Longitudinal Metagenomic Analysis of the Water and Soil from Gulf of Mexico Beaches Affected by the Deep Water Horizon Oil Spill

W. R. Widger[1‡], G. Golovko[2‡], A. Martinez[2‡], E. Ballesteros[2‡], J. Howard[2], Z. Xu[2], U. Pandya[2], V. Fofanov[6], M. Rojas[2], C. Bradburne[3], T. Hadfield[4], N. A. Olson[3], J. L. Santarpia[3,5], Y. Fofanov[1,2]

[1]Department of Biology and Biochemistry and the [2]Department of Computer Science, University of Houston, Houston, TX; [3]Johns Hopkins University Applied Physics Laboratory, Laurel, MD; [4]Midwest Research Institute, Palm Bay, FL; [5]Department of Civil and Environmental Engineering, University of Maryland Baltimore County, Baltimore MD; Eureka Genomics, Inc., Hercules, CA. ‡Equal contribution to the manuscript.

Summary

Estimates of $7x10^5$ cubic meters of crude oil were released into the Gulf of Mexico as a consequence of the April 20th, 2010 Deep Water Horizon drilling rig explosion, leaving thousands of square miles of earth's surface covered in crude oil. Dispersants were used on large slicks and injected at the well head, resulting in oil being suspended throughout the water column[1]. Starting in June 2010, oil reached hundreds of miles of Louisiana, Alabama, Mississippi, and Florida shoreline disturbing the ecological balance and economic stability of the region. While visible damages are evident in the wildlife populations and marine estuaries, the most significant affect may be on the most basic level of the ecosystems: the bacterial and plankton populations.

We present results from high throughput DNA sequencing of close-to-shore water and beach soil samples before and during the appearance of oil in Louisiana and Mississippi. Sixteen samples were taken over a two month period at approximately two week intervals from Grand Isle, LA and Gulfport, MS and were sequenced using the Illumina GAIIx platform. Significant genomic-based population fluctuations were observed in the soil and water samples. These included large spikes in the human pathogen *Vibrio cholera*, a sharp increase in *Rickettsiales* sp., and decrease of *Synechococus* sp. in water samples. Analysis of the contiguous *de-novo* assembled DNAs (contigs) from the samples also suggested the loss of biodiversity in water samples by the time oil appeared at the shores in both locations. Our observations lead us to the conclusion that oil strongly influenced microbial population dynamics, had a striking impact on the phytoplankton and other flora present prior to the appearance of oil, and that the microbial community had not recovered to pre-spill conditions by the end of our observational period.

Introduction

Microbial communities are an essential but vulnerable part of any ecosystem. The basic metabolic activities of microbial communities represent the fundamental status of any environment[2]. Abrupt and severe changes in the microbial metabolism can produce long term effects on the entire ecosystem[3]. Near-shore water and soil microbial communities are likely to experience significant damage from oil contamination as well as human intervention in response to an oil spill.

Until recently, the major obstacles in the analysis of environmental metagenomes were their complexity and the need to isolate and culture individual microorganisms[4]. The first successful estimation of microbial diversity was based on massive sequencing of highly conserved microbial genes such as 16S rRNA[5,6]. Over the last 10 years this method has been successfully applied to the analysis of microbial composition of human and environmental metagenomes[7-15]. Next Generation Sequencing (NGS) technologies such as 454 Sequencing Systems (Roche), SOLiD (Life Technologies), and Illumina's Genome Analyzer allow deep sequencing-based analysis of the entire genomic material (metagenome) present in environmental[8,12] or clinical samples[7,13,16] and have resulted in the discovery of thousands of new genes and metabolic pathways[4,13,16-18].

In order to estimate effects of oil and dispersants (such as COREXIT 9500), on the near-shore water and soil microbial communities, we collected soil (sand) and water samples over a two month period (at approximately two week intervals) in Grand Isle, LA and Gulfport, MS (Supplementary Table 1). A total of sixteen samples were taken just before and after oil began appearing at the sampling locations. Sequencing of the collected

DNA was performed using the Illumina GAIIx Genome Analyzer. Over the course of the study, a total of 431 million 36-base length reads, were produced. After excluding low quality and low complexity reads, 319 million remained for analysis. Raw and filtered reads, quality statistics, and filtration summaries are available on the project's supplementary website (www.bioinfo.uh.edu/Oil_Spill).

Methods Summary

Analysis of sequencing data was based on the combination of three complementary approaches (Figure 1): (*Pipeline **a***) direct search for each of 300+ million reads in GenBank and two specialized databases containing bacterial and viral sequences; (*Pipeline **b***) mapping (alignment) of all the reads to selected microbial genomes in order to estimate their relative abundance across samples; (*Pipeline **c***) and *de-novo* assembly of the most abundant genomic sequences (contigs) using all reads followed by mapping the individual reads from each sample to each contig to estimate their relative abundance.

In *Pipeline **a*** (direct search for each read in GenBank; Figure 1a), the MegaBLAST[19] suite was used to search the non-redundant nucleotide (*nt*) database for each sub-sequence (read) from each sample. Low complexity and repeated region filters were applied in the search, in order to increase its specificity and exclude non-conclusive reads from future consideration. A total of 2,657,868 unique reads were found to have homologous subsequences in 228,123 different genomic sequence entries in GenBank with specific accession numbers. Next, the highest *e*-value score, was used to build a taxonomical association for each selected read and then the Taxonomy Common Tree program from (NCBI)[20,21] was employed to obtain numbers of unique reads associated with each taxonomical group in each sample (available on the supplementary website (*BLAST Hits with Taxonomy.xlsx*)

