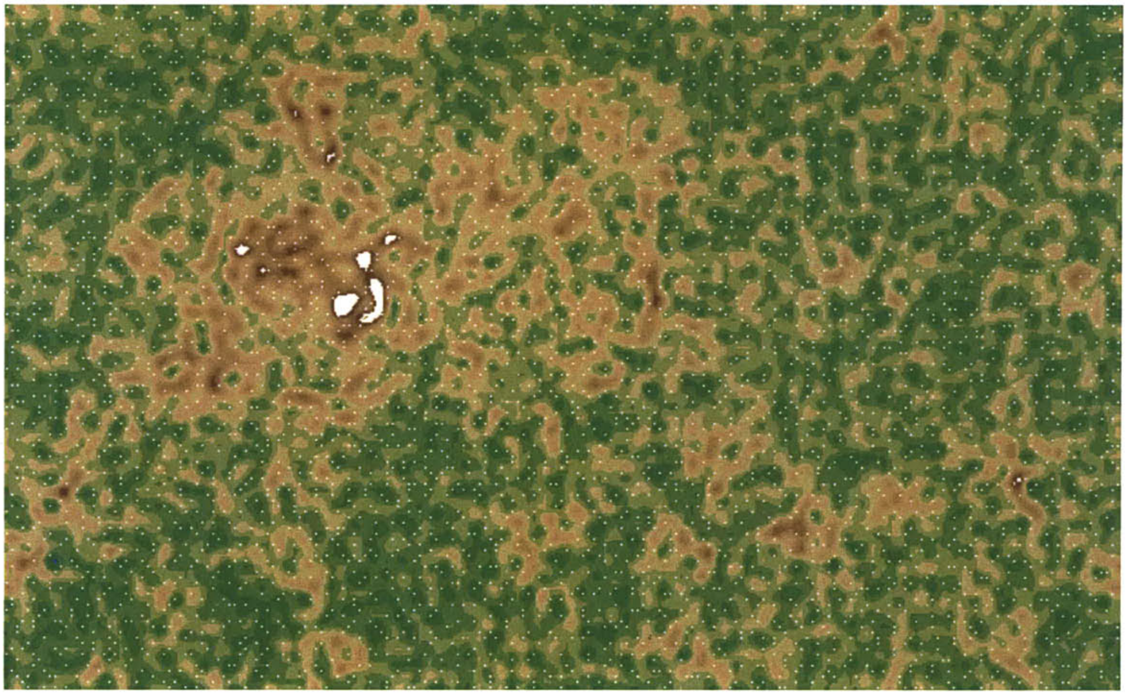
By and large the background is one large valley, with a few mountainous regions. As expected for input sequences on the order of 700 kb, emergent clustering of the sequences themselves are not clear.

Figure 1.11. Clustering of fosend sequences longer than 700 bp from HF200; 'mountain' sequences in Figure 6 removed.



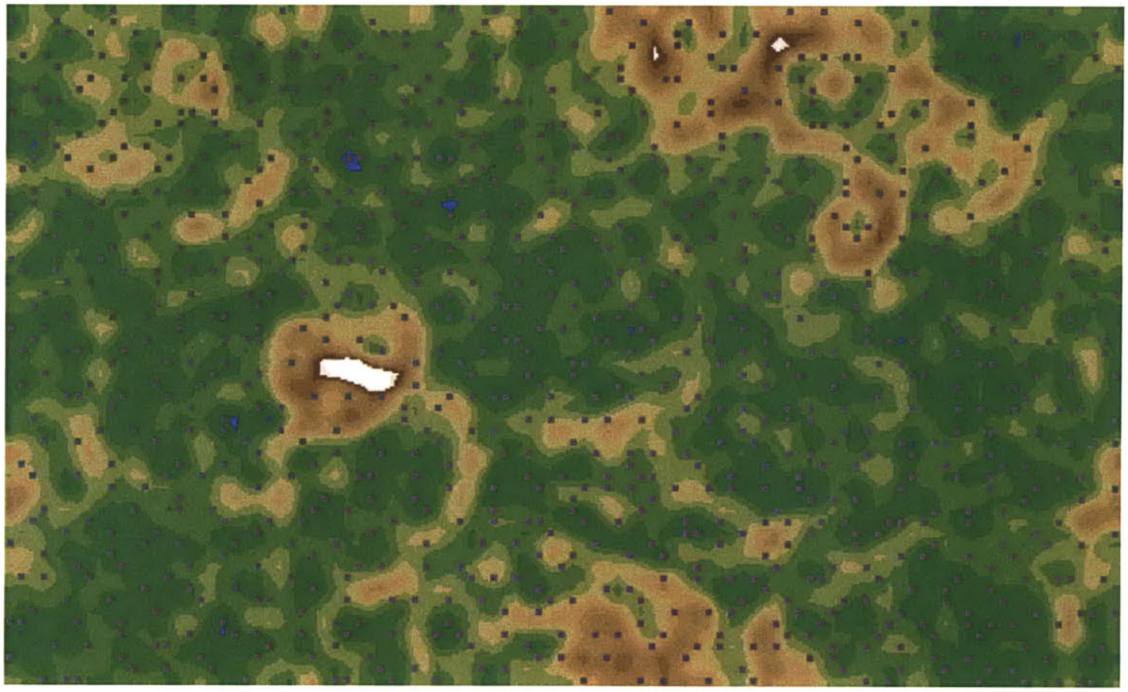
To view the emergent clusters from the HF200 fosmid end sequences, without the very mountainous sequences (those very unlike the others) obscuring the topography, several mountain sequences were removed from the ESOM input. Quality files were unavailable for the all the HF library sequences. Many of the mountain sequences were, upon inspection, junk sequence. Others were highly repetitive sequence, atypical of most bacterial and archaeal genomic regions. Included among these highly repetitive sequences were telomere sequences, $(TAACCC)_n$. Highly repetitive sequences are expected within the DeLong lab HF fosmid libraries, since several percent of the organisms represented in the libraries are eukaryotic (evidenced by 18S rRNA content). Several weakly bounded valleys are evident, but no clear boundaries, indicating clustering are clearly visible. 'Mountain' sequences removed from Figure 7 are listed in List 1.0 (see Appendix)

Figure 1.12. Clustering of joined fosmid end sequences from HF200 exceeding 1.5kb in length (Mountain sequences from Figure 10 removed.)



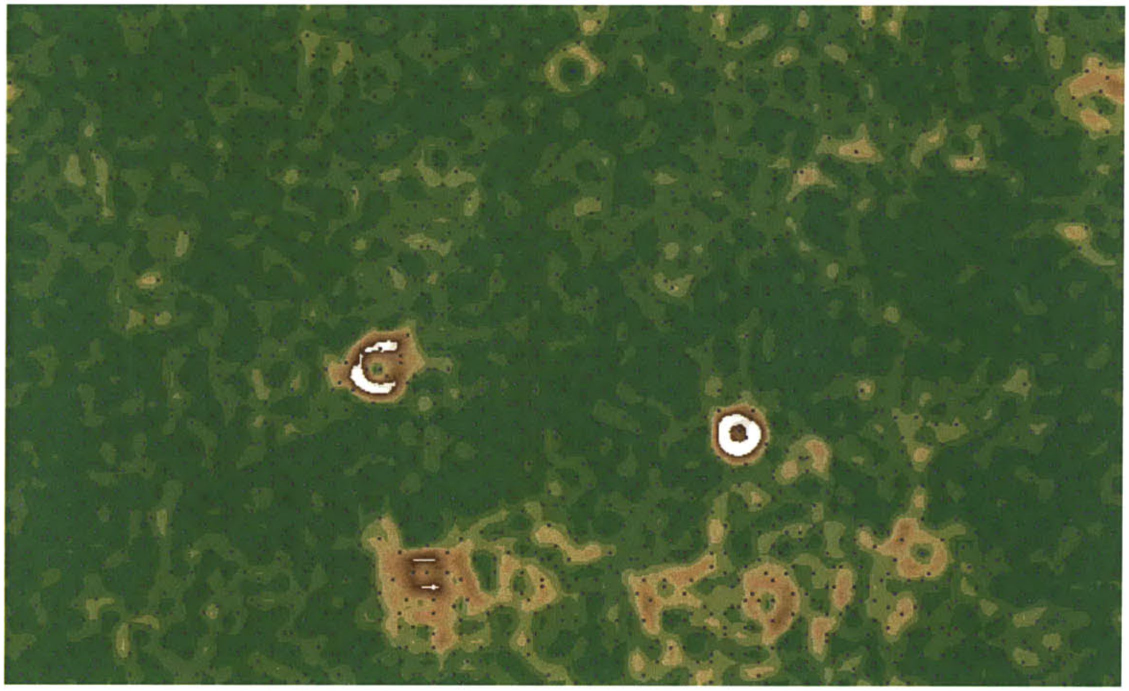
Clearly bounded valleys, indicative of sequences that clearly cluster together, are more visible here than they are in Figure 8.

Figure 1.13. Clustering of joined fosmid end sequences from H179 125m exceeding 1.5 kb in length. No sequences were removed.



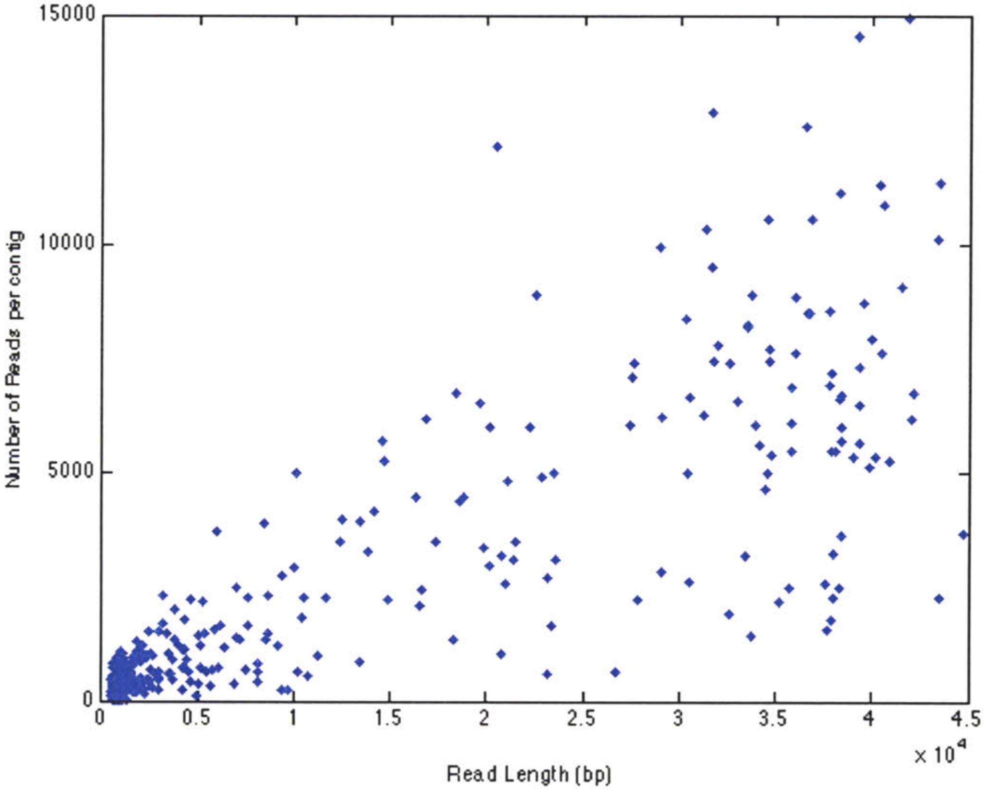
Only about 1000 fosmid ends fit the qualifications of having a total of 1500 bp with quality scores over 20. About four mountains are visible; no very clear emergent "valleys" are visible, though a few potential "valleys" could be hypothesized.

Figure 1.14. Clustering joined fosmid end sequences from H179 125m exceeding 1.4 kb in length. No sequences were removed.



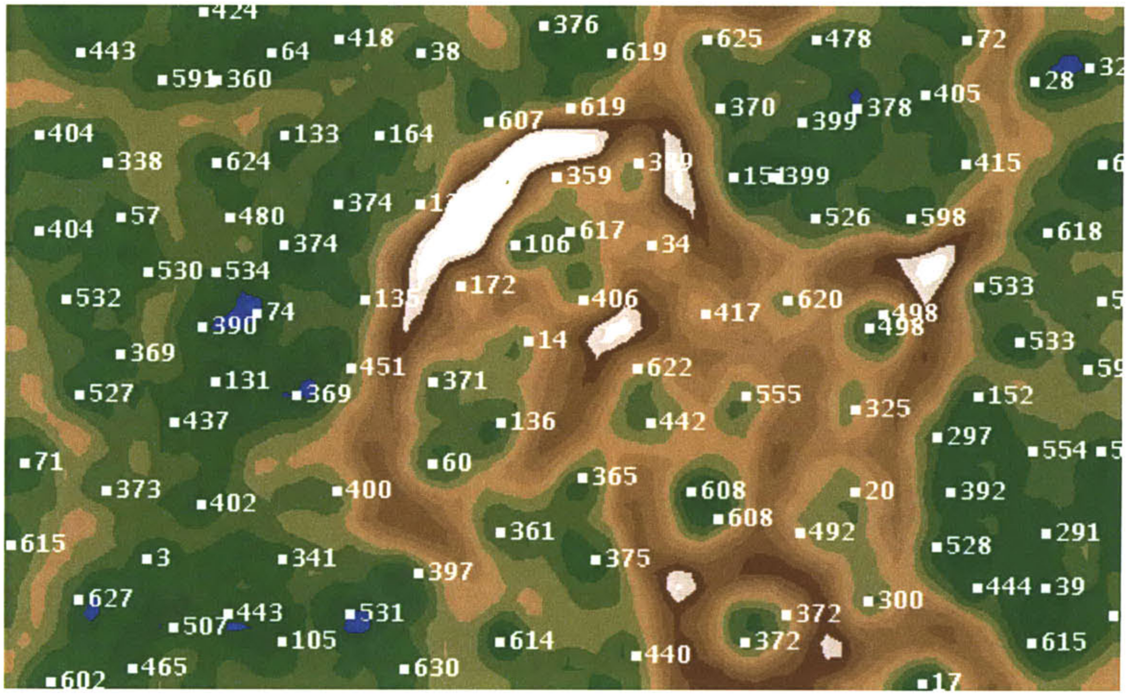
Note that no clear emergent boundaries are visible.

Figure 1.15. Coverage from raw sequence reads of contigs assembled by the gsAssembler software.



On the horizontal axis is the length of the contig. Notice a break at about 25 kb. Sequences above this length may well be full-length or nearly full-length original fosmid sequences. The vertical axis represents the number of raw sequence reads used to construct the each contig. If each contig had the same coverage, points would fall along a single diagonal line. Reasons for the variability in yield could include : different fosmids have potentially different effects on the *E.coli* host cell, causing some clones to replicate more or less efficiently than others. Cells were grown to slightly different densities, and all the fosmids pooled. Certain cell lysis and fosmid extraction steps may have had higher yields in some clones as compared to others.

Figure 2.1. Marine group II archaeal **fosmids binned by SOM clustering of tetranucleotide signatures.**



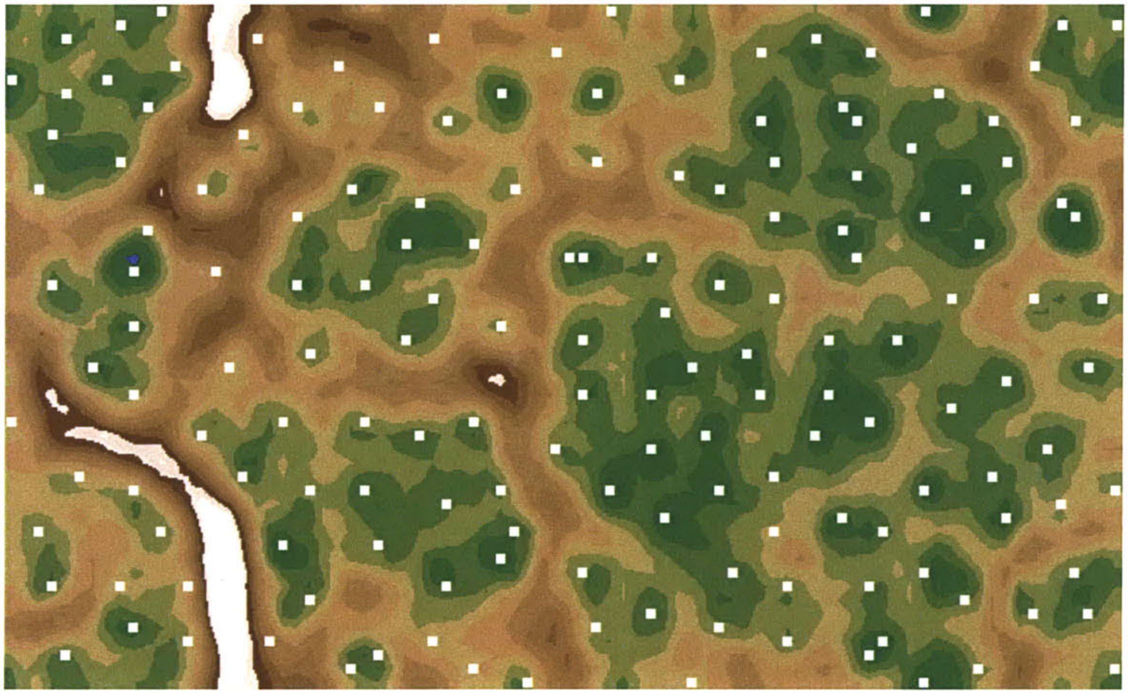
Clustering of 20 kb fragments of all 454-sequenced tetranucleotide signatures of contigs over 20 kb from HF200. Note that a majority of the sequences are in green valleys; a sizeable fraction, however, inhabit the hilly space in the center of the map. Note that when two labels are identical, that is because the contig is over 40 kb, so that taking 20 kb fragments produces two tetra signatures, with the same label (the contig/fosmid name).

