METAGENOME RECRUITMENT OF THE GLOBAL OCEAN SURVEY DATASET TO FOUR CLOSELY-RELATED SAR11 GENOMES

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

MICROBIOLOGY

MAY 2014

By

Eric S. Brucks

Thesis Committee:

Michael Rappé, Chairperson
Grieg Steward
Sean Callahan

Keywords: SAR11, Pelagibacter, Global Ocean Survey, Fragment Recruitment

Abstract:

The free-living marine bacterial clade known as SAR11 is thought to be one of the most abundant lineages of organisms on our planet. It is also one of the smallest free-living organisms, and possibly the ancestor of the very successful endosymbiont, mitochondria. Members of this group of bacteria possess an incredibly small genome that displays a high degree of synteny and high DNA recombination rate, which represents the upper limit for bacteria. This study examines the global metagenomic read recruitment of the genomes of four closely related SAR11 strains, HIMB4, HIMB5, HIMB83, and HIMB140, in publicly available coastal and open ocean metagenome libraries. Recruitment plots were used to identify and analyze clusters of genes with unusual recruitment patterns, and utilized 2,797,042 open ocean and 2,131,159 coastal high quality reads. It was found that strain HIMB83 possessed the most highly recruited genome and had the highest coastal preference of the four. A hypervariable region previously identified in SAR11, HVR2, was identified in all four strains, and was characterized by a large number of genes coding for cell wall and membrane components as well as several enzymes involved in carbohydrate metabolism. HVR1 was also present in all examined genomes where it was dominated by 12 proteins involved in the type II secretion and type IV pilus pathway. The recruitment of the proteorhodopsin gene present in each of the strains was similar, with a greater degree of conservation on the 3' end of the gene. A phage integrase and extra DNA polymerase was found within the genome of HIMB4, suggesting a possible lysogenic virus present within the genome. This study adds to the current understanding of SAR11 subgroup 1a in a global context, and provides a workflow for generating metagenome recruitment plots for future studies.

List of Tables

List of Figures

TABLE OF CONTENTS

Introduction:

SAR11, also known as the bacterial family Pelagibacteraceae, is the most abundant lineage of bacteria detected in the surface ocean via molecular techniques. SAR11 is an alphaproteobacterial clade within the bacterial order Rickettsiales (Williams *et al.,* 2007). It is possibly the most closely related free-living relative to mitochondria (Thrash *et al.,* 2011), although that phylogenetic placement has recently been called into question (e.g. Rodríguez-Ezpeleta and Embley, 2012). In several different study sites, SAR11 has been shown to compose over half of the microbial community in the euphotic zone of the ocean, and a significant fraction of the microbial community as deep as 3 km (e.g. Morris *et al.*, 2002; Eiler *et al.,* 2008). SAR11 has a relatively slow growth rate that is not atypical of natural seawater bacteria, between 0.19 and 0.71 d^{-1} in controlled laboratory conditions (Rappe *et al.,* 2002; Teira *et al.,* 2009). Both cultivation and cultivation-independent methods have shown that the SAR11 lineage consists of oligotrophic heterotrophs which possess no chlorophyll or bacteriochlorophyll. However, SAR11 has been shown to actively express the gene encoding for proteorhodopsin, which is a protein that utilizes light energy to subsidize normal ATP generation (Béjà *et al.,* 2001; Giovanonni *et al.,* 2005a), among other potential functions. Proteorhodopsin may allow SAR11 to maintain an active metabolism during periods of carbon starvation, which is a common phenomenon in the open ocean (Steindler *et al.,* 2011). Many other types of prokaryotes in the open ocean contain proteorhodopsin, which highlights its importance as an auxillary energy source (e.g. de la Torre *et al.,* 2003; Moran and Miller, 2007). Proteorhodopsin is divided into two categories based on absorbance spectra (green versus blue), and the abundance of each type varies regionally (Béjà *et al.,* 2001; Man *et al.,* 2003; Sabehi *et al.,* 2007). The SAR11 representatives generally contain the

proteorhodopsin family representing their region, with open ocean strains containing proteorhodopsin that absorbs blue light and coastal strains containing green light-absorbing proteorhodopsin (Grote *et al.*, 2012).