Pipeline ***b*** (mapping/alignment of all the reads to selected microbial genomes; Figure 1b) was designed to confirm presence and estimate abundance of *known* bacteria species based on the number and locations of the aligned (mapped) sequencing reads across available reference genomes. This approach assumed each bacterium, if actually present in the sample with detectable abundance, would have reads mapped evenly across a large portion of its genome. To make the average nucleotide coverage correctly represent the abundance, all the highly repetitive and/or commonly shared sequences between bacterial species (such as rRNA and tRNA genes) were excluded. To identify candidate bacteria, we merged reads from all 16 samples and performed a MegaBLAST search against the database of all publically available (as of July 2010) bacterial genomic sequences (Eureka Genomics, Inc.). Out of 156,774 different genomic sequence entries in the database, we selected 525 bacterial genomic sequences >100kb for which the total number of unique matching reads were at least 0.05% of the total number of nucleotides in the sequence where the maximum observed value was 6.83% (file *BLAST Hits and Selected Candidate Bacterial Sequences.xlsx* is available on the supplementary data website). Reads from all sixteen samples were mapped to candidate genomes. Since the average coverage across selected genomes was less than 1, highly repetitive and commonly shared bacterial sub-sequences were excluded by assigning zero coverage values to locations in the genome where coverage values were higher than five. To make sure selected genomes regions were covered evenly, each genome was subdivided into non-overlapping windows of constant width such that, on average, 10 reads were expected inside each window;

average coverage was calculated using 80% of windows (10% of the windows with the highest and 10% of the windows with the lowest coverage were excluded as outliers; in each sample the average coverage for each genome was assigned to zero if the total number of windows was less than 10 or the total number of windows with zero coverage was less than 50%. These procedures left us with 188 genomes with non-zero coverage in at least one of sixteen samples and this file (*Candidate Genome Sequence Coverages.xlsx)* is available on the supplementary data website.

Since genomic sequences for the majority of environmental microorganisms are not available, we chose to also employ a *de-novo* assembly approach (*Pipeline* ***c***; Figure 1c) to identify the most abundant genomic sequences in the samples. To maximize the coverage density, we combined reads from all four locations and used the Dwight assembler developed at the University of Houston Center for Biomedical and Environmental Genomics [www.bioinfo.uh.edu]. The assembler generated 34,765 contigs ranging in lengths from 100 to 9,669 nucleotides. The BLAST algorithm was used to identify the possible origin of these assembled sequences and searches were carried out against viral and bacterial databases containing a taxonomically organized collection of 600,000 publically available genomic sequences (Eureka Genomics, Inc.) as well as against the GenBank collection of non redundant nucleotide sequences using BLAST[20]. Reads from each of sixteen samples were mapped separately to each contig and average coverage values were calculated excluding positions with the highest 10% and the lowest 10% coverage. The average coverage for each contig was then normalized by the number of unique reads in each sample (*Contigs BLAST and Coverage.xlsx* is available on the supplementary website).

Results

In the direct GenBank search results (*Pipeline a*), the majority (84-99%) of unique and conclusive reads were identified as bacterial or eukaryotic. The most abundant between these two groups varied between date and location (Figure 2). The proportion of archaea and virus associated reads varied from approximately 1% to 6.5% of the total number of unique, conclusive reads in each sample. Out of all sixteen samples the largest number of conclusive search results mapped to complete genomic sequences of *Marinobacter aquaeolei*, *Synechococcus* sp, *Propionibacterium acnes*, *Candidatus Pelagibacter*, *Vibrio cholera*, *Haliangium ochraceum*, *Synechococcus* phage S-RSM4, *Synechococcus* cyanophage syn9, *Nitrosopumilus maritimus* (archaea), and *Thalassiosira pseudonana* (eukaryote, diatom). A full list of genomes is available on the supplementary website.

Approximately 18% of unique search results pointed to chloroplast, mitochondria, or ribosomal RNA (Figure 3). Interestingly, a high correlation (r^2=0.97) was observed between the number of reads associated with mitochondria and chloroplasts (Figure 1, Supplement). It is important to mention, that the presence of common genes between *Synechococcus sp.* and chloroplast[22,23] such as *psa*A and *psb*A could introduce bias in the estimation of the abundance of both groups of organisms. To eliminate this possibility, we calculated (using *Pipeline **b***) nucleotide by nucleotide coverage density for several chloroplast genomes (data not shown) including *Thalassiosira pseudonana*, *Phaeodactylum tricornutum*, and *Odontella sinensis* and, since no significant coverage bias was observed across these genomes we concluded the data presented in Figure 3 are a reasonable estimation of the concentration of chloroplast genomic material in the sample.

The archaea associated reads were dominated by *Nitrosopumilus maritimus,* a common archaeon living in seawater and responsible for oxidizing ammonium to nitrite; and *Natrialba magadii,* an extremophile adapted to alkaline and hypersaline conditions (pH 9.5 and 3.5 M NaCl) as well as high temperatures and the presence of solvents[24,25]. Before oil (and possibly dispersant) reached the shore on June 14, 2010, *N. maritimus* was dominant in beach soil, while *N. magadii* was the most abundant archaeon in near-shore water. The appearance of oil and a significant decline of *N. maritimus* was observed to occur simultaneously in water at both locations. This was followed by a large increase two weeks later (Figure 3). During the following two weeks, the relative amount of *N. maritimus* increased in Grand Isle, which was more affected by oil and recovery activities than Gulfport, which declined to near pre-disaster concentrations. The relative abundance of *N. magadii* in water was found to decline immediately after oil reached the shore followed by sharp decreases two and four weeks later in the Gulfport location. In summary, four weeks after the disaster, the proportion of reads associated with these two organisms was still significantly elevated in water (29 fold in Grand Isle and 2.5 fold in Gulfport). The ratio between reads associated with these two species changed 100+ fold in the heavily affected Grand Isle water, from an initial almost equal ratio (46/71 *N. maritimus to N. magadii)* before the disaster to 2,801/41 by July 29, 2010). These observations suggest recovery of the archaea population is not complete, particularly in the water at Grand Isle.

Mammals (primarily human and rodents, such as *Mus musculus*) were one of the most abundant classes among identified Eukaryotes (Figure 2, Supplement). In all water samples, the proportions of these sequences decreased after oil arrived on the shores. In soil samples, a similar decrease was observed in the heavily affected Grand Isle samples, which may be explained by the extensive cleaning activities in the area. These activities may have both disturbed animal habitats and restricted public access to the area. In two groups of data: non-mammal eukaryotes and total samples when mammal-associated reads were excluded, we observed *dinoflagellates* associated reads (mostly *Heterocapsa triquetra* and *Heterocapsa rotunda*) were present in a significantly higher fraction in water but not soil samples. The proportion of these reads increased continuously over the four weeks of observations (Figure 4). A similar pattern was observed for the *Plasmodium* sp, which was dominated by reads associated with the human pathogens *P. knowlesi* and *P. falciparum*. The proportion of reads associated with green plants (dominated by *Micromonas* sp.) dropped in the water when oil reached the shore and did not fully recover during the observation time. In contrast, the proportion of diatom-associated reads (primarily *Thalssiosira pseudonana* and *Cylindrothecsa fusiformis*) dropped sharply in all samples after oil arrived. In the Grand Isle area, the proportion of diatoms and other algae grew continuously during the observation period, primarily due to the increase in the abundance of *T. pseudonan.* During this same period, the proportion of *C. fusiformis* decreased, such that the overall balance between these two diatomes species changed in favor of *T. pseudonana.* The proportion of fungi associated reads (including *Candida dubliniensis* and *Lodderomyces elongisporus*, and *Penicillum chysogenum*) increased at both locations when oil arrived (primary by increased *P. chysogenum*) and did not fully recover to pre-spill conditions during the observation time, especially in the soil samples.