Contigs represented by points in the hilly regions:

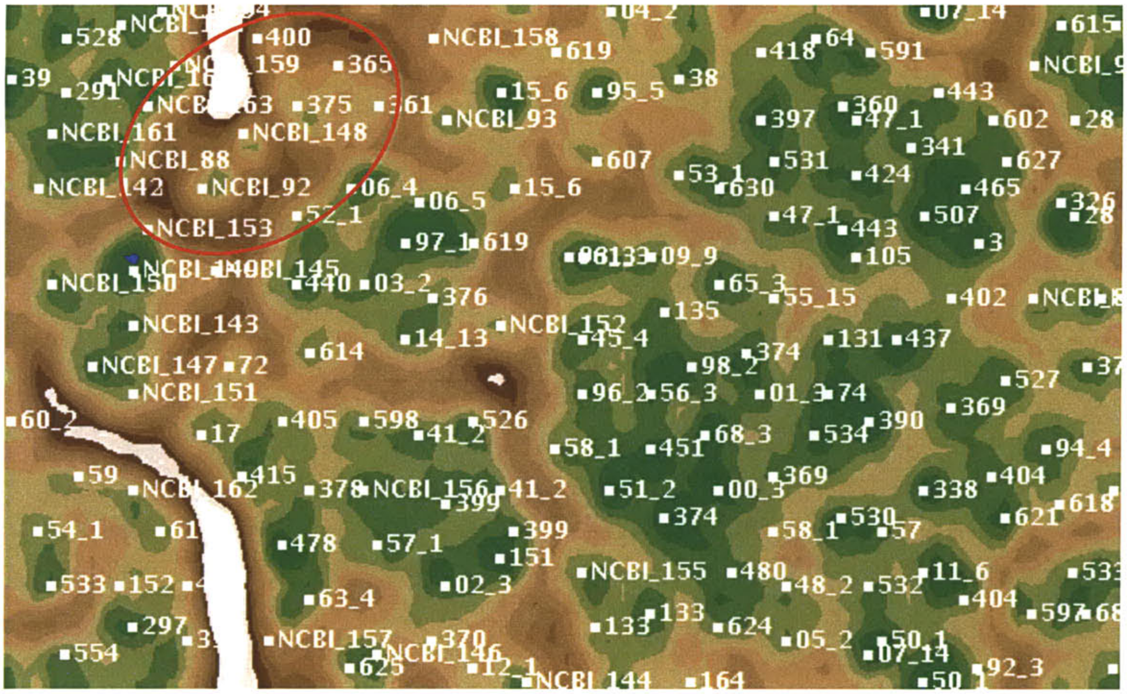
14, 20, 34, 60, 106, 136, 172, 300, 325, 359, 371, 372, 389, 406, 417, 442, 492, 498, 555, 608, 617, 620, 622.

Of these, only 3 show any statistically significant homology to *any* fosend in the HF200 library: 14, 172, and 325. And of these, only 172 shows homology at the *end* of the sequence to an existing end-sequence, meaning only 1 of these 23 contigs might match to a fosmid end sequence.

Figure 2.2. Marine group II archaeal fosmids binned by SOM clustering of tetranucleotide signatures, outliers removed.

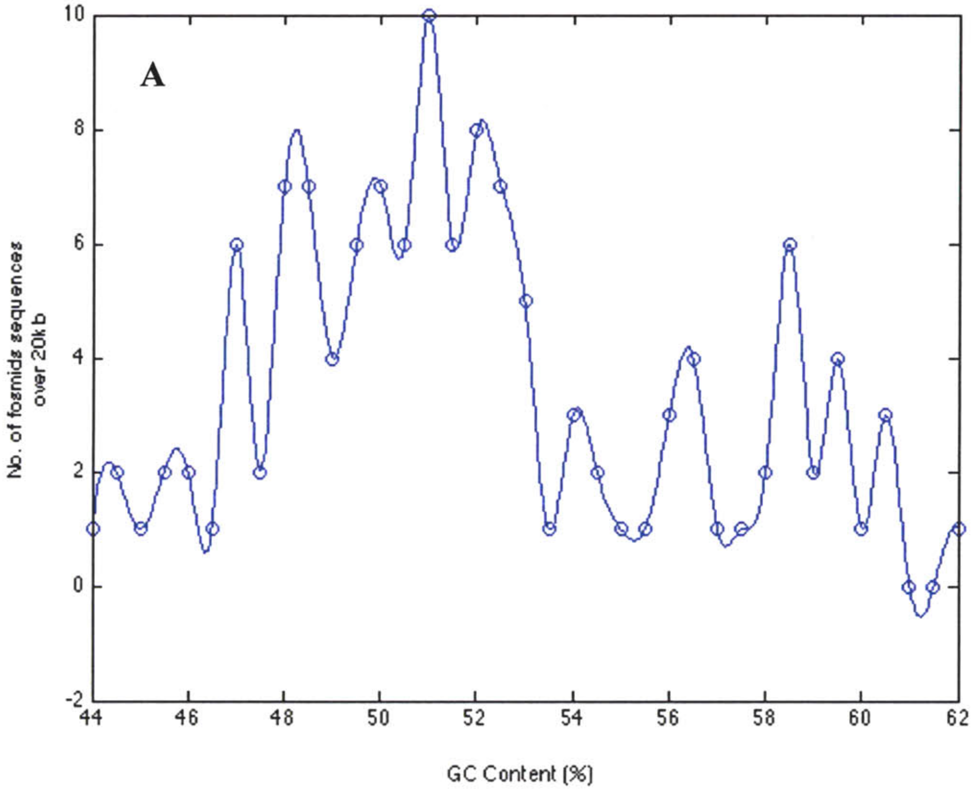


Clustering of all 454-sequenced tetranucleotide signatures of contigs over 20 kb from HF200, with 20kb fragments of all putative group II archaeal sequences over 20 kb as background. Labels are omitted, because the topography is clearer without them.



Clustering of all 454-sequenced tetranucleotide signatures of contigs over 20 kb from HF200, with labels. There are only 2 clear clusters that do not belong with the remainder of the sequences, in the upper left hand portion of the map, circled in red. Those belonging to one cluster: 361, 365, 375, 400. Those belonging to another: NCBI_92 (APKG2M17), NCBI_148 (HF4000_137B17). Blast analysis confirms that the above fragments have no clear euryarchaeal markers.

Figure 2.3. Group II fosmids binned by GC content



Bin size for Figure 2.3A: 0.5% GC. Noise is evident, but two humps, one centered at about 51% GC, the other at about 58% GC could be postulated.

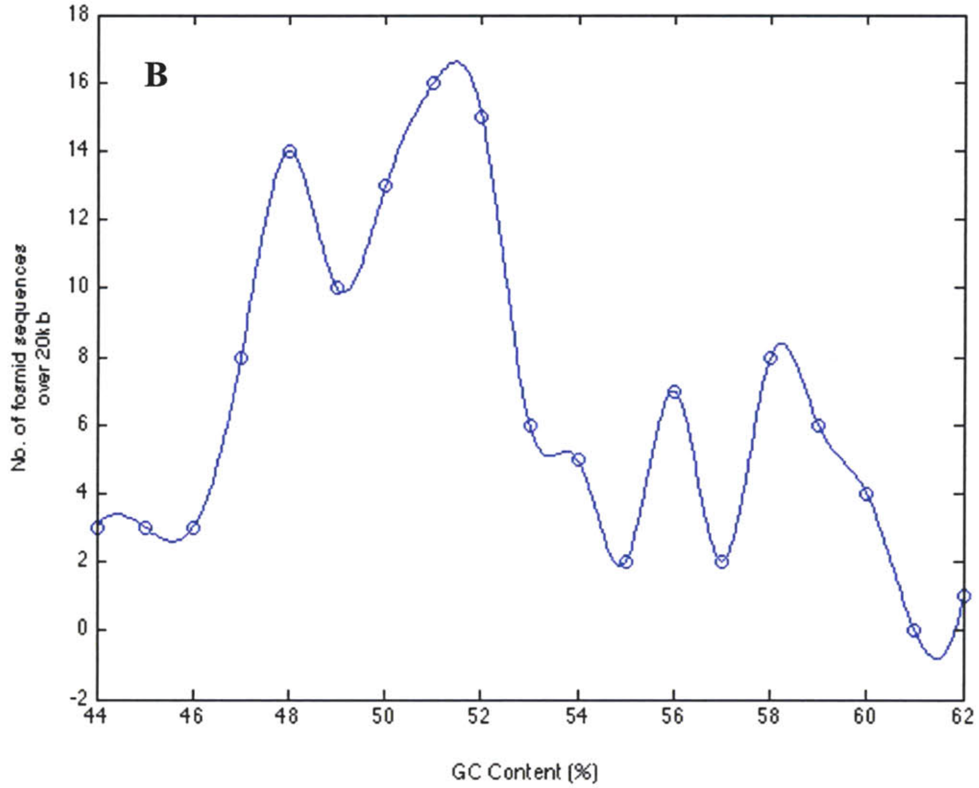


Figure 2.3B bin size: 1% GC. Once again, an above-random bump occurs from about 48-52% GC. It is unclear whether the remaining variation is real or noise-induced. A potential approach to resolve the issue is the variation in GC content over 20 kb in a single *in situ* genome. If that variation is less than 1%, these peaks may be real. Yet, even if the variation in GC content over a genome segment over 20 kb is less than 1%, the variation in peak height is on the order expected for poisson fluctuations (usually valid when randomly sampling small numbers of a sample). Standard deviation ~ √N.

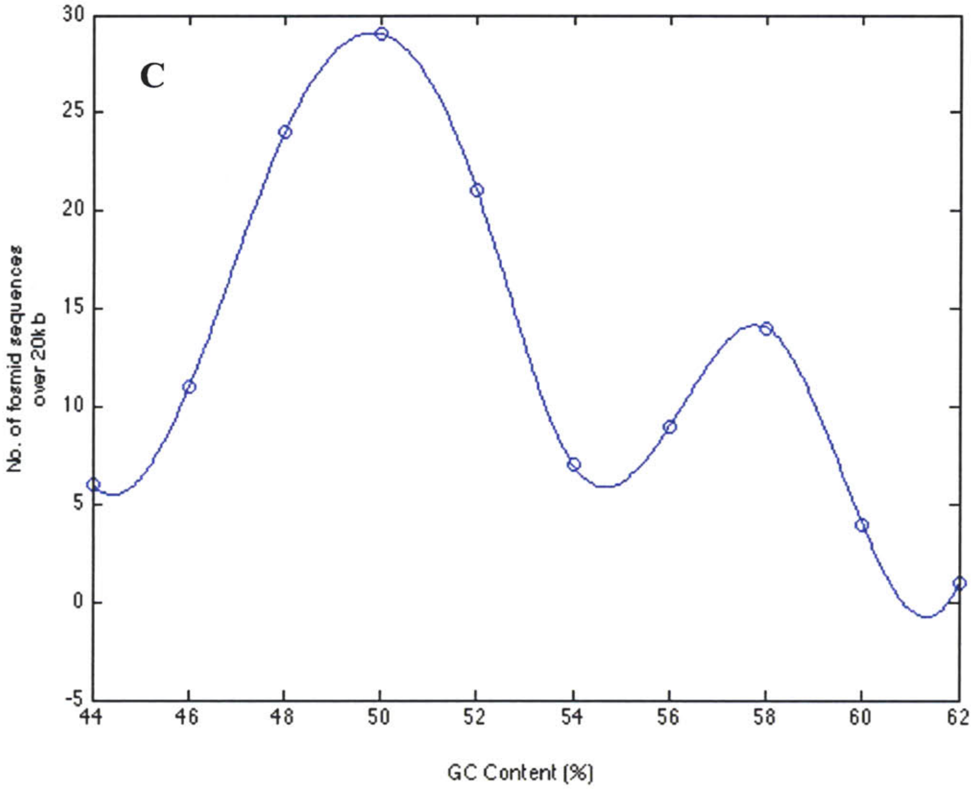
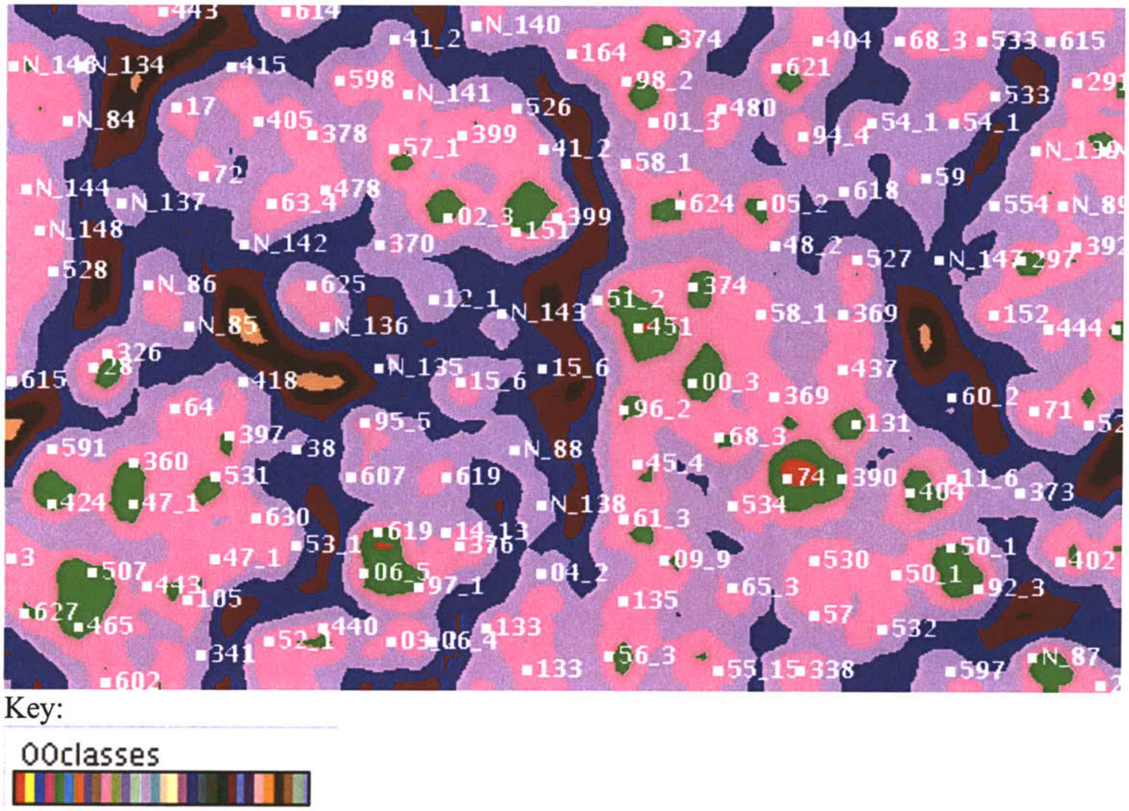
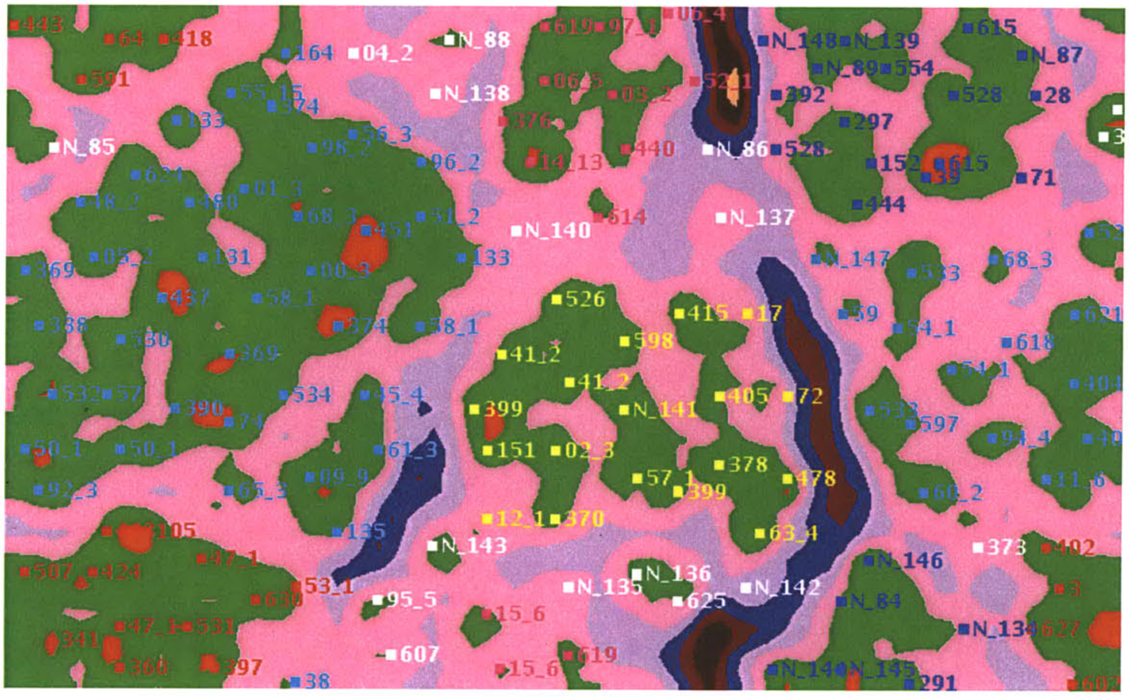


Figure 2.3C bin size: 2% GC. 55% GC emerges as a natural boundary between the two populations. With this increased bin size, both peaks are statistically significant.