SAR11 is characterized by a minimalist evolutionary strategy. For example, the organism itself is tiny (Fig. 1); at just 0.01 μm^3, SAR11 cells are some of the smallest free-living organisms discovered to date (Rappé *et al.*, 2002). Among those characterized so far, the average SAR11 strain contains only ~1400 genes on its ~1.4 Mbp genome, with a high degree of co-localization of related genes (synteny) and low GC content (Giovanonni *et al.*, 2005b; Tamames 2009; Grote *et al.*, 2012). The small cell and genome size and low GC content has also been observed in many parasitic strains (Fraser *et al.*, 1995; Shingenobu *et al.*, 2000; Merhej and Raoult 2011). The benefit of being small and vibrioid is ascribed to a high surface area to volume ratio. Many small genomes have a low GC content presumably because the genomes contain a reduced set of DNA recombination and repair enzymes, which leaves them exposed to the mutational bias towards AT pairs (Van Leuven and McCutcheon 2012). The most abundant phototrophic genus in the ocean, *Prochlorococcus* also contains a small genome, which further argues for the effectiveness of this evolutionary strategy in the open ocean environment (Dufresne *et al.*, 2003, Rocap *et al.*, 2003). Although the property of streamlined DNA is conserved across all members of the lineage, the individual components of the genome vary significantly between SAR11 subclades. For example, HIMB114 and HIMB59 both have a similar genome size and number of genes, but the 742 shared orthologous genes are outnumbered by the 813 genes contained in only one of the strains. The conserved

core genome of all SAR11 strains sequenced so far contains 705 orthologous clusters, while the entire pan-genome discovered to date contains 2,558 orthologous clusters (Grote *et al.*, 2012).

The large amount of variation in the pan-genome might be attributed in part to the ~$2.4x10^{28}$ estimated individual SAR11 cells in the world's oceans (Morris *et al.*, 2002) and the significant phylogenetic diversity between major SAR11 sub-lineages, represented by up to 18% 16S rRNA gene divergence (Thrash *et al.*, 2011; Grote *et al.*, 2012). Within populations of SAR11 cells, this diversity is thought to be maintained at least in part by a high rate of homologous recombination relative to the background mutation rate. This ratio has been calculated to be approximately 60:1 in favor of homologous recombination, which is on the extreme end of the spectrum when compared to other prokaryotes (Vergin *et al.*, 2007; Vos and Didelot, 2009).

Based on 16S rRNA gene sequence analyses, the SAR11 clade can be divided into at least five distinct phylogenetic lineages, or subclades, that can be distinguished by their spatial, temporal, and/or geographic distribution (Fig. 2). Briefly, subclade I (consisting of lineages Ia and Ib) is a surface ocean group of ubiquitous geographic distribution. The subclade Ia lineage includes *Pelagibacter ubique* strain HTCC1062, the first isolated strain of SAR11 (Rappé *et al.*, 2002). Subclade II is the most prevalent lineage found in the mesopelagic open ocean, though it also contains a monophyletic group that is common in coastal water and surface waters of the open ocean. Subclade II is basal to subclades Ia and b. Subclade III consists of a coastal ocean/brackish water group (IIIa),

and the lone freshwater lineage of SAR11 (subclade IIIb). Subclades IV and V are of marine origin, and are basal to the remaining subclades (Grote *et al.*, 2012). Strains have been isolated from subclades Ia, II, IIIa, and V (Grote *et al.*, 2012).

This study is an investigation of the global recruitment of four subgroup 1a genomes using the available data sets from the Global Ocean Sampling (GOS) Expedition divided in to two environmental categories: open ocean and coastal. Recruitment in this case is the matching of metagenomic fragments to a genomic template organized by similarity score. In winter 2003, The *Sorcerer II* GOS Expedition set out to take metagenomic ocean samples from sites stretching from the tip of the Nova Scotia to the equatorial pacific (Rusch *et al.*, 2007). The data from this study is publically available via the Community Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis (CAMERA) website: http://camera.calit2.net/ (Sun *et al.*, 2010). The four closely related genomes selected for this study, HIMB4, HIMB5, HIMB83, and HIMB140, were isolated from Kaneohe Bay off of the coast of Oahu, Hawaii.