Two complementary approaches: direct GenBank search (*Pipeline* ***a***) and the average coverage of selected genomes (*Pipeline* ***b***) were employed to estimate the abundance of bacteria species. *Pipeline* ***b*** was used to select the most abundant bacteria where coverage was consistently distributed across the genomic sequences. Figure 5 shows an example of average nucleotide coverage across constant size windows for *Synechococus* RCC307 and *Vibrio cholera* MJ-1236 chromosome 1. The high coverage regions correspond to common bacterial sequences and repeatable genomic sequences including rRNA operons. The average coverage density expressed significant correlation with the total numbers of unique reads pointing to the same taxa as the GenBank search (Figure 3, Supplement), indicating the results from both approaches are in agreement.

Both direct GenBank search and the average coverage of selected genomes approaches indicate that the concentration of oxygen producing bacteria i.e. *Synechococus* sp. (dominating *Cyanobactera* group), decreases in water when oil reaches the sampling areas and by the end of the sampling period it nearly recovered to its original proportion (although still less at Gulfport; Figure 3). The *Proprionibacteium* species (including human associated *P. acnes*) concentration dropped in the water 100+ fold and did not recover in the water, presumably due to limited access to the area for recreational purposes. In the water samples, the *Rickettsiales* order was dominated by the *Candidatus Pelagibacter ubique* HTCC1062 genome. After a slight decrease in abundance after oil first arrived at shore, the relative concentration of these water associated bacteria began increasing sharply (up to 10 fold) and did not return to their original proportion during the observation period. The most striking observation, however, was a spike in the

Vibrio cholera associated reads, especially in Grand Isle soil and water (500+ fold), which was significantly more affected by oil than Gulf Port.

The majority of reads pointing to DNA Viruses, identified by direct search (*Pipeline a*), mapped to *Synechococcus* phage S-RSM4, *Synechococcus* cyanophage syn9, and *Prochlorococcus* phage P-SSM2. The absence of significant bias in the average coverage across these genomes (*Pipeline **b***) indicates the presence of these genomes was estimated with reasonable accuracy. Over the observation period, the ratio between the proportions of *Synechococcus* cyanophage syn9 and *Prochlorococcus* phage P-SSM2 was significantly changed in the more heavily contaminated Grand Isle water (Figure 3).

Genomic sequences for the majority of environmental microorganisms are unavailable, hence only a small fraction of reads were found to return conclusive results in the direct GenBank search or mapped to selected genomes. De-novo assembly (*Pipeline c*) was designed to identify the most abundant subsequences in the samples. Because of computer hardware limitations in the assembly process, we merged reads from each location and performed the assembly process for each of the locations separately. Reads from each sample were mapped to all contigs to estimate changes in coverage between samples. In individual samples, the coverage for many of the assembled contigs was high (over 1000x). Not surprisingly, many of the assembled contigs had no significant match to the viral and bacterial databases used, or in GenBank.

Interestingly, the total number of contigs present (i.e. those with non-zero coverage) appear to be significantly different across water samples, even when the origin of many of the assembled contigs is unknown. The appearance of oil correlated with a general decrease in the coverage and total number of contigs present in the samples, while only a few specific contigs increased in coverage (Figure 6). These observations can be explained by a loss of biodiversity across the most abundant organism groups in the samples.

Pair-wise comparison of the contig coverage densities from the water samples (Figure 4, Supplement) show differences in the compositions of the most abundant genomic sequences among the water samples, including large changes at both locations corresponding to the appearance of oil. Data from samples collected on the last day of observations (eight weeks after oil reached shores) showed more similarity to samples taken before the appearance of oil than to other samples taken in the presence of oil. This suggests microbial communities at both locations were "responding" to the appearance of oil in a similar manner.

Discussion and Conclusions

The debate over how much oil remains in the gulf water column, on the sea bed, and in sand at the beaches continues. Federal scientists have reported seventy four percent of the oil may have been removed by skimming and bio-remediation efforts but oil layers close to the bottom and micro dispersed oil droplets skew these calculations[26]. The large amount of dispersed oil and the resulting changes in microbial populations may still have effects on the ecological balance in the region.

Presented results suggest the microbial populations in soil and water were significantly disturbed by the appearance of oil, and possibly dispersant, and did not return to normalcy during the eight weeks of observation. The emergence of *V. cholera,* coincident with the first appearance of oil on the shores, and changes in the concentration of other potential human pathogens suggest regular monitoring of the bacterial population in areas affected by disasters is desperately needed. The observed loss of the biodiversity and possible loss of oxygenic photosynthetic bacteria such as *Synechococcus* sp. suggest the possibility of oxygen depletion due to the loss of oxygen producing microorganisms as well as oxygen consumption by bacteria capable of metabolizing at least a fraction of the oil. The long term damage to the ecosystem including the basic food chain is uncertain and requires future research.

Reference List

1 "Deepwater Horizon unified command, US scientific teams refine estimates of oil flow From BP's well prior to capping. Gulf of Mexico oil spill response," www.deepwaterhorizonresponse.com_go_doc_2931_840475_.htm.

2 V. Torsvik and L. Ovreas, "Microbial diversity and function in soil: From genes to ecosystems," Current Opinion in Microbiology 5(3), 240-245 (2002).

3 A. J. Symstad, et al., "Long-term and large-scale perspectives on the relationship between biodiversity and ecosystem functioning," BioScience 53(1), 89-98 (2003).

4 J. C. Venter, et al., "Environmental genome shotgun sequencing of the Sargasso Sea," Science 304(5667), 66-74 (2004).

5 W. T. Liu, et al., "Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA," Appl. Environ Microb. 63(11), 4516-4522 (1997).