Figure 2.4. Further clustering of HF200 assembled contigs, non-marine group II archaeal sequences removed.



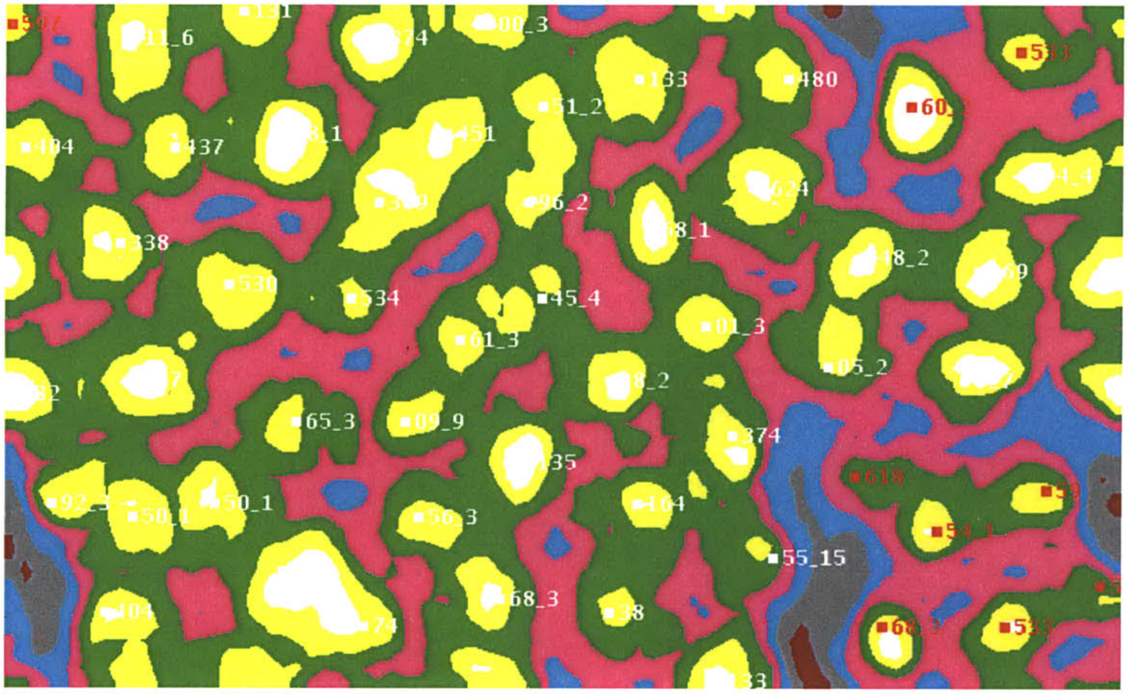
A different map coloring is used, because it makes the visualization easier. Note that the NCBI_XX labels in this figure may differ, with respect to the fosmids they label, than those in the previous figure. Five clusters emerge (see Appendix, List 2.1).



Key:

A second clustering, with the different classes labeled with different colors. The topography looks slightly different, because 93_3 and 07_14 are omitted. These two have strong exact overlap with other fosmids (93_3 with 61_3; 07_14 with 50_1). N_147, by looking at the first map in this figure, should be white (an outier), not light blue.

Figure 2.5. Refining Cluster 5 of the marine group II *Archaea*.



The wingzcolors made the misfits most clearly visible. The key for the new color scheme:

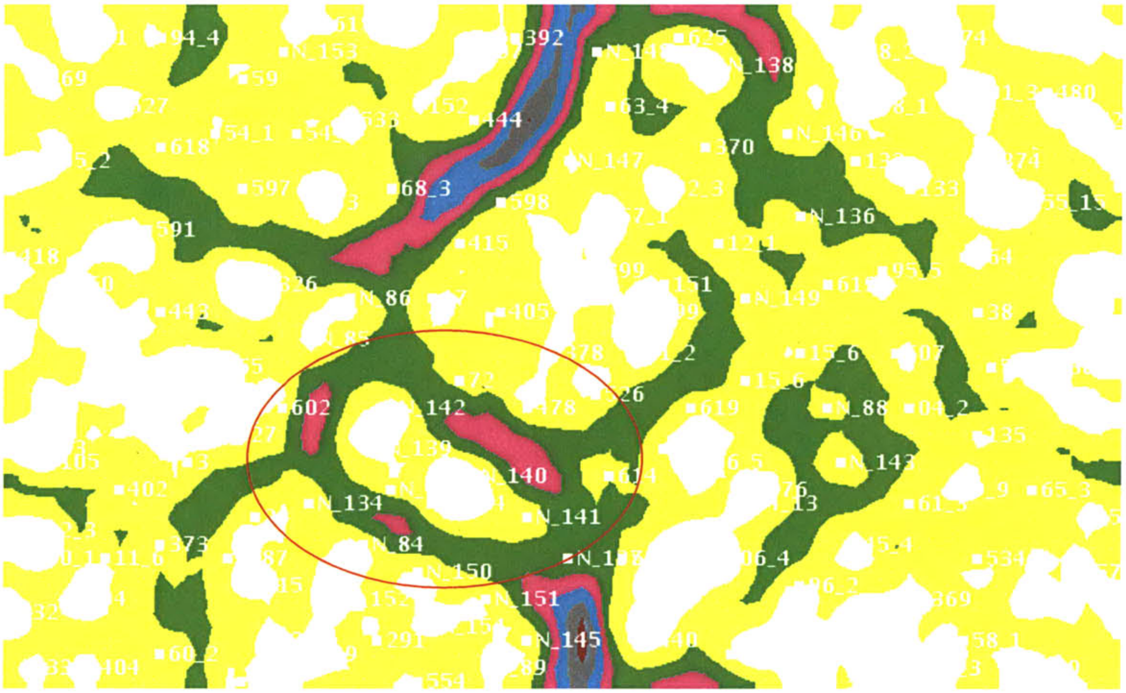
wingzcolors

⬜ 🟨 🟩 🟥 🟦 ▨

The following, colored in red on the above map, cluster away from the rest of Cluster 5:

Contig Name	*% GC*
contig00059	59.1
contig00533	58.2
contig00597	56.0
contig00618	52.7
4003654_fasta.screen.Contig1	57.9
4003660_fasta.screen.Contig2	58.1
4003668_fasta.screen.Contig3	54.2

Figure 2.6. Bathypelagic water cluster of marine group II *Archaea*:



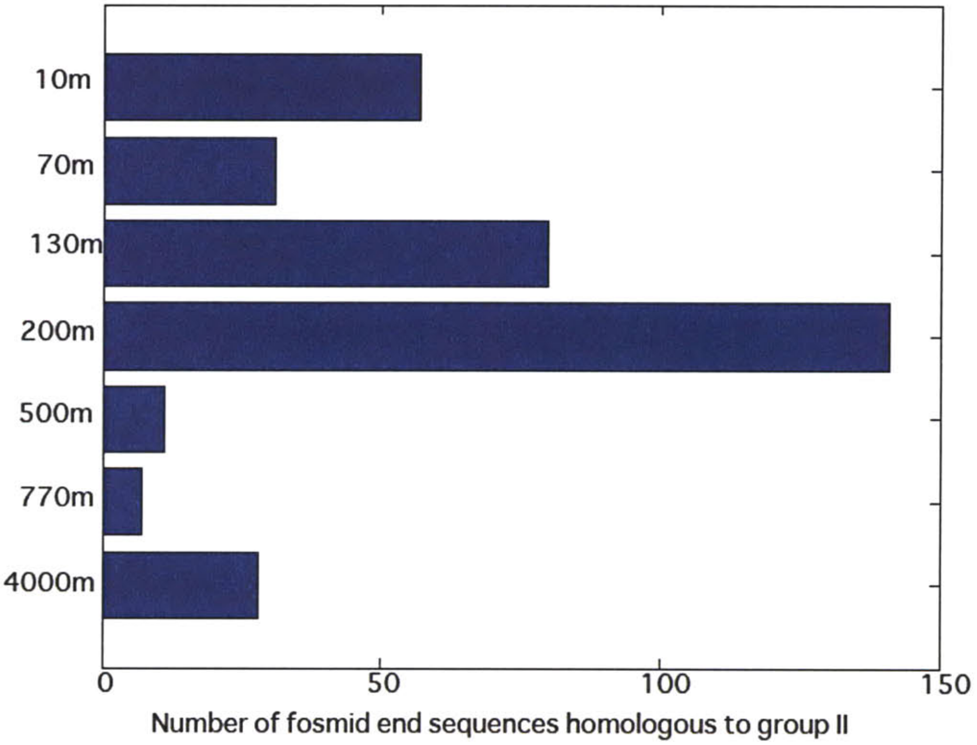
Though the labels are not easily read, the above plot clearly shows a clustering including deep water fosmids, with no photic zone sequences in the same cluster. The cluster has been circled above in red.

The fosmids in that cluster, with GC content listed are:

Cluster 6

Short Name	Full Fosmid Name	%GC
N_135	DeepAnt-JyKC7	57.1
N_139	HF4000_001N02	56.6
N_140	HF4000_ANIW133F6	56.7
N_142	HF4000_ANIW137P11	55.0
N_144	HF4000_ANIW141C7	56.7

Figure 2.7. Representation of marine group II archaeal presence in depth profiles for the HF, HOT179 and HOT186 cruises.

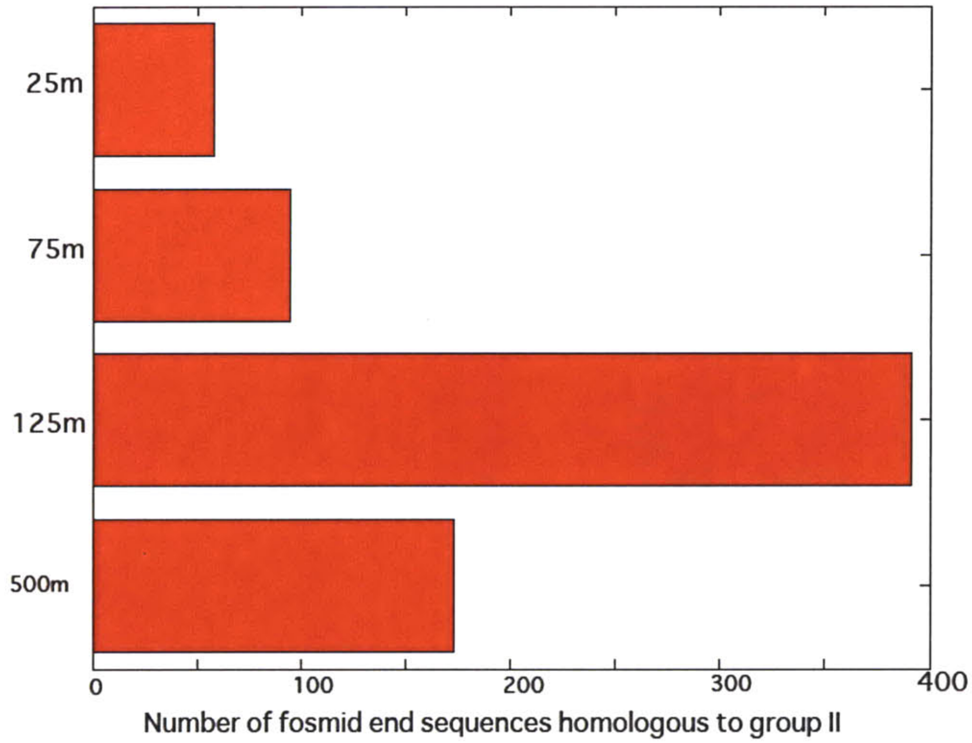


A fosmid end sequence was defined as homologous to an existing group II archaeal sequence if the region of homology covered at least 93% of the end-sequence length, if the percent ID over that region was at least 85%, and if the bitscore was at least 190 – the last condition ensures that perfectly matching, but very short sequences (such as those present in the HOT179 and HOT186 libraries), were not counted.

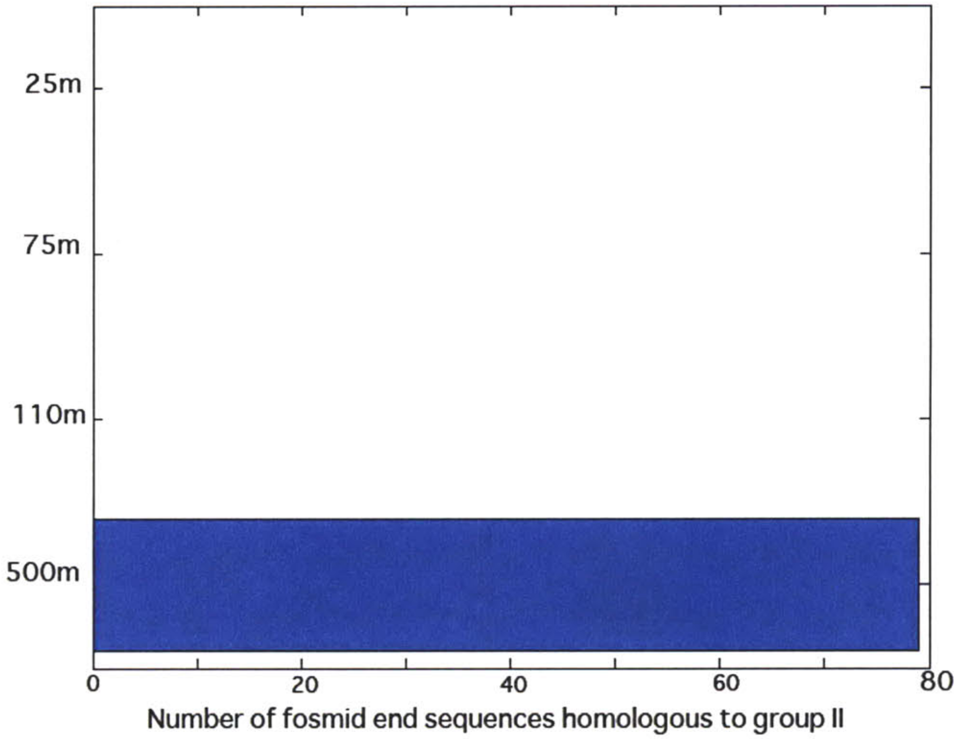
The bars shown above are actual values. Note that the library sizes varied from library to library, and that the volume of seawater necessary to achieve the same biomass in the deep is much greater than that in the photic zone.

Note also that the current pool of group II archaeal sequences is heavily biased to those from HF130 and HF200. Thus the actual group II archaeal presence appears much more prevalent at those two depths. The numbers of group II archaeal 16S sequences found in the DNA sampled at each depth, from [3] are given below for reference.