Small, conserved genomes often contain hotspots of gene exchange analogous to pathogenicity islands in parasitic bacteria which account for most of the variability between closely related strains (Coleman *et al.*, 2006). Previous studies have found between one and four islands of low recruitment in SAR11 genomes (Wilhelm *et al.*, 2007; Rodriguez-Valera *et al.*, 2009; Grote *et al.*, 2012). These regions have low recruitment presumably because the genes within frequently recombine and mutate, which drastically reduces the number of fragments recruited above the similarity score

cutoff. The most prominent of the low recruitment regions is Hypervariable Region Two (HVR2). This region is located directly adjacent to the 16S rRNA coding region on all sequenced strains except HIMB59, where it is found between two tRNA genes (Grote *et al.*, 2012). This region contains genes that are hypothesized to participate in unique metabolic pathways, which may account for some of the unique metabolic features of individual SAR11 strains. The remaining hypervariable regions identified in HTCC1062 and HTCC1002, named HVR1, HVR3, and HVR4, have not been identified in other strains (Wilhelm *et al.*, 2007; Grote *et al.*, 2012).

There were two objectives for this study. The first objective was to determine if fragment recruitment could be used to discern the habitat preference of the four SAR11 subgroup 1a strains. The second objective was to investigate regions in each genome that recruit fragments at unusual frequencies, similarity scores, or environmental ratios. The content of those regions may help to identify metabolic differences between the strains or explain the relative coastal:open ocean abundance ratios.

Materials and Methods:

Creation of coastal and open ocean custom databases from the Global Ocean Sampling (GOS) database. The GOS Database (Rusch *et al.*, 2007) was accessed through the Community Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis (CAMERA) website (Sun *et al.*, 2010). The raw nucleotide sequence reads from 14 open ocean (GS000a, GS000b, GS000c, GS000d, GS001a, GS001b, GS001c, GS017, GS018, GS022, GS023, GS026, GS037, GS047) and 18 coastal (GS002, GS003, GS004, GS007,

GS008, GS009, GS010, GS013, GS014, GS015, GS019, GS021, GS027, GS028, GS029, GS034, GS035, GS036) GOS samples were downloaded as custom FASTA files (Table 1). The samples were collected, stored, and sequenced as part of the *Sorcerer II* GOS expedition (Rusch *et al.,* 2007). In order to decrease the overall size of each database, custom scripts were used to remove repetitive descriptive elements, such as literature citations.

Core and pan genome analysis. To determine the shared vs. unique gene fraction within the four closely related SAR11 subclade Ia strains isolated from coastal Hawaii seawater (HIMB4, HIMB5, HIMB83, and HIMB140), the phylogenetic profiler within the Integrated Microbial Genomes Expert Review (IMG-ER) system, developed by the Joint Genome Institute (JGI; Walnut Creek, CA, USA), was used (Markowitz *et al.* 2009). The cutoff E-value for identifying a homolog was 10^{-5}, and the minimum percent identity was 30%, which are the default values for identifying shared genes in the IMG program. In addition to identifying homologs shared by all four strains (the core genome), genes shared by two or three strains were also determined using the same parameters, as well as the unique genes for each strain. The unique genes from each strain were selected from the genome and highlighted on the genomic map created in IMG-ER.

Creation of a viewable index file and fragment recruitment plot. The workflow used in this study, outlined in Figure 3, is based on the workflow utilized by CAMERA to generate recruitment plots. The genome sequences of HIMB4, HIMB5, HIMB83 and HIMB140 were retrieved from the IMG-ER database (Markowitz *et al.*, 2009) as FASTA