6 V. Wintzingerode, U. B. Gobel, and E. Stackebrandt, "Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis," FEMS Microbiol. Rev. 21(3), 213-229 (1997).

7 A. Suau, et al., "Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut," Appl. Environ Microb. 65(11), 4799-4807 (1999).

8 A. B. Martin-Cuadrado, 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 J 2(8), 865-886 (2008).

9 M. Kamekura, "Diversity of extremely halophilic bacteria," Extremeophiles 2(3), 289-295 (1998)

10 H. G. Martin, et al., "Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities," Nat Biotech 24(10), 1263-1269 (2006).

11 D. H. Parks and R. G. Beiko, "Identifying biologically relevant differences between metagenomic communities," Bioinformatics 26(6), 715-721 (2010).

12 R. V. Thurber, et al., "Metagenomic analysis of stressed coral holobionts," Environmental Microbiology 11(8), 2148-2163 (2009).

13 P. J. Turnbaugh, et al., "An obesity-associated gut microbiome with increased capacity for energy harvest," Nature 444(7122), 1027-1131 (2006).

14 P. J. Turnbaugh, et al., "A core gut microbiome in obese and lean twins," Nature 457(7228), 480-484 (2009).

15 T. C. Hazen, et al., "Deep-sea oil plume enriches indigenous oil-degrading bacteria," Science 330(6001), 204-208 (2010).

16 D. Willner, et al., "Metagenomic analysis of respiratory tract DNA viral communities in Cystic Fibrosis and non-Cystic Fibrosis individuals," PLos One 4(10), e7370 (2009).

17 B. J. Baker, et al., "Insights into the diversity of Eukaryotes in acid mine drainage biofilm communities," Appl. Environ Microb. 75(7), 2192-2199 (2009).

18 R. Edwards, et al., "Using pyrosequencing to shed light on deep mine microbial ecology," BMC Genomics 7(1), 57 (2006).

19 Y Zheng, C. Xu, and J. Xue, "A simple greedy algorithm for a class of shuttle transportation problems," Optim. Lett. 3(4), 491-497 (2009).

20 E. W Sayers, et al., "Database resources of the National Center for Biotechnology Information," Nucleic Acids Res 37, D5-D15 (2009).

21 D. A. Benson, et al., "GenBank," Nucleic Acids Res 37, D26-D31 (2009).

22 K. Shinozaki, et al., "The complete nucleotide sequence of the tobacco chloroplast genome," Plant Molecular Biology Reporter 4(3), 111-148 (1986).

32 S. Robbens, et al., "The complete chloroplast and mitochondrial DNA sequence of *ostreococcus tauri*: Organelle genomes of the smallest Eukaryote are examples of compaction," Molecular Biology and Evolution 24(4), 956-968 (2007).

24 M. Konneke, et al., "Isolation of an autotrophic ammonia-oxidizing marine archaeon," Nature 437(7058), 543-546 (2005).

25 M. Kamekura, et al., "Diversity of alkaliphilic halobacteria: Proposals for transfer of *Natronobacterium vacuolatum*, *Natronobacterium magadii*, and *Natronobacterium pharaonis* to *Halorubrum, Natrialba, and Natronomonas* Gen. Nov., respectively, as *Halorubrum vacuolatum* Comb. Nov., *Natrialba magadii* Comb. Nov., and *Natronomonas pharaonis* Comb. Nov., respectively," Int. J. Syst. Bacteriol 47, 853-857 (1997).

26 R. Camilli, et al., "Tracking hydrocarbon plume transport and biodegradation at Deepwater Horizon," Science 330(6001), 201-204 (2010).

End Notes

The environmental samples were collected in collaboration with Midwest Research Institute (MRI) and the University of Houston Center for Bio-Medical and Environmental Genomics (CBMEG). This work is partially funded by the Department of Homeland Security Science and Technology Directorate (UH-G101151), Johns Hopkins University Applied Physics Lab National Security Technology, and the University of Houston (CBMEG).

Author Contribution Statement

W. R. Widger, G. Golovko, A. Martinez and E. Ballesteros contributed equally to the writing of the manuscript. W. R. Widger was responsible for analysis and interpretation of bacterial and viral composition using all three pipelines. G. Golovko was responsible for pipeline algorithm development and implementation, resulting data analysis and interpretation. A. Martinez, was responsible for pipeline a algorithm development and implementation, resulting data analysis and interpretation. E. Ballesteros, was responsible for pipeline c algorithm development and implementation, resulting data analysis and interpretation. J. Howard, was responsible for sample collection, resulting data analysis and interpretation. Z. Xu and U. Pandya were responsible for Sample preparation, DNA isolation, resulting data analysis and interpretation. V. Fofanov designed and implemented the viral and bacterial databases and the mapping/alignment algorithms, as wellas resulting data analysis and interpretation. M. Rojas was responsible for data storage, analysis, and interpretation. C. Bradburne aided in the analysis and interpretation of bacterial and viral composition using all there pipelines. T. Hadfield aided in sample collection, resulting data analysis and interpretation. N. A. Olson, J. L. Santarpia and Y. Fofanov guided algorithm design, data analysis and interpretation, and manuscript development.

Figure Captions

Figure 1. Three complimentary approaches to metagenomic analysis.