Library	No. of 16S rRNA sequences (numbers taken from [3])	No. of Sequences over 100bp
HF10	21	7,829
HF70	13	10,978
HF130	32	6,754
HF200	22	8,241
HF500	11	9,003
HF770	9	11,473
HF4000	13	11,212

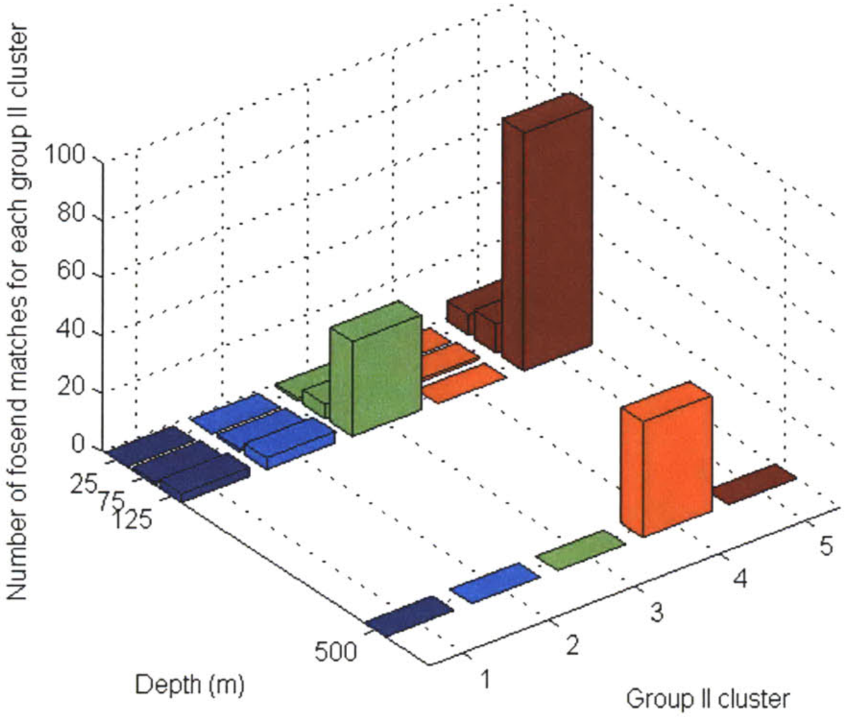


HOT179 profile. The total number of sequences over 100 bp, after filtering out low quality scores at each depth was (in parentheses): 25m(20,705); 75m(19,801); 125m(27,307); 500m(27,181).

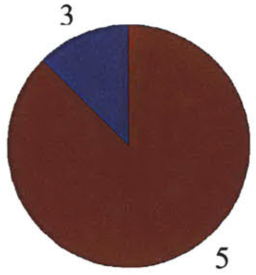


HOT186 profile. The total number of sequences over 100bp, after filtering out low quality scores at each depth was (in parentheses): 25m(26,237); 75m(17,524); 110m(21,824); 500m(25,613).

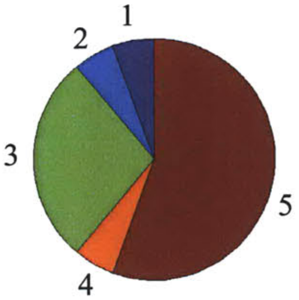
Figure 2.8. Depth profiles of the 6 emergent clusters of the marine group II archaea.



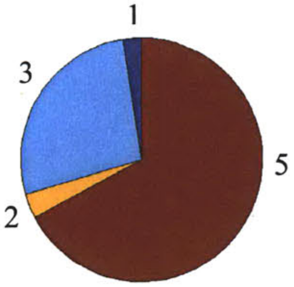
HOT179 Group II Archaeal Depth Profile. A match from an end sequence to a group II sequence was defined as blast hit whose length covered over 97 % of the end sequence length, over which the percent ID was over 98%, and the bitscore of which was at least 190. The number of matches broken down by contigs within clusters are printed in List 2.3 (see Appendix). Below, normalized to the total number of group II archaeal matches, is the breakdown, in pie chart form, for clusters.



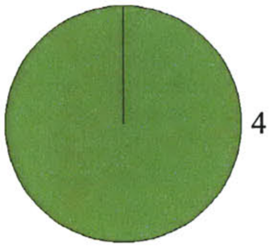
HOT179 25m



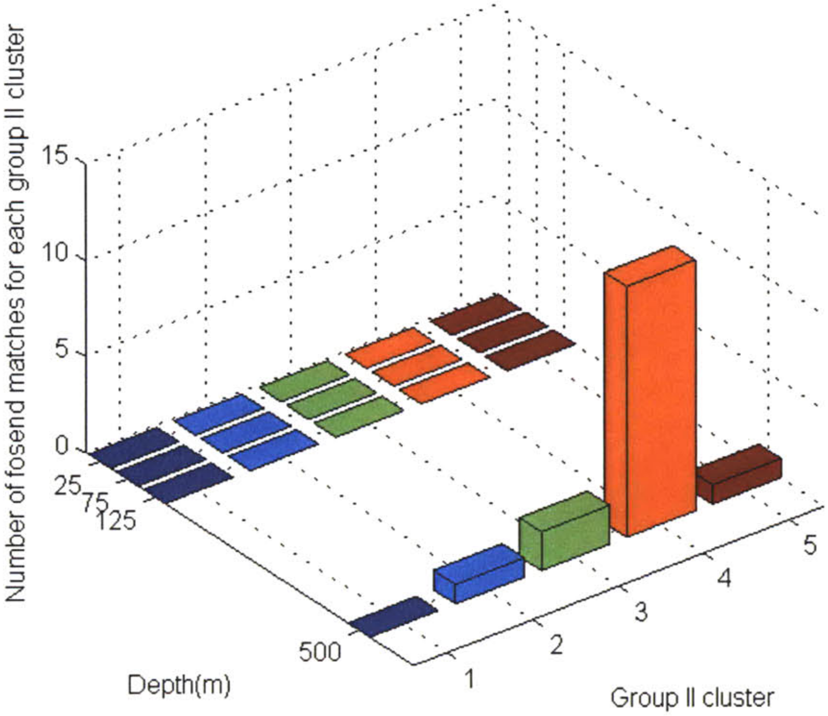
HOT179 75m



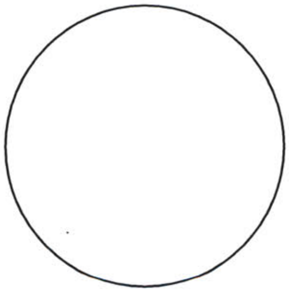
HOT179 125m



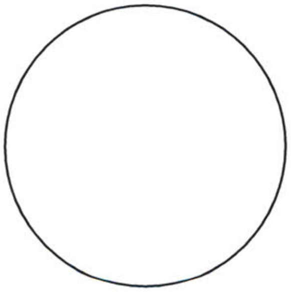
HOT179 500m



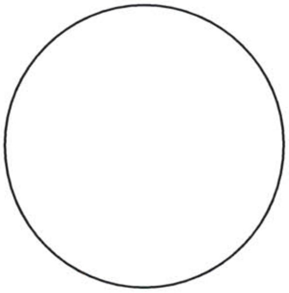
HOT 186 clusters. Numbers of matches to contigs within clusters are printed in List 2.4 (see Appendix).



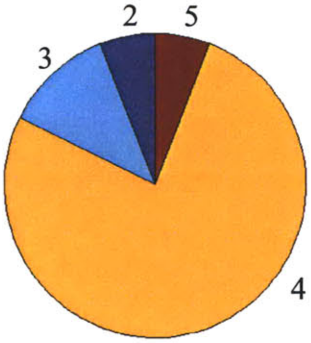
HOT186 25m



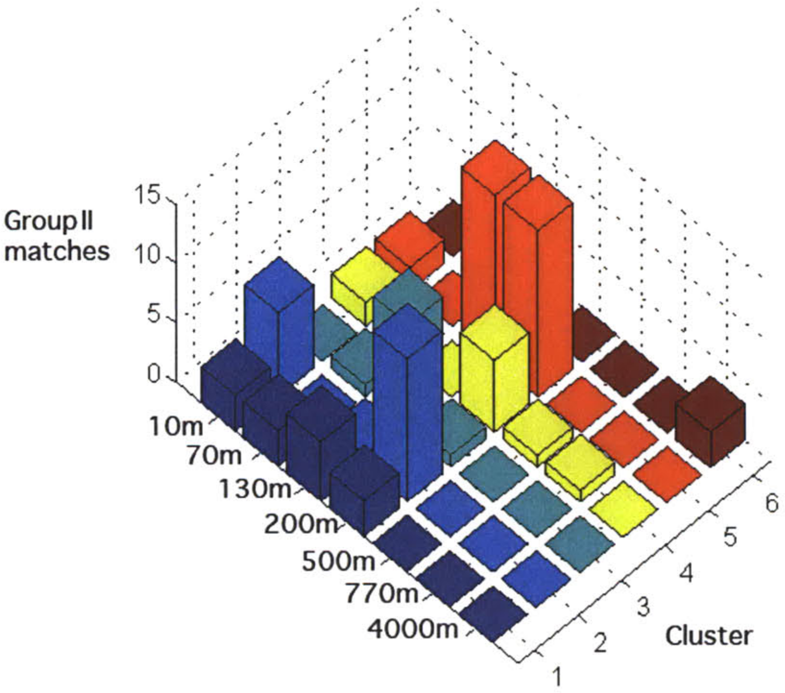
HOT186 75m



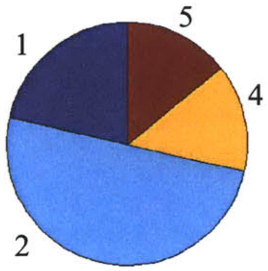
HOT186 110m



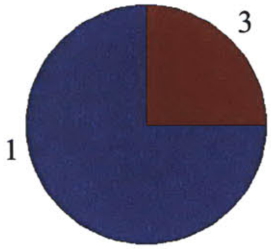
HOT186 500m



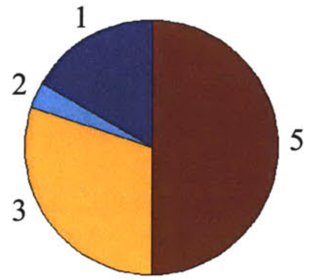
HF clusters. The depths are not drawn to scale, to make viewing the bars easier.
The numbers of matches to contigs within clusters are printed in List 2.5 (see Appendix).



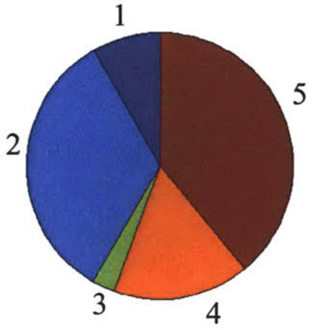
HF 10m



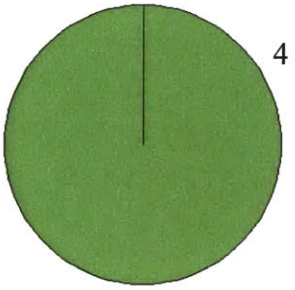
HF 70m



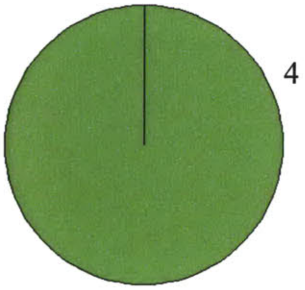
HF 130m



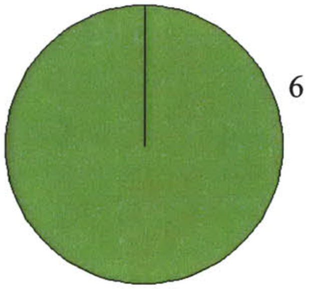
HF 200m



HF 500m



HF 770m



HF 4000m

Tables

Table 2.1. Comparison of marine group II archaeal clustering via SOM to clustering via GC content

Cluster	Number of sequences	Average GC content (%)	High GC value (%)	Low GC value (%)
1	17	49.8	52.6	46.2
2	19	51.7	53.4	48.9
3	11	46.4	48.3	43.3
4	19	59.0	62.1	55.4
5	48	51.8	59.1	47.2
5, outliers removed	41	50.9	56.0	47.2
6	5	56.4	57.1	55.0

Appendix

List 1.0 Sequences in Figure 1.10 that cluster away from the "valley," that is those sequences whose tetranucleotide signatures are statistically unlike the other input sequences.

Dominant "Mountain" regions in group II archaeal clusters :
ASNG1001.g2
ASNG1004.g2
ASNG1008.g2
ASNG1010.g2
ASNG1013.g2
ASNG1017.b2
ASNG1021.g2
ASNG1022.b2
ASNG1028.b2
ASNG1032.g2
ASNG1034.b2
ASNG1048.g2
ASNG1050.b2
ASNG1055.b2
ASNG1063.g2
ASNG1064.b2
ASNG1067.b2
ASNG1067.g2
ASNG1074.b2
ASNG1075.b2
ASNG1080.g2
ASNG1084.b2
ASNG1086.g2
ASNG1091.b2
ASNG1097.b2
ASNG1100.b2
ASNG1109.b2
ASNG1110.b2
ASNG1120.b2
ASNG1136.b2
ASNG1136.g2
ASNG1150.g2
ASNG1405.g2
ASNG1436.b2
ASNG1439.b2
ASNG1546.g2
ASNG1550.g2
ASNG1554.b2
ASNG1562.b2
ASNG1569.g02
ASNG1569.g2
ASNG1570.g02
ASNG1570.g2
ASNG1577.g02
ASNG1609.b2
ASNG1618.b2
ASNG1619.g2
ASNG1645.g2
ASNG1650.b2

ASNG1659.b2
ASNG1669.b2
ASNG1684.b2
ASNG1690.g2
ASNG1691.g2
ASNG1715.b2
ASNG1720.g02
ASNG1741.b2
ASNG1742.b2
ASNG1746.g2
ASNG1765.b2
ASNG1773.b2
ASNG1786.g2
ASNG1787.b2
ASNG1823.g02
ASNG1829.g2
ASNG1833.b2
ASNG1837.b2
ASNG1838.b2
ASNG1842.b2
ASNG1844.b2
ASNG1861.b2
ASNG1869.b2
ASNG1971.g2
ASNG2078.g2
ASNG2106.g2
ASNG2115.g2
ASNG2190.g2
ASNG2261.b2
ASNG2289.g2
ASNG2317.b2
ASNG2320.b2
ASNG2323.b2
ASNG2329.b2
ASNG2336.g2
ASNG2339.b2
ASNG2345.b2
ASNG2369.g2
ASNG2373.g2
ASNG2400.g2
ASNG2401.g2
ASNG2432.b2
ASNG2441.g2
ASNG2455.g2
ASNG2465.b2
ASNG2474.g2
ASNG2485.g2
ASNG2513.b2
ASNG2530.g2
ASNG2615.g2
ASNG2692.g2
ASNG2693.b2
ASNG2705.b2
ASNG2708.g2
ASNG2721.b2
ASNG2735.b2

ASNG2735.g2
ASNG2749.b2
ASNG2759.g2
ASNG2765.b2
ASNG2768.b2
ASNG2777.b2
ASNG2777.g2
ASNG2781.g2
ASNG2784.b2
ASNG2784.g2
ASNG2789.b2
ASNG2808.b2
ASNG2816.b2
ASNG2816.g2
ASNG2846.g2
ASNG2850.b2
ASNG2850.g2
ASNG2853.b2
ASNG2868.g2
ASNG2869.b2
ASNG2872.b2
ASNG2879.b2
ASNG2888.b2
ASNG2916.g2
ASNG2924.g2
ASNG2925.b2
ASNG2928.g2
ASNG2939.b2
ASNG2976.g2
ASNG2977.b2
ASNG2983.b2
ASNG2985.b2
ASNG2992.b2
ASNG2994.b2
ASNG2998.g2
ASNG3000.g2
ASNG3007.b2
ASNG3007.g2
ASNG3009.g2
ASNG3015.g2
ASNG3016.b2
ASNG3016.g2
ASNG3022.g2
ASNG3024.b2
ASNG3032.b2
ASNG3032.g2
ASNG3041.b2
ASNG3048.b2
ASNG3048.g2
ASNG3071.b2
ASNG3104.g2
ASNG3121.g2
ASNG3168.b2
ASNG3196.g2
ASNG3201.b2
ASNG3217.b2

ASNG3217.g2
ASNG3232.g2
ASNG3264.b2
ASNG3264.g2
ASNG3289.g2
ASNG3291.b2
ASNG3321.b2
ASNG3329.b2
ASNG3377.b2
ASNG3418.b2
ASNG3456.b2
ASNG3469.b2
ASNG3752.g2
ASNG3835.g2
ASNG3899.b2
ASNG3908.b2
ASNG3929.g2
ASNG4018.b2
ASNG4018.g2
ASNG4099.g2
ASNG4121.b2
ASNG4122.g2
ASNG4197.g2
ASNG4213.b2
ASNG4217.b2
ASNG4223.b2
ASNG4223.g2
ASNG469.b2
ASNG502.b2
ASNG515.b2
ASNG520.b2
ASNG520.g2
ASNG579.b2
ASNG590.b2
ASNG659.g2
ASNG680.b2
ASNG724.g2
ASNG772.b2
ASNG772.g2
ASNG774.b2
ASNG776.b2
ASNG785.g2
ASNG787.g2
ASNG788.b2
ASNG801.b2
ASNG804.g2
ASNG808.g2
ASNG809.b2
ASNG818.b2
ASNG823.g2
ASNG829.b2
ASNG829.g2
ASNG855.g2
ASNG871.g2
ASNG875.g2
ASNG877.b2

ASNG877.g2
ASNG901.b2
ASNG904.g2
ASNG918.b2
ASNG923.b2
ASNG928.b2
ASNG929.b2
ASNG934.b2
ASNG937.b2
ASNG945.b2
ASNG957.b2
ASNG958.g2
ASNG959.g2
ASNG960.b2
ASNG961.b2
ASNG964.b2
ASNG969.b2
ASNG986.g2
ASNG989.g2
ASNG992.b2
ASNG994.b2
ASNG996.g2
HF200_05_38TF
HF200_05_49TR
HF200_05_64TF
HF200_05_90TF
HF200_15_50TR
HF200_20_45TR
HF200_20_74TR
HF200_25_52TR
HF200_35_49TR
HF200_40_63TF
HF200_55_06TF

Other mountain regions:
ASNG1515.g2
ASNG1600.g02
ASNG1600.g2
ASNG1604.g02
ASNG1604.g2
ASNG2012.b2
ASNG2143.g2
ASNG2740.g2
ASNG2793.g2
ASNG3156.b2
ASNG3186.g2
ASNG3242.g2
ASNG3351.g2
ASNG3378.b2
ASNG3386.b2
ASNG3439.g2
ASNG3474.b2
ASNG3550.b2
ASNG3550.g2
ASNG3591.b2
ASNG3722.g2

ASNG3752.b2
ASNG3885.b2
ASNG3886.g2
ASNG3904.b2
ASNG3916.b2
ASNG3921.g2
ASNG3977.g2
ASNG4163.b2
ASNG4194.g2
ASNG557.g2
ASNG664.b2
ASNG728.g2
ASNG777.b2
ASNG793.b2
ASNG796.b2
ASNG992.g2
HF200_05_31TF
HF200_15_13TF
HF200_35_09TF
HF200_40_90TR
HF200_55_22TF
HF200_55_34TR

List 1.1 Closest fosmid end-sequence blastx hits, to either nr or HF130 seed sequences, of the HF200 fosmids that were 454 sequenced.