files. The program FR-HIT was used in order to align fragment hits from the two custom metagenome databases (one “open ocean” and one “coastal”) with each respective genome, running each reference genome against each sample database (Niu *et al.*, 2011) with the following parameters: E-value cutoff of 10 , k-mer size of 11, k-mer overlap size of 8, sequence identity threshold of 75%, minimum read length of 21 bp, and maximum number of failed attempts of 20. Both direct and complementary chains were used. FR-HIT utilizes small nucleotide fragments, so it is important to use a high E-value cutoff, and to rely on sequence identity for eliminating poor potential matches. The program was run using the ‘nohup’ and ‘nice’ commands, which allowed for the utilization of all cores and to check progress and retrieve files while FR-HIT was running at maximum speed. The representative strains of SAR11 used in this study did not all have valid gi numbers except for HIMB5 (GI 406705597). In cases such as these, FR-HIT utilizes the string of numbers occupying the taxon ID region of the FASTA header. FR-HIT files were altered to utilize an *E. coli* genome as an annotation template instead using the following sed command:

```
sed -s
"s/{default_gi_number_for_SAR11}/gi\|110640213\|ref\|NC\_00
8253.1\|/g" {input} > {output}
```

Where the input and output files are both fragment recruitment ‘.fr’ files, and the default_gi_number_for_SAR11 is the string of ten numbers on the third to last column of each fragment recruitment read of the raw input file which should be identical to the taxon ID. This command replaces the default gi numbers with the gi number for *Escherichia coli* strain 536, meaning that the locations of recruited fragments are accurate for each SAR11 strain, but the annotations link to the corresponding loci on the *E. coli*

genome. This process was validated using the unaltered HIMB5 gi number in the same pipeline. A path to the resulting output files was put into an MGAviewer alignment index file in open ocean-coastal pairs, which allows MGAviewer to merge the two distinct .fr files into one viewable index file (Zhu *et al.*, 2012). A separate site for each strain was created using the createsite.py command in the python package and inputting the alignment index files as the argument (eg. $ createsite.py alignmentfile1 outputfolder). This process utilized some supporting components of the separate biopython package. The resulting index file was viewed using Mozilla Firefox. The end result was a fragment recruitment plot with open-ocean and coastal fragment groups separated by color.

Identification of important recruitment regions. Each fragment recruitment plot was manually examined in a 25kbp-wide MGA viewer window moving along the genome in approximately 5kbp increments. Regions that did not fit the normal recruitment patterns of the strain being examined were recorded. Those regions were then matched to their correct annotations in IMG-ER using the scaffold viewing tool. The regions reported in this study were a selected subset of those initially recorded, and that selection was based on both the degree of deviance from the recruitment norm and gene content of the region. Proteorhodopsin was identified in each recruitment plot using the IMG-ER gene search by function.

Determination of fragment recruitment for different similarity scores. The same FR-HIT files used to create the fragment recruitment plots were used to determine the number of fragments recruited in each similarity score range at 5% intervals from 75% to

100%. The following commands were used to count the number of instances fragments recruited at given thresholds:

80% `$ egrep "(8|9|10)[[:digit:]]\.[[:digit:]][[:digit:]]%" -c filename.fr`

85% `$ egrep "(85|86|87|88|89|90|91|92|93|94|95|96|97|98|99||10)\.[[:digit:]][[:digit:]]%" -c filename.fr`

90% `$ egrep "(8|9|10)[[:digit:]]\.[[:digit:]][[:digit:]]%" -c filename.fr`

95% `$ egrep "(95|96|97|98|99||10)\.[[:digit:]][[:digit:]]%" -c filename.fr`

In each of these cases filename.fr is the FR-HIT file containing the hits to be counted. The commands search for all instances where the given percentage or higher is reported and counts them, returning an integer value representing the number of hits. All FR-HIT fragments that satisfied the similarity score were counted, and since the lower threshold for fragment length was 20bp, the egrep command counted only fragments of that size or larger. Because the majority of matched fragments were small, long fragments which take up a large portion of the visual diagram did not affect the recruitment trends as strongly as expected, resulting in apparent discrepancies between the MGAviewer diagram and the numerical recruitment statistics. To insure that the egrep commands were returning

accurate results, a concatenated sample FR-Hit file was created with a known number of recruited fragments in each range. Each command was successfully checked against the sample file.

Results:

Metadata. The metagenome samples used in this study were collected from Canada to near French Polynesia. The latitudinal and longitudinal range of the coastal samples were 32º09'30"N to 10º07'53"S and 64º30'00"W to 91º49'01"W, respectively. The latitudinal and longitudinal range of the open ocean samples were 44º8'14"N to 1º13'42"S and 63º38'40"W to 91º49'01"W respectively (Fig. 4). A total of 2,797,042 high quality sequence reads were included in the custom open ocean database, while 2,131,159 were included in the coastal version.