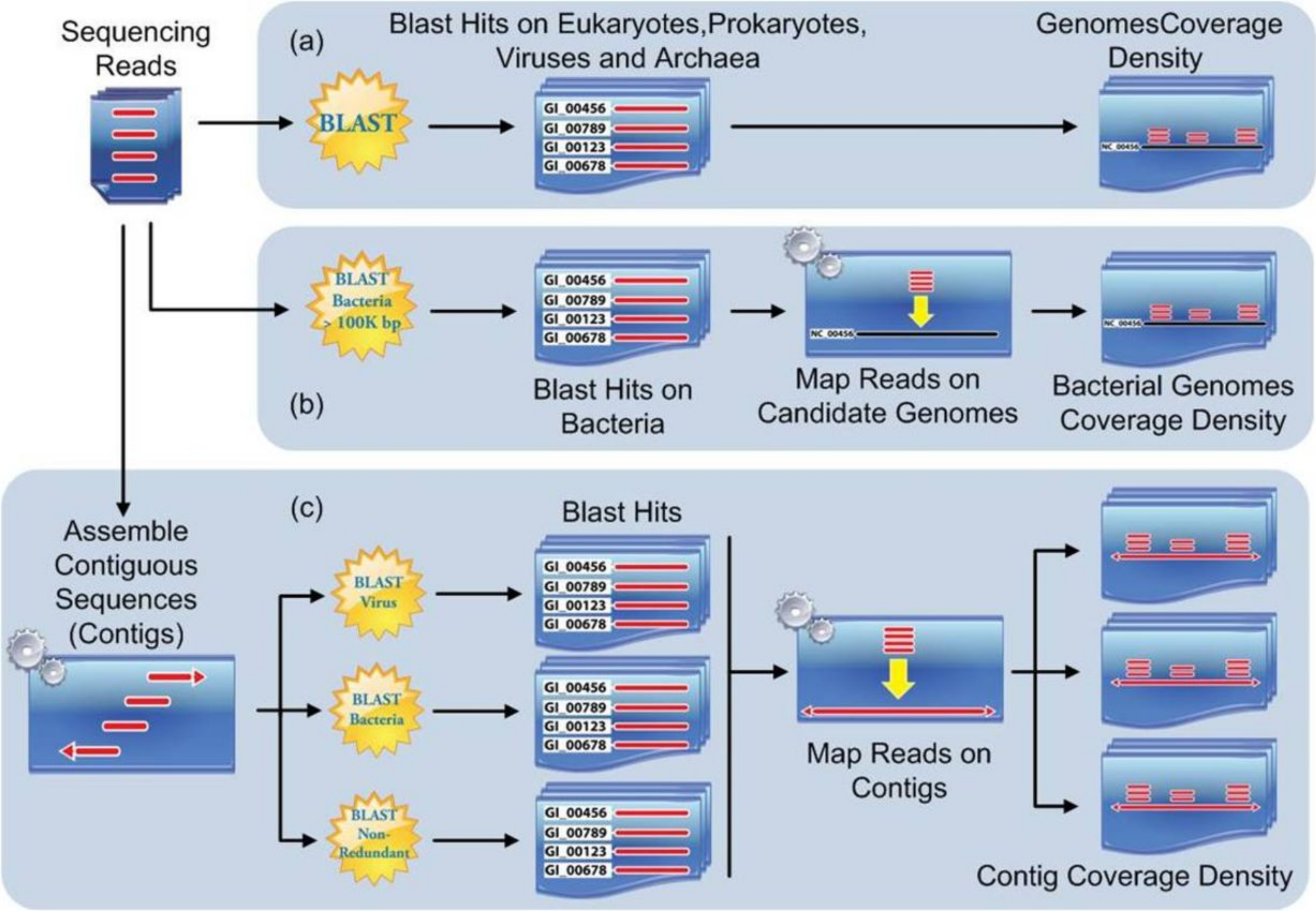
Figure 2. The relative proportion of conclusive reads assigned to major divisions across sixteen soil and water samples.

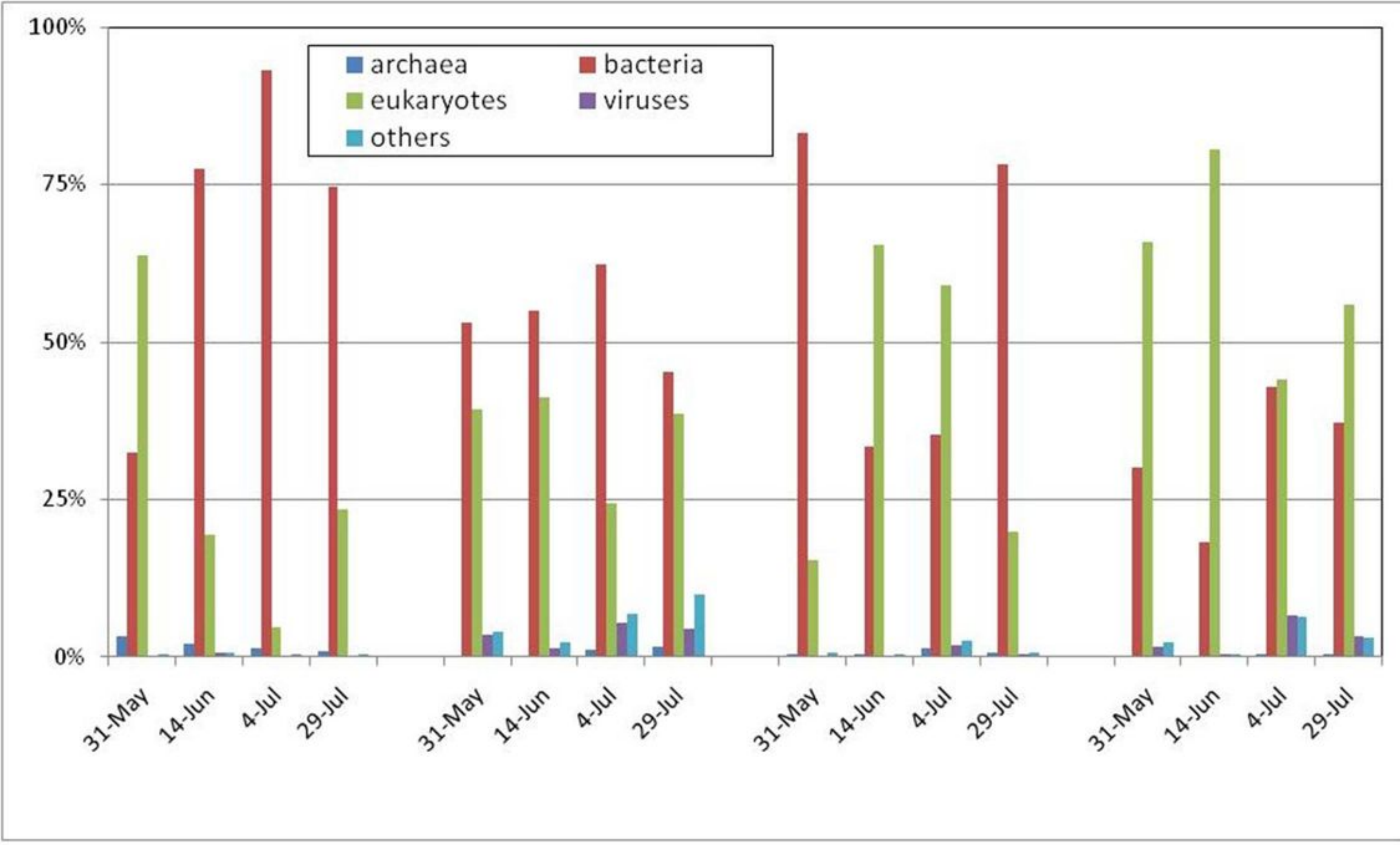
Figure 3. Trends in reads associated with bacteria, archea, virus and phage, and other taxa at the four sample locations.