From among those end sequences that clustered within the group II archaeal seed sequences, I selected those fosmids for sequencing whose end sequences showed strong BLAST homology either to euryarchaeal housekeeping genes in nr, or to the seed fosmids themselves. Below are the top blastx or tblastx hits for the fosmids I had sequenced:

ASNG1177.b2 ribosomal protein L4 [uncultured marine group II euryarchaeote KM3-130-D10] E: = e-105

ASNG1177.g2 4Fe-4S ferredoxin iron-sulfur binding protein [uncultured marine group II euryarchaeote KM3-72-G3] E: 1e-69

ASNG1201.b2 AF268611_5 translation initiatin factor 6 [uncultured marine group II euryarchaeote 37F11] E: 2e-22

ASNG1201.g2 TPR repeat-containing protein [Methanosaeta thermophila PT] E: 3e-11

ASNG1226.b2 putative ATP synthase alpha/beta family, nucleotide-binding domain protein [uncultured marine microorganism HF4000_APKG8D23] E: = 9e-82

ASNG1226.g2 hypothetical protein ALOHA_HF4000APKG8D23ctg1g1 [uncultured marine microorganism HF4000_APKG8D23] E: 1e-69

ASNG1234.b2 putative ATP synthase alpha/beta family, nucleotide-binding domain protein [uncultured marine microorganism HF4000_APKG8D23] E: e-99

ASNG1234.g2 hypothetical protein ALOHA_HF4000APKG8D23ctg1g1 [uncultured marine microorganism HF4000_APKG8D23] E: 1e-69

ASNG1297.b2 4003660_fasta.screen.Contig2 E: e-100

ASNG1297.g2 4003655_fasta.screen.Contig15 E: e-150

ASNG1322.b2 4003655_fasta.screen.Contig15 E: e-150

ASNG1322.g2 4003716_fasta.screen.Contig3 E: e-107

alcohol dehydrogenase [Salinispora arenicola CNS-205] E: 3e-37

oxidoreductase, zinc-binding protein [Algoriphagus sp. PR1] E: 5e-35

ASNG1327.b2 4003703_fasta.screen.Contig2 E: e-112

Ribulose-phosphate 3-epimerase [Akkermansia muciniphila ATCC BAA-835] E: 5e-25

ASNG1327.g2 4003693_fasta.screen.Contig3 E: 4e-84

ASNG1375.b2 phosphoribosylglycinamide formyltransferase-like protein [uncultured organism HF10_3D09] E: 4e-46

ASNG1375.g2 4003662_fasta.screen.Contig4 E: e-109

novel protein similar to vertebrate uridine monophosphate synthetase (UMPS, zgc:55702) [Danio rerio] E: = 1e-27

orotidine-5'-monophosphate decarboxylase [Coccidioides posadasii] E: = 3e-27

ASNG1385.b2 4003648_fasta.screen.Contig2 E: e-169

glutamate synthase domain 2 [uncultured archaeon GZfos23H7] E: = 3e-96

ASNG1415.b2 4003661_fasta.screen.Contig3 E: e-103

ASNG1415.g2 hypothetical protein [uncultured organism HF10_3D09] E: 3e-22

ASNG1490.b2 4003711_fasta.screen.Contig6 E: e-143

ribosomal protein, L1P family [Aciduliprofundum boonei T469] E: 9e-18

ASNG1490.g2 4003706_fasta.screen.Contig5 E: 7e-28

ASNG1506.b2 4003703_fasta.screen.Contig2 E: e-125

glutaredoxin-like protein [Thiobacillus denitrificans ATCC 25259] E: 8e-22

ASNG1506.g2 hypothetical protein PGUG_03728 [Pichia guilliermondii ATCC 6260] E: 0.25
ASNG1556.b2 Oligosaccharyl transferase STT3 subunit family [Aciduliprofundum boonei T469] E: 2e-32
ASNG1556.g02 4003705_fasta.screen.Contig2 E: 6e-76
ASNG1556.g2 4003705_fasta.screen.Contig2 E: 2e-91
glycine dehydrogenase subunit 1 [Pyrococcus abyssi GE5] E: 1e-45
Glycine cleavage system P-protein [Thermococcus barophilus MP] E: 4e-45
ASNG1584.b2 tRNA (adenine-N(1)-)-methyltransferase [Acidothermus cellulolyticus 11B] E: 1e-16
ASNG1584.g02 4003698_fasta.screen.Contig2 E: e-152
ASNG1584.g2 4003698_fasta.screen.Contig2 E: e-163
ABC transporter ATP-binding protein [Symbiobacterium thermophilum IAM 14863] E: 4e-72
FeS assembly ATPase SufC [Rubrobacter xylanophilus DSM 9941] E: 4e-72
ASNG1617.b2 hypothetical protein ALOHA_HF4000ANIW137G21ctg1g2 [uncultured marine microorganism HF4000_ANIW137G21] E: = 2e-98
ASNG1617.g02 metallo-beta-lactamase superfamily protein [Stigmatella aurantiaca DW4/3-1] E: 1e-18
ASNG1617.g2 metallo-beta-lactamase superfamily protein [Stigmatella aurantiaca DW4/3-1] E: 1e-09
ASNG1633.b2 hypothetical protein ALOHA_HF4000ANIW133F6ctg1g8 [uncultured marine microorganism HF4000_ANIW133F6] E: 5e-61
ASNG1633.g02 4003707_fasta.screen.Contig14 E: 3e-43
ASNG1633.g2 4003650_fasta.screen.Contig1 E: 8e-77
ASNG1644.b2 4003695_fasta.screen.Contig5 E: 3e-71
Bacterial pre-peptidase C-terminal domain family [Thermococcus barophilus MP] E: 6e-07
ASNG1644.g02 putative SRP54-type protein, GTPase domain protein [uncultured marine microorganism HF4000_APKG2H5] E: 3e-66
ASNG1644.g2 putative SRP54-type protein, GTPase domain protein [uncultured marine microorganism HF4000_APKG2H5] E: 4e-92
ASNG1655.b2 4003655_fasta.screen.Contig15 E: 4e-28
peptidil-prolyl cis-trans isomerase [Clostridium acetobutylicum ATCC 824] E: 6e-15
ASNG1655.g02 4003696_fasta.screen.Contig2 E: 5e-26
ASNG1655.g2 4003700_fasta.screen.Contig2 E: 3e-35
biotin/lipoic acid binding domain-containing protein [Myxococcus xanthus DK 1622] E: 5e-08
putative acetyl-CoA carboxylase biotin carboxyl carrier protein subunit [Pyrococcus furiosus DSM 3638] E: 3e-06
ASNG1693.b2 hypothetical protein [uncultured euryarchaeote Alv-FOS4] E: = 4e-14
ASNG1693.g02 translation initiation factor IF-2 [Methanopyrus kandleri AV19] E: = 6e-41
ASNG1693.g2 translation initiation factor IF-2 [uncultured methanogenic archaeon RC-I] E: 5e-55
ASNG1734.b2 4003668_fasta.screen.Contig3 E: e-144
AAA ATPase containing von Willebrand factor type A (vWA) protein-like omain [Herpetosiphon aurantiacus ATCC 23779] E: 8e-08
ASNG1734.g02 4003647_fasta.screen.Contig1 E: 0.0
ASNG1734.g2 4003647_fasta.screen.Contig1 E: 0.0

putative binding-protein-dependent transport system inner membrane component [uncultured marine microorganism HF4000_ANIW141L21] E: = 8e-54

ASNG1769.b2 putative uncharacterized BCR, YnfA/UPF0060 family protein [uncultured marine microorganism HF4000_ANIW133F6] E: = 9e-70

ASNG1769.g02 DNA polymerase II, large subunit DP2 [Aciduliprofundum boonei T469] E: = 2e-23

ASNG1769.g2 DNA polymerase II large subunit [Methanosaeta thermophila PT] E: 5e-42

ASNG1823.b2 hypothetical protein AF2040 [Archaeoglobus fulgidus DSM 4304] E: 0.011

ASNG1823.g02 4003696_fasta.screen.Contig1 E: 0.58

ASNG1823.g2 4003696_fasta.screen.Contig1 E: 6e-50

putative pyrophosphohydrolase [Corynebacterium urealyticum DSM 7109] E: 7e-19

putative mutator mutT protein [Clostridium botulinum B1 str. Okra] E: 3e-18

ASNG1830.b2 4003655_fasta.screen.Contig15 E: e-114

eukaryotic peptide chain release factor subunit 1 [uncultured marine group III euryarchaeote AD1000-40-D7] E: 7e-50

ASNG1830.g2 hypothetical protein ALOHA_HF4000ANIW141L21ctg1g14 [uncultured marine microorganism HF4000_ANIW141L21] E: = e-114

ASNG1872.b2 4003703_fasta.screen.Contig2 E: 5e-73

tRNA-splicing endonuclease positive effector [Micromonas pusilla CCMP1545] E: 8e-25

ASNG1872.g2 putative phosphoenolpyruvate carboxykinase [uncultured marine microorganism HF4000_ANIW133F6] E: 5e-21

ASNG1878.b2 4003652_fasta.screen.Contig1 E: 3e-65

regulator of chromosome condensation, RCC1 [Roseiflexus sp. RS-1] E: = 4e-47

ASNG1878.g2 4003694_fasta.screen.Contig4 E: e-100

phosphoserine aminotransferase [Desulfotalea psychrophila LSv54] E: 1e-77

ASNG1882.b2 4003648_fasta.screen.Contig2 E: 7e-25

putative membrane-associated protease [Clostridium botulinum A3 str. Loch Maree] E: 2e-11

ASNG1882.g2 tRNA pseudouridine synthase A [Thermococcus onnurineus NA1] E: 2e-06

ASNG1950.b2 4003707_fasta.screen.Contig14 E: 3e-84

ASNG1950.g2 hypothetical protein ALOHA_HF4000APKG10F17ctg1g36 [uncultured marine microorganism HF4000_APKG10F17] E: = 3e-33

ASNG1965.b2 4003653_fasta.screen.Contig1 E: 0.0

D-aminoacylase [Rhodopirellula baltica SH 1] E: = 5e-27

ASNG1965.g2 Mn2+-dependent serine/threonine protein kinase [Methanopyrus kandleri AV19] E: 8e-16

ASNG2011.b2 4003695_fasta.screen.Contig5 E: 5e-65

Patched family [Aciduliprofundum boonei T469] E: 1e-12

ASNG2011.g2 4003704_fasta.screen.Contig2 E: e-157

thymidylate synthase [uncultured marine bacterium HF10_29C11] E: = e-106

ASNG2037.b2 4003653_fasta.screen.Contig1 E: e-174

phosphatase [Pelobacter carbinolicus DSM 2380] E: 4e-16

ASNG2037.g2 protein kinase domain [Synechococcus sp. PCC 7335] E: 6e-13

ASNG2071.b2 4003698_fasta.screen.Contig2 E: e-166

putative PKD domain protein [uncultured marine microorganism HF4000_APKG8D23] E: 3e-88

ASNG2071.g2 4003641_fasta.screen.Contig2 E: 7e-06

ASNG2075.b2 4003651_fasta.screen.Contig2 E: e-141

DNA-directed RNA polymerase subunit alpha [Methanosphaera stadtmanae DSM 3091] E: 2e-56

ASNG2075.g2 cobalt-precorrin-6x reductase [Frankia alni ACN14a] E: 0.015

ASNG2133.b2 hypothetical protein ALOHA_HF4000APKG2H5ctg1g12 [uncultured marine microorganism HF4000_APKG2H5] E: 9e-63

ASNG2133.g2 putative transketolase, C-terminal domain protein [uncultured marine microorganism HF4000_APKG1C9] E: 1e-55

ASNG2186.b2 4003657_fasta.screen.Contig1 E: 1e-66

L-lysine 6-aminotransferase [Stenotrophomonas sp. SKA14] E: 7e-38

ASNG2186.g2 4003654_fasta.screen.Contig1 E: e-133

ABC transporter ATP-binding protein [Symbiobacterium thermophilum IAM 14863] E: 8e-40

ASNG2245.b2 4003690_fasta.screen.Contig3 E: 4e-63

DNA mismatch repair protein [Salinibacter ruber DSM 13855] E: = 5e-56

ASNG2245.g2 V-type ATP synthase subunit A [Methanoculleus marisnigri JR1] E: e-126

ASNG2258.b2 putative FG-GAP repeat protein [uncultured marine microorganism HF4000_ANIW141L21] E: = e-132

ASNG2258.g2 4003668_fasta.screen.Contig3 E: 7e-79

ASNG2266.b2 4003696_fasta.screen.Contig2 E: 4e-90

methylmalonyl-CoA mutase, subunit alpha, C-terminus [Pyrococcus abyssi GE5] E: 1e-29

ASNG2266.g2 4003696_fasta.screen.Contig1 E: 5e-58

Pyroglutamyl-peptidase I [Thermus aquaticus Y51MC23] E: 9e-07

ASNG2430.b2 excinuclease ATPase subunit [uncultured organism HF10_3D09] E: = 3e-65

ASNG2430.g2 DNA topoisomerase VI, B subunit [Aciduliprofundum boonei T469] E: 8e-35

ASNG2646.b2 glutamyl-tRNA(Gln) amidotransferase subunit D [Methanocaldococcus jannaschii DSM 2661] E: 6e-09

ASNG2646.g2 hypothetical protein ALOHA_HF4000ANIW137G21ctg1g21 [uncultured marine microorganism HF4000_ANIW137G21] E: 1e-50

ASNG2819.b2 putative DEAD 2 [uncultured marine microorganism HF4000_ANIW141L21] E: 3e-43

ASNG2819.g2 AF268611_25 queuine tRNA-ribosyltransferase [uncultured marine group II euryarchaeote 37F11] E: 6e-64

ASNG2826.b2 4003643_fasta.screen.Contig1 E: e-102

PKD domain containing protein [Natrialba magadii ATCC 43099] E: 4e-06

ASNG2826.g2 4003702_fasta.screen.Contig3 E: 3e-40

putative X-Pro dipeptidyl-peptidase (S15 family) [uncultured marine microorganism HF4000_ANIW137J11] E: 1e-32

ASNG2962.b2 4003704_fasta.screen.Contig2 E: 1e-28

ASNG2962.g2 4003651_fasta.screen.Contig2 E: e-169

DNA-directed RNA polymerase, beta subunit, putative [Aciduliprofundum boonei T469] E: 4e-93

ASNG3023.b2 4003706_fasta.screen.Contig5 E: 3e-68

putative amino acid or sugar ABC transport system, permease protein [Labrenzia alexandrii DFL-11] E: 1e-09

ASNG3023.g2 putative Thrombospondin type 3 repeat [uncultured marine microorganism HF4000_001N02] E: = 2e-48

ASNG3084.b2 acetyl-CoA acetyltransferase [Candidatus Korarchaeum cryptofilum OPF8] E: 3e-19

ASNG3084.g2 4003665_fasta.screen.Contig3 E: 0.0

AAA family ATPase, CDC48 subfamily protein [Methanococcoides burtonii DSM 6242] E: 9e-09

ASNG3147.b2 putative PKD domain protein [uncultured marine microorganism HF4000_APKG8D23] E: 1e-87

ASNG3147.g2 putative Ketopantoate reductase PanE/ApbA [uncultured marine microorganism HF4000_APKG8D23] E: 6e-19

ASNG3149.b2 4003653_fasta.screen.Contig1 E: e-102

ASNG3149.g2 putative RecF/RecN/SMC N terminal domain protein [uncultured marine microorganism HF4000_APKG8D23] E: 8e-63