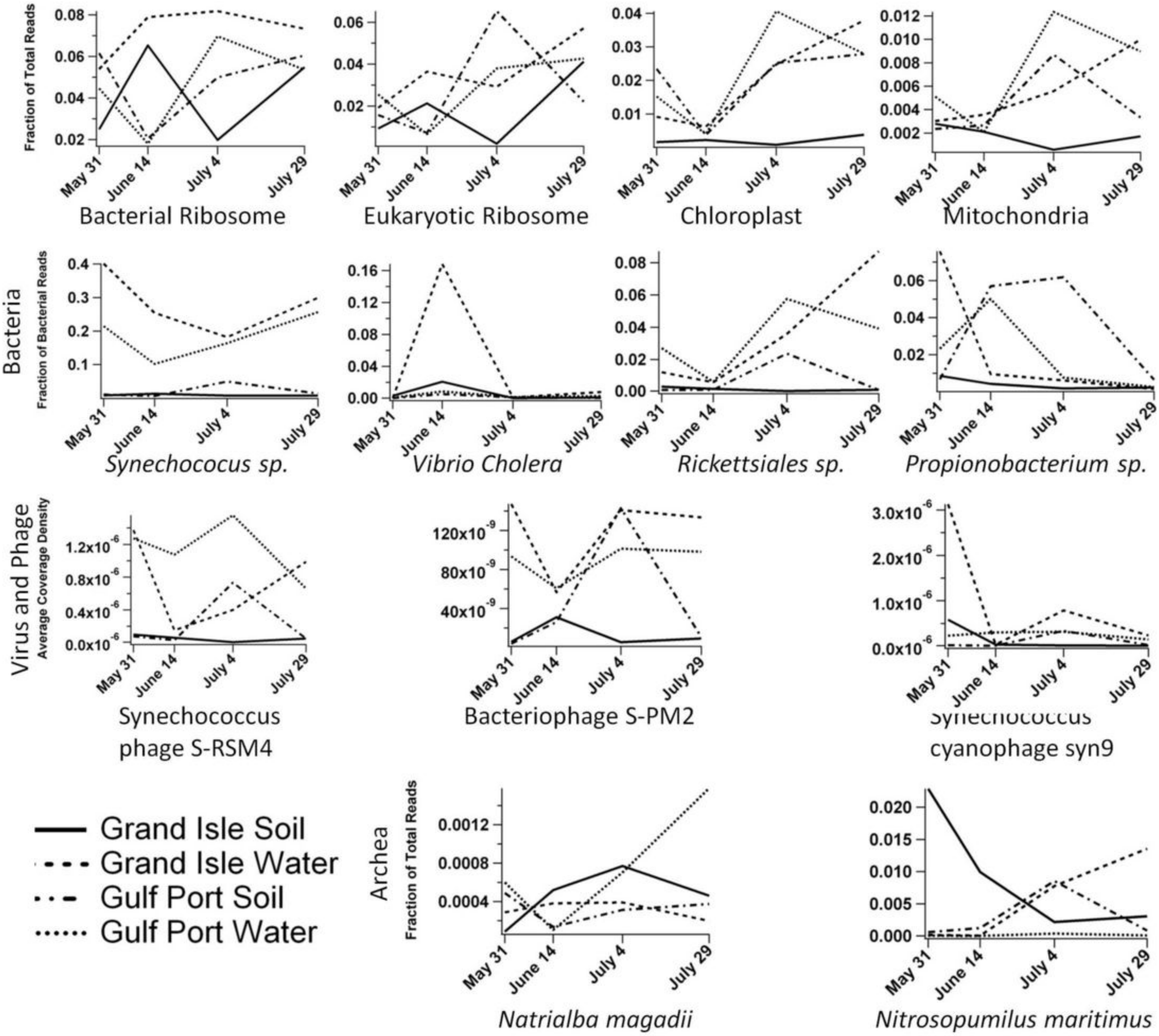
Figure 4. Percent of unique conclusive reads in samples (mammals excluded) associated with dinoflagellates, fungi, algae, and green plants.

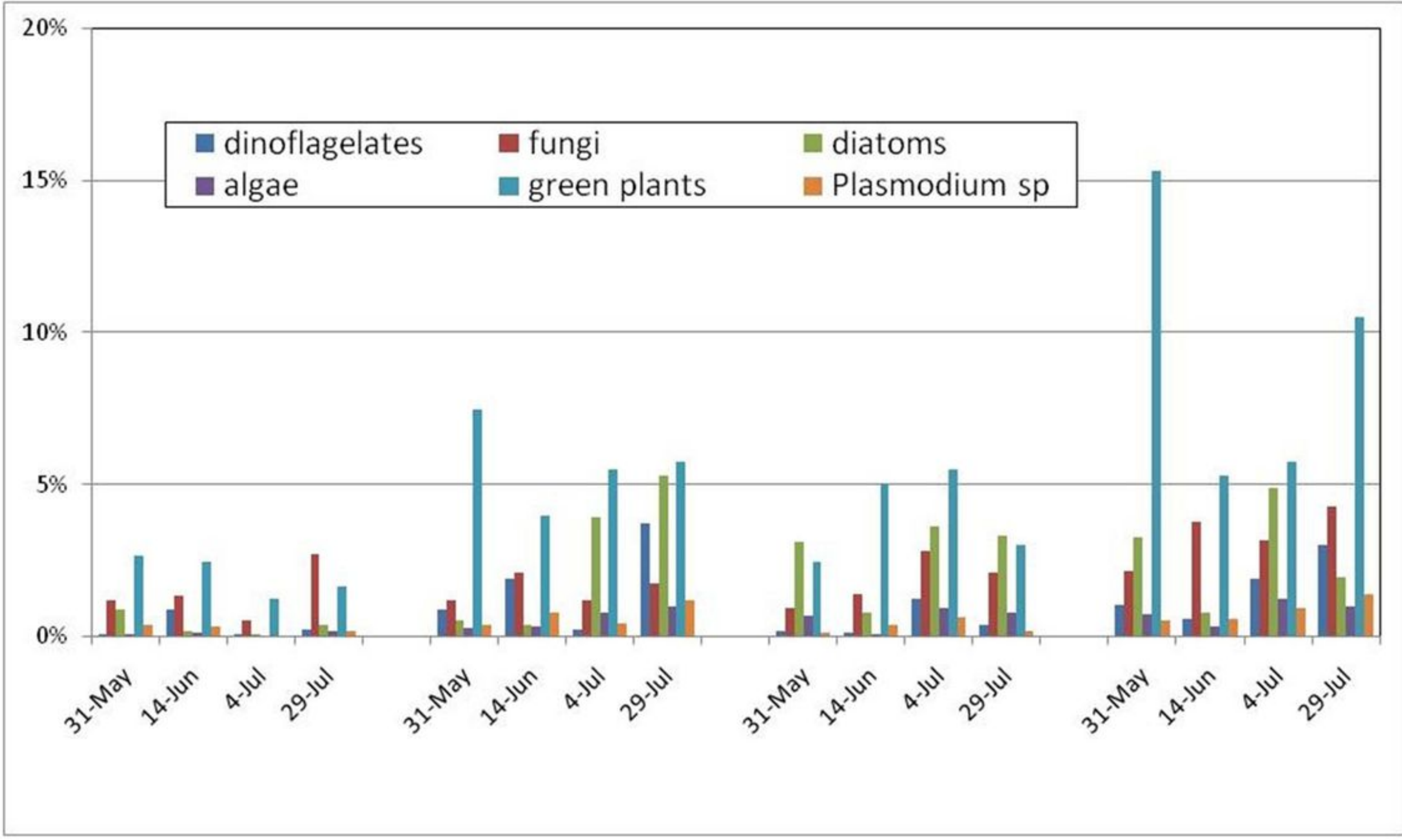
Figure 5. The nucleotide-by-nucleotide coverage density for *Synechococus* (top) and *Vibrio cholera* (bottom) genomes. High coverage regions correspond to *r*RNA for both organisms and *psb*A, *psa*AB for *Synechococcus*. Low or no coverage regions corresponding to heavily mutated or lost regions.

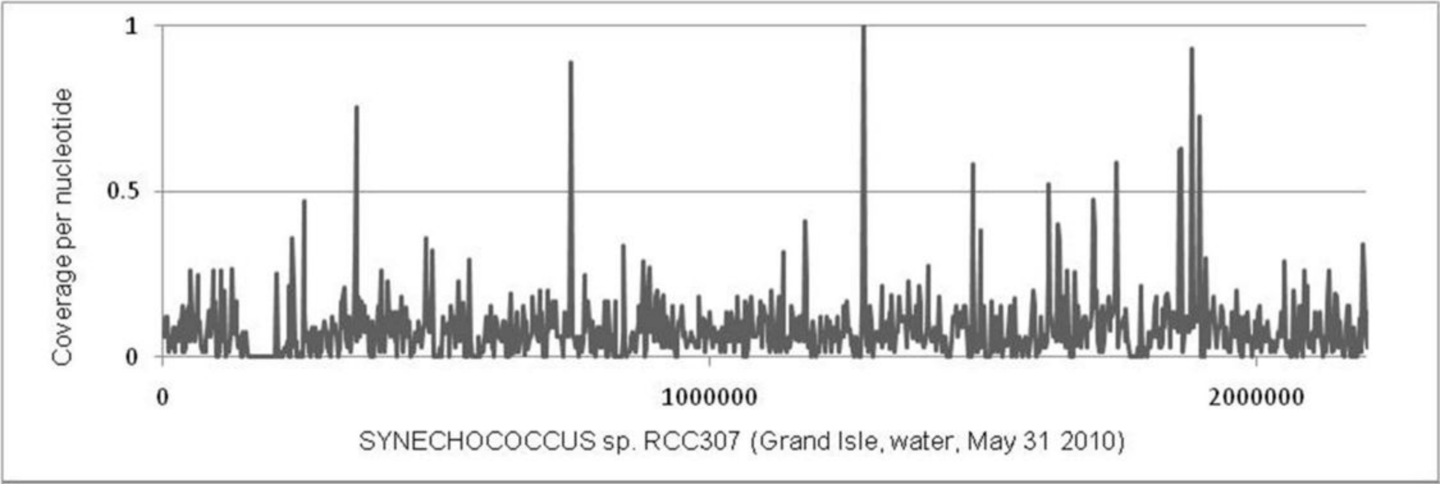
Figure 6. The distribution of contig coverage across water samples.