ASNG3159.b2 thymidylate kinase [Thermococcus onnurineus NA1] E: 6e-09

ASNG3159.g2 putative Ketopantoate reductase PanE/ApbA [uncultured marine microorganism HF4000_APKG8D23] E: 8e-37

ASNG3171.b2 thermosome beta subunit [Aeropyrum pernix K1] E: = 3e-32

ASNG3171.g2 4003693_fasta.screen.Contig3 E: 7e-75

ASNG3172.b2 rRNA methylase [uncultured marine group II euryarchaeote HF70_59C08] E: 3e-44

ASNG3172.g2 4003714_fasta.screen.Contig13 E: 2e-61

ribonucleotide-diphosphate reductase subunit alpha [Leptospira borgpetersenii serovar Hardjo-bovis L550] E: 1e-14

ASNG3208.b2 putative DEAD/DEAH box helicase [uncultured marine microorganism HF4000_APKG2H5] E: e-135

ASNG3208.g2 putative uncharacterized ACR, COG2107 [uncultured marine microorganism HF4000_APKG1C9] E: 6e-34

ASNG3228.b2 4003649_fasta.screen.Contig2 E: e-109

ASNG3228.g2 hypothetical protein ALOHA_HF1029C11.0035 [uncultured marine bacterium HF10_29C11] E: 2e-49

ASNG3235.b2 4003695_fasta.screen.Contig5 E: 1e-04

ASNG3235.g2 4003668_fasta.screen.Contig3 E: e-109

predicted calcium-binding protein [Pseudoalteromonas tunicata D2] E: 7e-14

AAA ATPase containing von Willebrand factor type A (vWA) protein-like omain [Herpetosiphon aurantiacus ATCC 23779] E: 6e-11

ASNG3277.b2 4003660_fasta.screen.Contig2 E: 0.0

acetyl-CoA carboxylase biotin carboxylase subunit [Rhodococcus jostii RHA1] E: = 7e-61

ASNG3277.g2 putative X-Pro dipeptidyl-peptidase (S15 family) [uncultured marine microorganism HF4000_ANIW137J11] E: 9e-07

ASNG3284.b2 4003709_fasta.screen.Contig9 E: e-109

pyruvate dehydrogenase subunit E1 [Planctomyces maris DSM 8797] E: = 3e-91

ASNG3284.g2 glycosyltransferase [Cenarchaeum symbiosum A] E: 2e-06

ASNG3288.b2 4003657_fasta.screen.Contig1 E: e-106

ASNG3288.g2 4003653_fasta.screen.Contig1 E: 3e-12

ASNG3300.b2 4003694_fasta.screen.Contig4 E: 0.10

ASNG3300.g2 4003653_fasta.screen.Contig1 E: e-161

ASNG3302.b2 putative transglutaminase-like superfamily protein [uncultured marine microorganism HF4000_ANIW141L21] E: 3e-81

ASNG3302.g2 AF268611_35 giant membrane protein [uncultured marine group II euryarchaeote 37F11] E: 9e-17

ASNG3307.b2 30S ribosomal protein S19e [Methanocorpusculum labreanum Z] E: 3e-29

ASNG3307.g2 putative DNA primase small subunit [uncultured marine microorganism HF4000_APKG1C9] E: 2e-37

ASNG3309.b2 4003694_fasta.screen.Contig4 E: 0.0

pyrroline-5-carboxylate reductase [Thermobifida fusca YX] E: 4e-44

ASNG3309.g2 4003706_fasta.screen.Contig5 E: 0.67

ASNG3311.b2 assimilatory nitrate reductase large subunit [uncultured marine group II euryarchaeote DeepAnt-15E7] E: 5e-72

ASNG3311.g2 carbamoyl-phosphate synthase ammonia chain [uncultured marine group II euryarchaeote AD1000-18-D2] E: e-103

ASNG3316.b2 AF268611_22 unknown [uncultured marine group II euryarchaeote 37F11] E: 1e-20

ASNG3316.g2 4003660_fasta.screen.Contig2 E: 0.052

ASNG3329.b2 4003666_fasta.screen.Contig3 E: 2.3

ASNG3329.g2 4003655_fasta.screen.Contig15 E: e-169

ASNG3336.b2 4003696_fasta.screen.Contig2 E: e-155

ASNG3336.g2 AF268611_35 giant membrane protein [uncultured marine group II euryarchaeote 37F11] E: 1e-17

ASNG3347.b2 4003641_fasta.screen.Contig2 E: 3e-98

conserved repeat domain protein [gamma proteobacterium NOR5-3] E: 7e-04

ASNG3347.g2 putative elongation factor Tu GTP binding domain protein [uncultured marine crenarchaeote HF4000_APKG10I20] E: e-138

ASNG3350.b2 Oligosaccharyl transferase STT3 subunit family [Aciduliprofundum boonei T469] E: 1e-05

ASNG3350.g2 4003716_fasta.screen.Contig3 E: 1e-31

ASNG3363.b2 4003694_fasta.screen.Contig4 E: 0.0

ASNG3363.g2 endopeptidase La [Halobacterium salinarum R1] E: 8e-31

ASNG3383.b2 DNA polymerase B region [Halorubrum lacusprofundi ATCC 49239] E: 2e-37

ASNG3383.g2 translation initiation factor SUI1 [uncultured marine group II euryarchaeote HF70_59C08] E: 1e-41

ASNG3389.b2 4003662_fasta.screen.Contig3 E: e-109

Type IIB DNA topoisomerase family [Aciduliprofundum boonei T469] E: = 7e-60

ASNG3389.g2 glycosyl transferase group 1 [Natrialba magadii ATCC 43099] E: 5e-04

ASNG3403.b2 putative ABC transporter [uncultured marine microorganism HF4000_ANIW137G21] E: = 5e-18

ASNG3403.g2 domain repeat protein [uncultured marine group II euryarchaeote HF70_39H11] E: 5e-46

ASNG3415.b2 ribosomal protein L32 [uncultured marine group II euryarchaeote KM3-85-F5] E: 2e-63

ASNG3415.g2 phospholipid-lipopolysaccharide ABC transporter [uncultured marine group II euryarchaeote KM3-130-D10] E: e-141

ASNG3433.b2 aspartate/tyrosine/aromatic aminotransferase [uncultured euryarchaeote Alv-FOS1] E: 5e-40

ASNG3433.g2 putative FG-GAP repeat protein [uncultured marine microorganism HF4000_ANIW141L21] E: 7e-48

ASNG3505.b2 4003663_fasta.screen.Contig4 E: 2e-31

ASNG3505.g2 putative 4Fe-4S binding domain protein [uncultured marine microorganism HF4000_ANIW137G21] E: 3e-36

ASNG3517.b2 4003704_fasta.screen.Contig2 E: 6e-47

putative DNA primase small subunit [uncultured marine microorganism HF4000_APKG1C9] E: = 7e-24

ASNG3517.g2 50S ribosomal protein L39e [Thermococcus kodakarensis KOD1] E: 7e-15

ASNG3627.b2 iron dependent transcriptional repressor [uncultured marine bacterium HF10_29C11] E: 8e-24

ASNG3627.g2 conserved domain protein [Thermococcus barophilus MP] E: = 2e-20

ASNG3646.b2 4003697_fasta.screen.Contig1 E: 1e-82

ASNG3646.g2 4003711_fasta.screen.Contig6 E: 4e-59

ASNG3647.b2 4003655_fasta.screen.Contig15 E: 1e-72

ASNG3647.g2 4003696_fasta.screen.Contig1 E: e-111

ASNG3673.b2 putative AIR synthase related protein, C-terminal domain protein [uncultured marine microorganism HF4000_001N02] E: 5e-50

ASNG3673.g2 4003650_fasta.screen.Contig1 E: 4e-55

major facilitator superfamily MFS_1 [Halothermothrix orenii H 168] E: = 1e-15

ASNG3680.b2 4003655_fasta.screen.Contig15 E: 2e-49

hypothetical membrane protein, conserved [Thermococcus onnurineus NA1] E: 3e-09

ASNG3680.g2 transcription initiation factor IIB [Thermococcus onnurineus NA1] E: 1e-31

ASNG3700.b2 A Chain A, [T. Acidophilum] Thermosome-Mg-Adp-Alf3 Complex E: 8e-36

ASNG3700.g2 4003654_fasta.screen.Contig1 E: e-124

carbamoyl-phosphate synthase L chain, ATP-binding [Frankia sp. CcI3] E: 9e-76

ASNG3756.b2 putative peptidase family M28 [uncultured marine microorganism HF4000_ANIW137J11] E: 2e-34

ASNG3756.g2 4003654_fasta.screen.Contig1 E: 2e-69

ASNG3798.b2 carbamoyl-phosphate synthase ammonia chain [uncultured marine group II euryarchaeote AD1000-18-D2] E: e-155

ASNG3798.g2 4003655_fasta.screen.Contig15 E: e-112

branched-chain alpha-keto acid dehydrogenase E2 component [Sphingomonas wittichii RW1] E: = 2e-47

ASNG3808.b2 4003692_fasta.screen.Contig3 E: e-101

endonuclease IV [Staphylothermus marinus F1] E: 2e-10

ASNG3808.g2 4003656_fasta.screen.Contig2 E: e-143

Aldehyde oxidase and xanthine dehydrogenase, molybdopterin binding protein [Plesiocystis pacifica SIR-1] E: 6e-56

ASNG3822.b2 4003702_fasta.screen.Contig3 E: e-101

HAD family hydrolase [Thermotoga lettingae TMO] E: = 2e-42

ASNG3822.g2 4003714_fasta.screen.Contig13 E: e-154

B12-dependent ribonucleoside diphosphate reductase [Thermomicrobium roseum DSM 5159] E: = 2e-87

ASNG3827.b2 4003693_fasta.screen.Contig3 E: 1e-04

ASNG3827.g2 4003704_fasta.screen.Contig2 E: 1e-51

putative transketolase, C-terminal domain protein [uncultured marine microorganism HF4000_APKG1C9] E: 5e-27

ASNG3832.b2 4003692_fasta.screen.Contig3 E: 5e-49

ASNG3832.g2 4003656_fasta.screen.Contig2 E: e-143

Aldehyde oxidase and xanthine dehydrogenase, molybdopterin binding protein [Plesiocystis pacifica SIR-1] E: 5e-53

ASNG3842.b2 putative ATPase RIL [Methanosaeta thermophila PT] E: 8e-58

ASNG3842.g2 assimilatory nitrate reductase large subunit [uncultured marine group II euryarchaeote DeepAnt-15E7] E: 9e-34

ASNG3861.b2 4003707_fasta.screen.Contig14 E: 2e-33

MDR-type permease [Bdellovibrio bacteriovorus HD100] E: 3e-25

ASNG3861.g2 4003698_fasta.screen.Contig2 E: e-123

peptidase [uncultured marine group II euryarchaeote KM3-85-F5] E: 2e-04

ASNG3874.b2 4003697_fasta.screen.Contig1 E: 6e-35

ASNG3874.g2 4003697_fasta.screen.Contig1 E: 1e-32

ASNG389.b2 cell division protein FtsZ [Methanosarcina acetivorans C2A] E: 1e-77

ASNG389.g2 ribosomal protein, L1P family [Aciduliprofundum boonei T469] E: 2e-54

ASNG3941.b2 Type IIB DNA topoisomerase family [Aciduliprofundum boonei T469] E: 7e-95

ASNG3941.g2 4003662_fasta.screen.Contig4 E: 9e-85

response regulator receiver protein [Desulfonatronospira thiodismutans ASO3-1] E: 6e-06

ASNG3957.b2 4003700_fasta.screen.Contig2 E: 3e-65

adenylate kinase [Alkalilimnicola ehrlichei MLHE-1] E: 2e-24

ASNG3957.g2 AF268611_35 giant membrane protein [uncultured marine group II euryarchaeote 37F11] E: 1e-12

ASNG397.b2 flotillin 1 [uncultured marine group II euryarchaeote KM3-72-G3] E: 2e-49

ASNG397.g2 ribosomal protein L3 [uncultured marine group II euryarchaeote DeepAnt-JyKC7] E: = e-101

ASNG3981.b2 4003664_fasta.screen.Contig6 E: 0.054

ASNG3981.g2 4003656_fasta.screen.Contig2 E: e-100

ASNG4051.b2 hypothetical protein ALOHA_HF4000001N02ctg1g22 [uncultured marine microorganism HF4000_001N02] E: = 1e-45

ASNG4051.g2 4003662_fasta.screen.Contig5 E: e-155

ASNG4053.b2 hypothetical protein RGAI101_3796 [Roseobacter sp. GAI101] E: 2e-14

ASNG4053.g2 4003665_fasta.screen.Contig3 E: e-155

holliday junction resolvase Hjc, putative [Aciduliprofundum boonei T469] E: = 7e-14

ASNG4112.b2 peptidase, M50 family, putative [Aciduliprofundum boonei T469] E: 1e-16

ASNG4112.g2 DEAD/DEAH box helicase domain protein [Natrialba magadii ATCC 43099] E: 3e-12

ASNG432.b2 4003714_fasta.screen.Contig13 E: 2e-68

ASNG432.g2 4003662_fasta.screen.Contig5 E: e-108

ASNG529.b2 putative cation transport protein [uncultured marine microorganism HF4000_APKG2H5] E: 3e-46

ASNG529.g2 putative tRNA synthetase class II core domain (G, H, P, S and T) [uncultured marine microorganism HF4000_APKG1C9] E: 4e-90

ASNG544.b2 peptidase [uncultured marine group II euryarchaeote DeepAnt-JyKC7] E: = 9e-47

ASNG544.g2 hypothetical protein [uncultured marine group II euryarchaeote KM3-130-D10] E: 2e-37

ASNG605.b2 molybdopterin oxidoreductase Fe4S4 region [uncultured marine group II euryarchaeote KM3-130-D10] E: 4e-96

ASNG605.g2 4003655_fasta.screen.Contig15 E: 2e-47

lipoyl synthase [Deinococcus radiodurans R1] E: 2e-20

ASNG637.b2 hypothetical protein ALOHA_HF4000ANIW141L21ctg1g16 [uncultured marine microorganism HF4000_ANIW141L21] E: 3e-43

ASNG637.b2 4003647_fasta.screen.Contig1 E: = 1e-26

ASNG639.b2 peptidase [uncultured marine group II euryarchaeote KM3-85-F5] E: e-116

ASNG639.g2 amino acid ABC transporter periplasmic amino acid-binding protein [uncultured marine group II euryarchaeote KM3-72-G3] E: 7e-30

ASNG667.b2 L-isoaspartyl protein carboxyl methyltransferase [Methanothermobacter thermautotrophicus str. Delta H] E: 5e-05

ASNG667.g2 4003705_fasta.screen.Contig2 E: e-167

serine hydroxymethyltransferase [Aciduliprofundum boonei T469] E: 7e-79

ASNG682.b2 4003661_fasta.screen.Contig E: = e-132

TPR repeat-like protein [uncultured organism HF10_3D09] E: 4e-29

ASNG682.g2 4003706_fasta.screen.Contig4 E: = 6e-23

Extracellular ligand-binding receptor [Desulfovibrio vulgaris str. 'Miyazaki F'] E: 1e-07

ASNG735.b2 site-specific recombinase, phage integrase family protein [Aciduliprofundum boonei T469] E: 8e-29

ASNG735.g2 4003657_fasta.screen.Contig1 E: = 1e-90

deoxyhypusine synthase [Aciduliprofundum boonei T469] E: 3e-31

ASNG740.b2 4003711_fasta.screen.Contig6 E: = e-178

putative DnaJ domain protein [uncultured marine microorganism HF4000_ANIW133F6] E: 8e-83

ASNG740.g2 4003697_fasta.screen.Contig1 E: = 1e-56

HF200_15_66TF CoB-CoM heterodisulfide reductase subunit B [uncultured marine group II euryarchaeote DeepAnt-15E7] E: = e-105

HF200_15_66TR 4Fe-4S ferredoxin iron-sulfur binding protein [uncultured marine group II euryarchaeote KM3-72-G3] E: 1e-98

HF200_15_82TF ribosomal protein L3 [uncultured marine group II euryarchaeote DeepAnt-JyKC7] E: 8e-50

HF200_15_82TR integral membrane protein [uncultured marine group II euryarchaeote KM3-85-F5] E: = 4e-39

HF200_15_85TF 4003662_fasta.screen.Contig5 E: = e-109

HF200_15_85TR 4003714_fasta.screen.Contig13 E: E: = e-112

putative RNA polymerases N / 8 kDa subunit [uncultured marine microorganism HF4000_001N02] E: 7e-28

HF200_15_87TF 4003704_fasta.screen.Contig2 E: = 4e-15

HF200_15_87TR 4003664_fasta.screen.Contig7 E: = e-113

putative KH domain protein [uncultured marine microorganism HF4000_APKG2H5] E: = 8e-61

HF200_20_49TF 4003653_fasta.screen.Contig1 E: e-135

D-aminoacylase [Rhodopirellula baltica SH 1] E: 2e-21

HF200_20_49TR hypothetical protein ALOHA_HF4000ANIW137G21ctg1g41 [uncultured marine microorganism HF4000_ANIW137G21] E: 1e-59