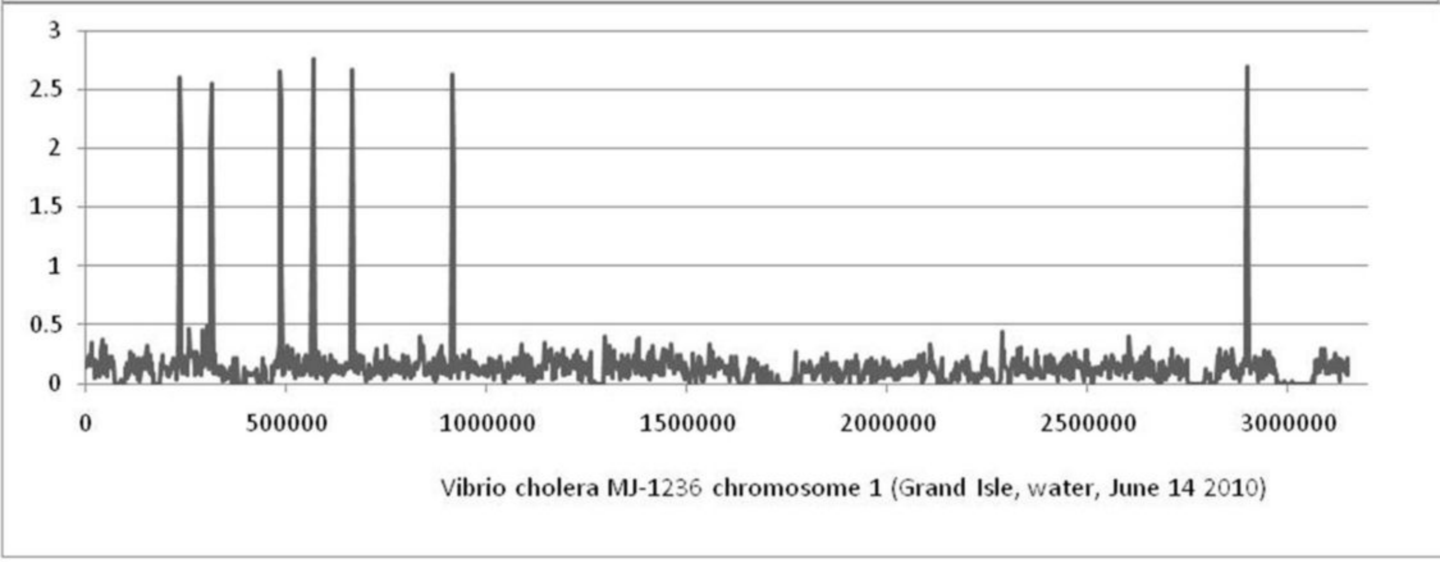

Sequencing Reads
(a)
BLAST
Blast Hits on Eukaryotes,Prokaryotes, Viruses and Archaea
GI_00456
GI_00789
GI_00123
GI_00678
GenomesCoverage Density
NC_00456
BLAST Bacteria > 100K bp
Blast Hits on Bacteria
Map Reads on Candidate Genomes
Bacterial Genomes Coverage Density
(b)
(c)
Assemble Contiguous Sequences (Contigs)
BLAST Virus
BLAST Bacteria
BLAST Non-Redundant
Blast Hits
Map Reads on Contigs
Contig Coverage Density

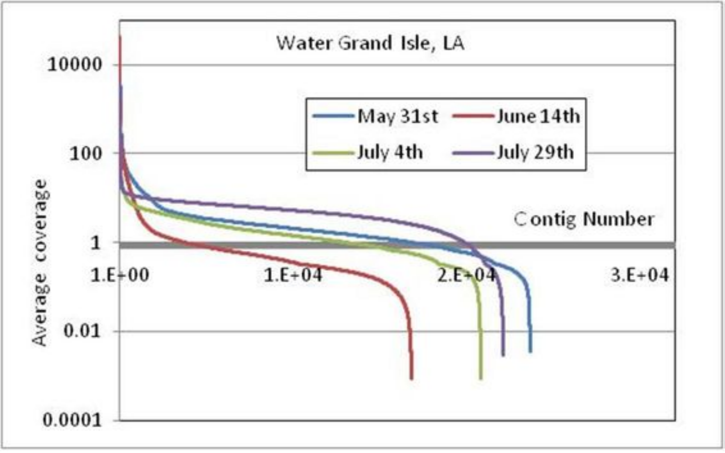

100%
75%
50%
25%
0%
archaea
bacteria
eukaryotes
viruses
others
31-May
14-Jun
4-Jul
29-Jul
31-May
14-Jun
4-Jul
29-Jul
31-May
14-Jun
4-Jul
29-Jul
31-May
14-Jun
4-Jul
29-Jul

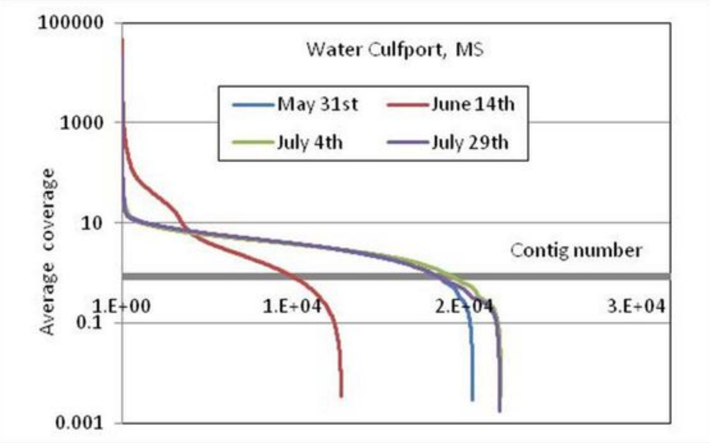

Fraction of Total Reads
0.08
0.06
0.04
0.02
May 31
June 14
July 4
July 29
Bacterial Ribosome
0.06
0.04
0.02
Eukaryotic Ribosome
0.04
0.03
0.02
0.01
Chloroplast
0.012
0.010
0.008
0.006
0.004
0.002
Mitochondria
Bacteria
Fraction of Bacterial Reads
0.4
0.3
0.2
0.1
Synechococus sp.
0.16
0.12
0.08
0.04
0.00
Vibrio Cholera
0.08
0.06
0.04
0.02
0.00
Rickettsiales sp.
0.06
0.04
0.02
Propionobacterium sp.
Virus and Phage
Average Coverage Density
1.2x10^-6
0.8x10^-6
0.4x10^-6
0.0x10^-6
Synechococcus phage S-RSM4
120x10^-9
80x10^-9
40x10^-9
Bacteriophage S-PM2
3.0x10^-6
2.0x10^-6
1.0x10^-6
0.0x10^-6
Synechococcus cyanophage syn9
Grand Isle Soil
Grand Isle Water
Gulf Port Soil
Gulf Port Water
Archea
Fraction of Total Reads
0.0012
0.0008
0.0004
Natrialba magadii
0.020
0.015
0.010
0.005
0.000
Nitrosopumilus maritimus


20%
15%
10%
5%
0%
dinoflagelates
fungi
diatoms
algae
green plants
Plasmodium sp
31-May
14-Jun
4-Jul
29-Jul
31-May
14-Jun
4-Jul
29-Jul
31-May
14-Jun
4-Jul
29-Jul
31-May
14-Jun
4-Jul
29-Jul


Coverage per nucleotide
1
0.5
0
0
1000000
2000000
SYNECHOCOCCUS sp. RCC307 (Grand Isle, water, May 31 2010)


3
2.5
2
1.5
1
0.5
0
0
500000
1000000
1500000
2000000
2500000
3000000
Vibrio cholera MJ-1236 chromosome 1 (Grand Isle, water, June 14 2010)


Water Grand Isle, LA
May 31st
June 14th
July 4th
July 29th
10000
100
1
0.01
0.0001
Average coverage
Contig Number
1.E+00
1.E+04
2.E+04
3.E+04


Water Culfport, MS
May 31st
June 14th
July 4th
July 29th
100000
1000
10
0.1
0.001
Average coverage
Contig number
1.E+00
1.E+04
2.E+04
3.E+04