HF200_25_12TF 4003697_fasta.screen.Contig1 E: = e-106

AF268611_11 membrane-associated subtilysin-type serine protease [uncultured marine group II euryarchaeote 37F11] E: 2e-38

HF200_25_12TR 4003711_fasta.screen.Contig6 E: = e-102

HF200_25_48TF 4003655_fasta.screen.Contig15 E: = e-105

peptide chain release factor eRF/aRF, subunit 1 [Aciduliprofundum boonei T469] E: 1e-34

HF200_25_48TR amino acid ABC transporter periplasmic amino acid-binding protein [uncultured marine group II euryarchaeote AD1000-18-D2] E: e-113

HF200_25_64TF 4003709_fasta.screen.Contig9 E: = e-146

carboxypeptidase Taq [Opitutus terrae PB90-1] E: 2e-52

HF200_25_64TR malate dehydrogenase [Halorubrum lacusprofundi ATCC 49239] E: 6e-73

HF200_25_67TF preprotein translocase secY subunit [uncultured marine group II euryarchaeote AD1000-18-D2] E: e-137

HF200_25_67TR synthetase class I [uncultured marine group II euryarchaeote EF100_57A08] E: = 7e-99

HF200_35_43TF 4003655_fasta.screen.Contig15 E: = 1e-74

2-oxoisovalerate dehydrogenase alpha subunit [Oceanicaulis alexandrii HTCC2633] E: 7e-44

HF200_35_43TR 4003716_fasta.screen.Contig3 E: = 1e-42

TraB/PrgY-like protein [uncultured marine group II euryarchaeote KM3-72-G3] E: 2e-35
HF200_35_46TF 4003662_fasta.screen.Contig4 E: = e-110
phosphoribosylformylglycinamidine synthase I [uncultured organism HF10_3D09] E: 5e-56
HF200_35_46TR ubiquinone biosynthesis protein [Methanosarcina acetivorans C2A] E: 7e-12
HF200_35_50TF carbamoyl-phosphate synthase ammonia chain [uncultured marine group II euryarchaeote AD1000-18-D2] E: e-140
HF200_35_50TR ribosomal protein L24 [uncultured marine group II euryarchaeote DeepAnt-15E7] E: 8e-63
HF200_35_54TF 4003656_fasta.screen.Contig3 E: = 2e-53
HF200_35_54TR 7-cyano-7-deazaguanine tRNA-ribosyltransferase [Methanococcus aeolicus Nankai-3] E: 6e-54
HF200_35_70TF 4003711_fasta.screen.Contig6 E: e-122
50S ribosomal protein L24A [Methanocorpusculum labreanum Z] E: 1e-13
HF200_35_70TR 4003652_fasta.screen.Contig1 E: = 4e-47
HF200_40_29TF putative phosphoenolpyruvate carboxykinase [uncultured marine microorganism HF4000_ANIW133F6] E: e-112
HF200_40_29TR 4003709_fasta.screen.Contig9 E: = e-123
DP2L_METBU DNA polymerase II large subunit (Pol II) E: 1e-59
HF200_40_34TF HF200_40_34TF 4003716_fasta.screen.Contig3 E: = e-113
FeS cluster containing hydrogenase components 1 [uncultured marine group II euryarchaeote EF100_57A08] E: = 2e-82
HF200_40_34TR amino acid ABC transporter periplasmic amino acid-binding protein [uncultured marine group II euryarchaeote KM3-72-G3] E: 2e-62
HF200_40_80TF ATPase, P-type (transporting), HAD superfamily, subfamily IC [Tolumonas auensis DSM 9187] E: 4e-32
HF200_40_80TR 4003712_fasta.screen.Contig1 E: = e-146
putative DEAD/DEAH box helicase [uncultured marine microorganism HF4000_APKG2H5] E: = 1e-74
HF200_40_88TF 4003692_fasta.screen.Contig3 E: = e-121
propionyl-CoA carboxylase, beta subunit [Plesiocystis pacifica SIR-1] E: 4e-70
HF200_40_88TR putative ABC transporter [uncultured marine microorganism HF4000_ANIW137G21] E: = 3e-52
HF200_45_03TF 4003715_fasta.screen.Contig6 E: = e-104
putative ATP synthase alpha/beta family, nucleotide-binding domain protein [uncultured marine microorganism HF4000_APKG8D23] E: 2e-44
HF200_45_03TR HF200_45_03TR 4003698_fasta.screen.Contig2 E: = e-101
putative DNA gyrase/topoisomerase IV, subunit A [uncultured marine microorganism HF4000_APKG8D23] E: = 1e-81
HF200_45_30TF phenylalanyl-tRNA synthetase, alpha subunit, putative [Aciduliprofundum boonei T469] E: 1e-57
HF200_45_30TR hypothetical protein Mlab_0629 [Methanocorpusculum labreanum Z] E: 5e-10
HF200_45_47TF 4003707_fasta.screen.Contig14 E: = e-133
hypothetical protein MXAN_1546 [Myxococcus xanthus DK 1622] E: = 4e-27
HF200_45_47TR 4003650_fasta.screen.Contig1 E: = 2e-62
HF200_45_67TF 4003704_fasta.screen.Contig2 E: = 3e-97
putative CBS domain protein [uncultured marine microorganism HF4000_APKG1C9] E: 5e-95

HF200_45_67TR 4003664_fasta.screen.Contig5 E: e-155
putative DEAD/DEAH box helicase [uncultured marine microorganism
HF4000_APKG2H5] E: e-130
HF200_55_04TF 4003706_fasta.screen.Contig5 E: = 3e-53
threonine aldolase [Symbiobacterium thermophilum IAM 14863] E: = 1e-34
HF200_55_04TR 4003707_fasta.screen.Contig14 E: e-121

Notice that most end sequences showed a blastx homology to existing group II sequences with an E-value greater than 1e-60.

The following list of fosmids was sequenced:
ASNG1177
ASNG1201
ASNG1226
ASNG1234
ASNG1297
ASNG1322
ASNG1327
ASNG1375
ASNG1385
ASNG1415
ASNG1490
ASNG1506
ASNG1556
ASNG1584
ASNG1617
ASNG1633
ASNG1644
ASNG1655
ASNG1693
ASNG1734
ASNG1769
ASNG1823
ASNG1830
ASNG1872
ASNG1878
ASNG1882
ASNG1950
ASNG1965
ASNG2011
ASNG2037
ASNG2071
ASNG2075
ASNG2133
ASNG2186
ASNG2245
ASNG2258
ASNG2266
ASNG2430
ASNG2646
ASNG2819
ASNG2826
ASNG2962
ASNG3023
ASNG3084
ASNG3147
ASNG3149
ASNG3159
ASNG3171

ASNG3172
ASNG3208
ASNG3228
ASNG3235
ASNG3277
ASNG3284
ASNG3288
ASNG3300
ASNG3302
ASNG3307
ASNG3309
ASNG3311
ASNG3316
ASNG3329
ASNG3336
ASNG3347
ASNG3350
ASNG3363
ASNG3383
ASNG3389
ASNG3403
ASNG3415
ASNG3433
ASNG3505
ASNG3517
ASNG3627
ASNG3646
ASNG3647
ASNG3673
ASNG3680
ASNG3700
ASNG3756
ASNG3798
ASNG3808
ASNG3822
ASNG3827
ASNG3832
ASNG3842
ASNG3861
ASNG3874
ASNG389
ASNG3941
ASNG3957
ASNG397
ASNG3981
ASNG4051
ASNG4053
ASNG4112
ASNG432
ASNG529
ASNG544
ASNG605
ASNG637
ASNG639
ASNG667
ASNG682
ASNG735
ASNG740
HF200_15_66
HF200_15_82
HF200_15_85
HF200_15_87

HF200_20_49
HF200_25_12
HF200_25_48
HF200_25_64
HF200_25_67
HF200_35_43
HF200_35_46
HF200_35_50
HF200_35_54
HF200_35_70
HF200_40_29
HF200_40_34
HF200_40_80
HF200_40_88
HF200_45_03
HF200_45_30
HF200_45_47
HF200_45_67
HF200_55_04

List 1.2 HF200 library fosmid names assigned full length (over 10 kb) contig names

The matches below are to *any* long contig, not just those deemed to be group II. There is a natural break at about 25 kbp. I guessed that sequences above this length are approximately full length. Of those 82 contigs, 52 were assigned, by blastn homology to the fosmid end sequences of the fosmids I had sequenced. Following is the list of sequenced fosmid names with assembly contig names, if applicable.
ASNG1177
ASNG1201
ASNG1226 contig00630
ASNG1234 contig00630
ASNG1297 contig00423 (contig only shows homology to one of the two fosmid ends)
ASNG1322 contig00374
ASNG1327 contig00003, contig00602
ASNG1375 contig00451
ASNG1385 contig00039 (contig only shows homology to one of the two fosmid ends)
ASNG1415 contig00621
ASNG1490 contig00074, contig00018
ASNG1506 contig00392
ASNG1556 contig00127, contig00418
ASNG1584 contig00105
ASNG1617 contig00554
ASNG1633 contig00074
ASNG1644 contig00431, contig00391
ASNG1655 contig00478 (contig only shows homology to one of the two fosmid ends)
ASNG1693 contig00399
ASNG1734 contig00360
ASNG1769 contig00534
ASNG1823 contig00127, contig00038
ASNG1830 contig00531
ASNG1872 contig00405
ASNG1878 contig00059 (most likely: but bitscore for ASNG1878.b2 is only 127)
ASNG1882
ASNG1950 contig00424, contig00132
ASNG1965 contig00507 (contig only shows homology to one of the two fosmid ends)
ASNG2011 contig00542 (contig only shows homology to one of the two fosmid ends)
ASNG2037 contig00341, contig00549
ASNG2071 contig00618
ASNG2075 contig00454 (contig only shows homology to one of the two fosmid ends)
ASNG2133
ASNG2186 contig00627
ASNG2245 contig00072

ASNG2258 contig00513, contig00604
ASNG2266
ASNG2430 contig00071, contig00444
ASNG2646
ASNG2819
ASNG2826 contig00370
ASNG2962 contig00615
ASNG3023 contig00424 (?? -- only one end shows homology, and the homology to that end is 5kb away from the end of the contig)
ASNG3084
ASNG3147
ASNG3149
ASNG3159
ASNG3171
ASNG3172
ASNG3208
ASNG3228
ASNG3235
ASNG3277
ASNG3284
ASNG3288
ASNG3300
ASNG3302
ASNG3307
ASNG3309
ASNG3311
ASNG3316
ASNG3329
ASNG3336
ASNG3347
ASNG3350
ASNG3363
ASNG3383
ASNG3389
ASNG3403
ASNG3415 contig00533 (?? -- only one end shows homology, and the homology to that end is 10kb away from the end
of the contig)
ASNG3433
ASNG3505 contig00480, contig00131
ASNG3517 contig00082, contig00441
ASNG3627 contig00454, contig00542
ASNG3646 contig00443
ASNG3647 contig00164 (contig only shows homology to one of the two fosmid ends)
ASNG3673 contig00017
ASNG3680 contig00064, contig00062
ASNG3700 contig00113 (contig only shows homology to one of the two fosmid ends)
ASNG3756 contig00532
ASNG3798 contig00373
ASNG3808 contig00415
ASNG3822 contig00402

ASNG3827 contig00607
ASNG3832 contig00415
ASNG3842
ASNG3861
ASNG3874
ASNG389 contig00528
ASNG3941
ASNG3957
ASNG397 contig00597, contig00380
ASNG3981
ASNG4051
ASNG4053
ASNG4112
ASNG432 contig00437
ASNG529
ASNG544 contig00530 (contig only shows homology to one of the two fosmid ends)
ASNG605 contig00478 (contig only shows homology to one of the two fosmid ends)
ASNG637 contig00061, contig00604
ASNG639 contig00533 (? -- contig only shows homology to one of the two fosmid ends, and bitscore at that is only 274)
ASNG667
ASNG682 contig00369
ASNG735 contig00057
ASNG740 contig00404
HF200_15_66 contig00316 (contig only shows homology to one of the two fosmid ends)
HF200_15_82 contig00624
HF200_15_85 contig00614
HF200_15_87 contig00133
HF200_20_49 contig00558 (contig only shows homology to one of the two fosmid ends)
HF200_25_12 contig00338, contig00134
HF200_25_48 contig00440
HF200_25_64 contig00135
HF200_25_67 contig00316 (contig only shows homology to one of the two fosmid ends)
HF200_35_43 contig00625
HF200_35_46 contig00527
HF200_35_50 contig00373 (1 kb in on contig), contig00039
HF200_35_54 contig00619
HF200_35_70 contig00390 (contig only shows homology to one of the two fosmid ends)
HF200_40_29 contig00003, contig00602 (1kb in on contig)
HF200_40_34 contig00533
HF200_40_80
HF200_40_88
HF200_45_03 contig00376
HF200_45_30 contig00297
HF200_45_47 contig00524, contig00390
HF200_45_67 contig00028
HF200_55_04 contig00075

Note the large, statistically significant, non-random empty section from ASNG3084 to ASNG3403. These colonies all come from, and are the only colonies picked, from plates 90, 91, 92, 93.

List 2.1 Fosmid/contig names of each sequence belonging to each marine group II cluster.

Cluster 1:

contig00017
contig00072
contig00151
contig00370
contig00378
contig00399
contig00399 Fragment 2
contig00405
contig00415
contig00478
contig00526
contig00598
4003641_fasta.screen.Contig2
4003641_fasta.screen.Contig2 Fragment 2
4003657_fasta.screen.Contig1
4003663_fasta.screen.Contig4
4003702_fasta.screen.Contig3
4003712_fasta.screen.Contig1
N_141 Uncultured marine group II euryarchaeote HF70_39H11

Cluster 2:

contig00003
contig00064
contig00105
contig00341
contig00360
contig00397
contig00402
contig00418
contig00424
contig00443
contig00443 Fragment 2
contig00465
contig00507
contig00531
contig00591
contig00602
contig00627
contig00630
4003647_fasta.screen.Contig1

4003647_fasta.screen.Contig1 Fragment 2
4003653_fasta.screen.Contig1

Cluster 3:

contig00376
contig00440
contig00614
contig00619
contig00619 Fragment 2
4003652_fasta.screen.Contig1
4003697_fasta.screen.Contig1
4003703_fasta.screen.Contig2
4003706_fasta.screen.Contig4
4003706_fasta.screen.Contig5
4003714_fasta.screen.Contig13
4003715_fasta.screen.Contig6
4003715_fasta.screen.Contig6 Fragment 2

Cluster 4:

contig00028
contig00039
contig00071
contig00152
contig00291
contig00297
contig00392
contig00444
contig00528
contig00528 Fragment 2
contig00554
contig00615
contig00615 Fragment 2
N_84 Uncultured marine group II euryarchaeote AD1000-18-D2
N_87 Uncultured marine microorganism HF4000_APKG2H5
N_89 Uncultured marine microorganism HF4000_APKG8D23
N_134 Uncultured marine group II euryarchaeote DeepAnt-15E7
N_139 Uncultured marine microorganism HF4000_ANIW141L21
N_144 Uncultured marine group II euryarchaeote KM3-130-D10
N_145 Uncultured marine group II euryarchaeote KM3-130-D10 Fragment 2
N_146 Uncultured marine group II euryarchaeote KM3-72-G3
N_148 Uncultured marine group II euryarchaeote SAT1000-15-B12

Cluster 5:

contig00038
contig00057
contig00059
contig00074
contig00131
contig00133
contig00133 Fragment 2
contig00135
contig00164
contig00338
contig00369
contig00369 Fragment 2

contig00374
contig00374 Fragment 2
contig00390
contig00404
contig00404 Fragment 2
contig00437
contig00451
contig00480
contig00527
contig00530
contig00532
contig00533
contig00533 Fragment 2
contig00534
contig00597
contig00618
contig00621
contig00624
4003645_fasta.screen.Contig4
4003648_fasta.screen.Contig2
4003650_fasta.screen.Contig1
4003650_fasta.screen.Contig1 Fragment 2
4003651_fasta.screen.Contig2
4003654_fasta.screen.Contig1
4003654_fasta.screen.Contig1 Fragment 2
4003655_fasta.screen.Contig15
4003656_fasta.screen.Contig3
4003658_fasta.screen.Contig1
4003658_fasta.screen.Contig1 Fragment 2
4003660_fasta.screen.Contig2
4003661_fasta.screen.Contig3
4003665_fasta.screen.Contig3
4003668_fasta.screen.Contig3
4003668_fasta.screen.Contig3 Fragment 2
4003692_fasta.screen.Contig3
4003693_fasta.screen.Contig3
4003694_fasta.screen.Contig4
4003696_fasta.screen.Contig2
4003698_fasta.screen.Contig2
4003700_fasta.screen.Contig3
4003701_fasta.screen.Contig3
4003705_fasta.screen.Contig2
4003707_fasta.screen.Contig14
4003709_fasta.screen.Contig9
4003711_fasta.screen.Contig6

Outliers:

contig00028 Fragment 2
contig00326
contig00373
contig00607
contig00625
N_85 Uncultured marine microorganism HF4000_APKG1C9
N_86 Uncultured marine microorganism HF4000_APKG1C9 Fragment 2
N_88 Uncultured marine microorganism HF4000_APKG7N23

4003695_fasta.screen.Contig5
4003704_fasta.screen.Contig2
N_135 Uncultured marine group II euryarchaeote EBAC37F11
N_136 Uncultured marine bacterium HF10_29C11
N_137 Uncultured marine microorganism HF4000_ANIW137G21
N_138 Uncultured marine microorganism HF4000_ANIW141A21
N_140 Uncultured organism HF70_19B12
N_142 Uncultured marine group II euryarchaeote HF70_59C08
N_143 Uncultured marine group II euryarchaeote KM3-136-D10
N_147 Uncultured marine group II euryarchaeote KM3-85-F5

List 2.2
Comparison of marine group II archaeal clustering via SOM to clustering via GC content

The clusters resolved in Figure 4 are printed again below.
Bold: Low GC content (less than 55 %)
Ordinary typeface: High GC content (over 55 %)

Note: GC content shown is for the full fosmid, not just 20kb fragments, those used for ESOM clustering.

Cluster 1:

contig00017	**52.5%**
contig00072	**52.6%**
contig00151	**48.8%**
contig00370	**47.4%**
contig00378	**49.9%**
contig00399	**48.7%**
contig00405	**51.1%**
contig00415	**50.8%**
contig00478	**51.4%**
contig00526	**50.7%**
contig00598	**49.9%**
4003641_fasta.screen.Contig2	**49.2%**
4003657_fasta.screen.Contig1	**48.2%**
4003663_fasta.screen.Contig4	**50.3%**
4003702_fasta.screen.Contig3	**48.7%**
4003712_fasta.screen.Contig1	**46.2%**
Uncultured marine group II euryarchaeote HF70_39H11	**50.0%**

Cluster 2:

contig00003	**53.2%**
contig00064	**51.8%**
contig00105	**52.3%**
contig00341	**51.7%**
contig00360	**51.1%**
contig00397	**50.7%**
contig00402	**52.4%**
contig00418	**50.3%**
contig00424	**51.6%**

contig00443	**52.2%**
contig00465	**52.7%**
contig00507	**52.3%**
contig00531	**51.7%**
contig00591	**52.3%**
contig00602	**53.2%**
contig00627	**53.4%**
contig00630	**50.1%**
4003647_fasta.screen.Contig1	**50.7%**
4003653_fasta.screen.Contig1	**48.9%**

Cluster 3:

contig00376	**46.0%**
contig00440	**47.4%**
contig00614	**48.3%**
contig00619	**44.9%**
4003652_fasta.screen.Contig1	**48.1%**
4003697_fasta.screen.Contig1	**45.9%**
4003703_fasta.screen.Contig2	**47.2%**
4003706_fasta.screen.Contig4	**47.2%**
4003706_fasta.screen.Contig5	**45.9%**
4003714_fasta.screen.Contig13	**46.5%**
4003715_fasta.screen.Contig6	**43.3%**

The fragments in Cluster 3 are not just in the low-GC cluster, they are all substantially lower than 55% GC in content; the highest GC value is 48.1%, the average, 46.7%.

Cluster 4:

contig00028	55.4%
contig00039	59.6%
contig00071	58.6%
contig00152	59.9%
contig00291	58.6%
contig00297	60.5%
contig00392	60.8%
contig00444	58.7%
contig00528	59.0%
contig00554	60.3%
contig00615	58.3%
Uncultured marine group II euryarchaeote AD1000-18-D2	58.8%
Uncultured marine microorganism HF4000_APKG2H5	56.1%
Uncultured marine microorganism HF4000_APKG8D23	62.1%
Uncultured marine group II euryarchaeote DeepAnt-15E7	56.0%
Uncultured marine microorganism HF4000_ANIW141L21	60.5%
Uncultured marine group II euryarchaeote KM3-130-D10	59.6%
Uncultured marine group II euryarchaeote KM3-72-G3	58.7%
Uncultured marine group II euryarchaeote SAT1000-15-B12	58.8%

Cluster 5:

contig00038	**47.2%**
contig00057	**54.1%**
contig00059	59.1%
contig00074	**51.0%**
contig00131	**51.5%**

contig00133	**48.7%**
contig00135	**48.2%**
contig00164	**48.5%**
contig00338	**52.8%**
contig00369	**52.1%**
contig00374	**50.1%**
contig00390	**51.3%**
contig00404	**53.6%**
contig00437	**52.6%**
contig00451	**51.9%**
contig00480	**51.8%**
contig00527	**54.3%**
contig00530	**51.5%**
contig00532	**53.0%**
contig00533	58.2%
contig00534	**50.0%**
contig00597	56.0%
contig00618	**52.7%**
contig00621	**54.7%**
contig00624	**50.8%**
4003645_fasta.screen.Contig4	**47.8%**
4003648_fasta.screen.Contig2	**53.1%**
4003650_fasta.screen.Contig1	**52.5%**
4003651_fasta.screen.Contig2	**49.4%**
4003654_fasta.screen.Contig1	57.9%
4003655_fasta.screen.Contig15	**49.5%**
4003656_fasta.screen.Contig3	**48.7%**
4003658_fasta.screen.Contig1	**50.2%**
4003660_fasta.screen.Contig2	58.1%
4003661_fasta.screen.Contig3	**49.2%**
4003665_fasta.screen.Contig3	**49.8%**
4003668_fasta.screen.Contig3	**54.2%**
4003692_fasta.screen.Contig3	**52.6%**
4003693_fasta.screen.Contig3	**49.3%**
4003694_fasta.screen.Contig4	56.0%
4003696_fasta.screen.Contig2	**47.7%**
4003698_fasta.screen.Contig2	**49.6%**
4003700_fasta.screen.Contig3	**49.7%**
4003701_fasta.screen.Contig3	**51.4%**
4003705_fasta.screen.Contig2	**51.4%**
4003707_fasta.screen.Contig14	**52.5%**
4003709_fasta.screen.Contig9	**48.1%**
4003711_fasta.screen.Contig6	**51.3%**

Outliers:

contig00028	Fragment 2
contig00326	**54.8%**
contig00373	57.0%
contig00607	**44.7%**
contig00625	**48.2%**
Uncultured marine microorganism HF4000_APKG1C9	**54.8%**
Uncultured marine microorganism HF4000_APKG7N23	**42.0%**
4003695_fasta.screen.Contig5	**45.1%**
4003704_fasta.screen.Contig2	**47.0%**
Uncultured marine group II euryarchaeote EBAC37F11	**44.3%**
Uncultured marine bacterium HF10_29C11	**47.6%**

Uncultured marine microorganism HF4000_ANIW137G21	56.9%
Uncultured marine microorganism HF4000_ANIW141A21	**44.3%**
Uncultured organism HF70_19B12	**48.7%**
Uncultured marine group II euryarchaeote HF70_59C08	**51.4%**
Uncultured marine group II euryarchaeote KM3-136-D10	**45.1%**
Uncultured marine group II euryarchaeote KM3-85-F5	60.0%

List 2.3 HOT179 fosmid end-sequence library matches to marine group II archaeal clusters, listed by matches to individual group II contigs/fosmids

HOT 179 25m:

Cluster 3

contig00376	1

Cluster 5

4003656_fasta.screen.Contig3	2
4003665_fasta.screen.Contig3	2
4003692_fasta.screen.Contig3	3

Outliers:

4003704_fasta.screen.Contig2	1
HF70_19B12	1
contig00625	3

HOT 179 75m:

Cluster 1

contig00526	1

Cluster 2

4003653_fasta.screen.Contig1	1

Cluster 3

4003706_fasta.screen.Contig4	1
4003714_fasta.screen.Contig13	1
contig00376	1
contig00619	2

Cluster 4

KM3-72-G3	1

Cluster 5

4003655_fasta.screen.Contig15	1
4003656_fasta.screen.Contig3	2
4003665_fasta.screen.Contig3	2
4003694_fasta.screen.Contig4	1
4003707_fasta.screen.Contig14	1
4003709_fasta.screen.Contig9	3

Cluster 5 outliers

4003660_fasta.screen.Contig2	3
contig00533	1

Outliers

4003704_fasta.screen.Contig2	1
HF10_29C11	1
HF70_19B12	1
contig00625	2

HOT179 125m:

Cluster 1

HF70_39H11	3

Cluster 2

4003653_fasta.screen.Contig1	3
contig00443	1

Cluster 3

4003697_fasta.screen.Contig1	3
4003703_fasta.screen.Contig2	5
4003706_fasta.screen.Contig4	5
4003706_fasta.screen.Contig5	3
4003714_fasta.screen.Contig13	5
contig00376	6
contig00619	6

Cluster 5

4003648_fasta.screen.Contig2	1
4003655_fasta.screen.Contig15	5
4003656_fasta.screen.Contig3	8
4003665_fasta.screen.Contig3	12
4003692_fasta.screen.Contig3	10
4003694_fasta.screen.Contig4	3
4003698_fasta.screen.Contig2	2
4003707_fasta.screen.Contig14	25
4003709_fasta.screen.Contig9	14
contig00164	1

contig00621	1

Cluster 5 Outliers

4003660_fasta.screen.Contig2	3
4003668_fasta.screen.Contig3	2
contig00533	2
contig00618	1

Outliers

4003695_fasta.screen.Contig5	2
4003704_fasta.screen.Contig2	19
HF70_59C08	5
contig00625	1

HOT179 500m:

Cluster 4

AD1000-18-D2	4
DeepAnt-15E7	1
KM3-130-D10	26
SAT1000-15-B12	8
contig00297	1

List 2.4 HOT186 fosmid end-sequence library matches to marine group II archaeal clusters, listed by matches to individual group II contigs/fosmids

HOT186 25m:
(nothing)

HOT186 75m:
(nothing)

HOT186 110m:
(nothing)

HOT186 500m:

Cluster 2

contig00397	1

Cluster 3

4003714_fasta.screen.Contig13	1
contig00376	1

Cluster 4

AD1000-18-D2	1
KM3-130-D10	10
SAT1000-15-B12	1
contig00039	1

Cluster 5

4003655_fasta.screen.Contig15	1

List 2.5 HF fosmid end-sequence library matches to marine group II archaeal clusters, listed by matches to individual group II contigs/fosmids

HF10:

Cluster 1:

contig00072	1
contig00151	1
contig00526	1

Cluster 2:

contig00105	2
contig00360	1
contig00397	2
contig00443	1
contig00531	1

Cluster 4:

contig00444	1
contig00554	1

Cluster 5:

contig00074	1
contig00369	1

Outliers:

HF70_19B12	2
HF10_29C11	1
contig00625	1

HF 70:

Cluster 1:

HF70_39H11	2
contig00526	1

Cluster 3:

4003652_fasta.screen.Contig1	1

Outliers:

HF70_59C08	3
HF70_19B12	4

HF 130:

Cluster 1:

4003641_fasta.screen.Contig2	2
4003657_fasta.screen.Contig1	1
4003712_fasta.screen.Contig1	2

Cluster 2:

4003647_fasta.screen.Contig1	1

Cluster 3:

4003652_fasta.screen.Contig1	2
4003697_fasta.screen.Contig1	1
4003703_fasta.screen.Contig2	1
4003706_fasta.screen.Contig4	1
4003714_fasta.screen.Contig13	3
contig00376	1

Cluster 5:

4003650_fasta.screen.Contig1	1
4003651_fasta.screen.Contig2	2
4003658_fasta.screen.Contig1	1
4003661_fasta.screen.Contig3	2
4003693_fasta.screen.Contig3	2
4003696_fasta.screen.Contig2	1
4003698_fasta.screen.Contig2	1
4003705_fasta.screen.Contig2	2
4003707_fasta.screen.Contig14	2
contig00374	1

Cluster 5 outliers:

4003654_fasta.screen.Contig1

Outliers:

4003704_fasta.screen.Contig2	2

HF 200:

Cluster 1:

contig00017	1
contig00072	1
contig00415	1

Cluster 2:

4003647_fasta.screen.Contig1	1	
contig00003		1
contig00105	2	
contig00360	2	
contig00402	1	
contig00531	1	
contig00591	1	
contig00627	1	
contig00630	2	

Cluster 3:

contig00376	1

Cluster 4:

contig00028	1
contig00297	2
contig00392	2
contig00615	1

Cluster 5:

contig00057	1
contig00133	1
contig00135	1
contig00369	2
contig00374	1
contig00480	1
contig00527	1
contig00532	1
contig00533	1
contig00618	1
contig00621	2
contig00624	1

Cluster 5 outliers:

contig00059	1

Outliers:

contig00326	2
contig00625	1

HF 500:

Cluster 4

KM3-130-D10	1

HF 770:

Cluster 4

SAT1000-15-B12	1

HF 4000:

Cluster 6:

HF4000_ANIW137P11	1	
HF4000_ANIW141C7	2	

Outliers:

HF4000_ANIW137G21	1	
HF4000_ANIW141A21	1	
HF4000_APKG1C9		2
HF4000_APKG7N23		2

Script 1. **Perl Script which takes sequences as input, and outputs the normalized percent abundance of each (non-redundant) tetranucleotide.**

```
#!/usr/bin/perl

use warnings;
use strict;

# NOTE: This program REQUIRES oligo file: /Users/rachel/oligo_list_length_4
# The file is a newline separated list of all tetranucleotides.

my $usage = "USAGE:
long_seq_file_to_fragments.tetranucleotide_format.pl INPUT_FILE  Fragment_Size_in_kb
note: input file should be a .seq file
";

my $fragment_size_in_kb = $ARGV[1];
my $input_file = $ARGV[0];

die $usage unless (defined $input_file and defined $fragment_size_in_kb);

my $fragment_size = $fragment_size_in_kb * 1000;
my $output_file = $input_file;
if ($output_file =~ /.seq/) {
   substr($output_file, index($output_file, ".seq"), 4, ""); # remove ".seq";
}
if ($fragment_size_in_kb < 1) {
   $output_file .= ".$fragment_size";
   $output_file .= "bp_fragments_tetranucleotide_counts";
}
else {
   $output_file .= ".$fragment_size_in_kb";
   $output_file .= "kb_fragments_tetranucleotide_counts";
```

```
}

my $oligo_file = "/Users/rachel/oligo_list_length_4";
open(OLIGO_FILE, "$oligo_file") || die "Can't open $oligo_file: $!\n";

# initializing default hash
my %initialized_hash;
while (my $line = <OLIGO_FILE>) {
   chomp $line;
   my $complimentary_oligo = $line;
   $complimentary_oligo =~ tr/ACGT/TGCA/;
   my $reverse_comp = reverse $complimentary_oligo;
   if ($line le $reverse_comp) {
      $initialized_hash{$line} = 0;
   }
   else {
$initialized_hash{$reverse_comp} = 0;
   }
}
close OLIGO_FILE;

open (INPUT, "$input_file") || die "Can't open $input_file: $!\n";
open (OUTPUT, ">$output_file") || die "Can't open $output_file: $!\n";

my $sequence;

while (my $header = <INPUT>) {
   chomp $header;
   my $sequence = <INPUT>;
   die "Error: Input format incorrect: each sequence file should contain a header file\n" unless
($header =~ /^>/);

   my $sequence_length = length $sequence;
   my $number_of_loops = $sequence_length / $fragment_size;
   if ($number_of_loops < 1) {
      die "ERROR:
Fragment size is longer than sequence.\n";
   }
   my @kb_fragments = $sequence =~ /[ACGTXN]{$fragment_size}/g;  # Break up sequence
into 1kb fragments
   my $fragment_counter = 0;
   foreach my $fragment (@kb_fragments) {
      $fragment_counter ++;
      my %oligos = %initialized_hash;
      for (1 .. 4) {
         my @partition = $fragment =~ /([ACGTNX]{4})/g;
         foreach my $oligo (@partition) {
            if ($oligo =~ /[NX]/) {
               $oligos{$oligo} ++;
            }
            else {
               my $compliment = $oligo;
               $compliment =~ tr/ACGT/TGCA/;
               my $reverse_compliment = reverse $compliment;
               if ($oligo le $reverse_compliment) {
                  $oligos{$oligo} ++;
               }
               else {
                  $oligos{$reverse_compliment} ++;
               }
            }
         }
          substr($fragment, 0, 1, "");
      }

      #  tetranucleotides have been counted: now need to normalize them:
      my $sum = 0;
      my $frequency;
      foreach my $oligo_count (keys %oligos) {
         $sum += $oligos{$oligo_count};
      }

      print OUTPUT "$header Fragment $fragment_counter\n";  # Header line
      foreach my $oligo_readout (sort keys %oligos) {
         if ($oligo_readout =~ /[NX]/) {}
         else {
            $frequency = $oligos{$oligo_readout} / $sum;
            print OUTPUT "$oligo_readout $frequency\n";
         }
      }
   }
}
close INPUT;
close OUTPUT;
```