Core and pan genome analysis. All strains examined in this study are closely-related members of SAR11, the most distantly related sharing 98.5% 16S rRNA gene sequence similarity (Fig. 5, Table 2). The core genome, comprised of homologous genes shared by all four SAR11 subclade 1a strains, consists of 1163 genes or 75.8-79.2% of the genes within each of the strains. The number of unique genes that bear no homology to genes in any of the other three strains ranged from 145 (HIMB140) to 100 (HIMB83) or between 6.8-9.7% of each genome (Table 3, Fig. 6). One large cluster of unique genes occurs adjacent to the rRNA genes in all four strains in HVR2 (Grote *et al.*, 2012), which is described in further detail below. The second region of low recruitment represented in all four strains is most likely HVR1 as described in Wilhelm *et al.* (2007), and occurs 200-

500kbp from HVR2. This region ranges in size from 52kbp (HIMB5) to 75kbp (HIMB4). Despite the lack of recruitment, the contents of this region are not unique genes. The majority of genes represented within this gap are involved in the type II secretory system and type IV pilus construction (Table 4). Although these genes are present in these four closely related strains, they are not all part of the greater SAR11 core genome; for example the related SAR11 strain HTCC9565 has only three genes coding for pilus construction. The remaining unique genes are spread throughout each genome as small clusters or individual genes without any discernible pattern. Coverage of the genome ranged from 0 to >1500x, with the highest coverage represented by regions coding rRNA and tRNA (Fig. 7).

General recruitment trends. All four strains examined in this study were similar in several regards. While more fragments with at least a 75% similarity score were recruited from open ocean samples, the majority of fragments with a higher degree of sequence similarity were coastal samples with the exception of rRNA, tRNA and a few genes with less than 200x coverage (Table 3). Each of the four genomes also exhibited two large regions of very low recruitment. HVR2 ranged in size from 29 KB (HIMB83) to 63 KB (HIMB140) and is adjacent to the ribosomal RNA operon. As mentioned above, the genes within this low recruitment region are almost exclusively unique (Table 5). These genes are most commonly involved in cell wall or membrane biogenesis and carbohydrate transport and metabolism. These categories are represented in every strain by multiple glycosyltransferases and nucleoside-diphosphate epimerases, which are often unique to the strains they are found in. Proteorhodopsin has a unique fragment signature

in all four strains examined in this study. The 3' end of the protein-coding strand is highly conserved in coastal samples, which compose the majority of the recruited proteorhodopsin fragments above 90% similarity.

HIMB4. The genome of HIMB4 recruited fragments with a slight bias towards the coastal database in the lower recruitment ranges, and a strong preference in the 95% and above range (Table 3). Several genes and syntenic gene clusters stood out as distinctive in this organism. There is high relative coastal recruitment for a mitomycin biosynthesis protein. Glycine dehydrogenase and a coenzyme of the glycine cleavage system have higher than normal coastal representation in the 82-86% similarity range, and have coastal fragments recruited above 90% similarity. Bacteria utilize mitomycin for many different purposes ranging in function from antibiotic precursor to transformation induction. Glycine dehydrogenase and the glycine cleavage system are important components in the metabolism of proteins containing glycine.

The HVR2 of HIMB4 is fairly large, and contains several proteins involved in the biosynthesis of sialic acid or a sialic acid derivative from a mannose substrate. That same region also has genes coding for several nucleoside-diphosphate-sugar epimerases. A smaller region of low recruitment in this strain has nine very small hypothetical proteins accompanying two large DNA replication and recombination enzymes, specifically DNA polymerase I with 3'-5' exonuclease/polymerase domains and phage integrase (Fig. 8). Neither of these genes or any other genes in this region are found in any of the other three strains. The HVR1 region present in HIMB4 is the largest of the four strains, containing nearly 50% more genes than the next largest, HIMB83 (Table 4). Like the other three strains, 12 genes in this region are involved in the type II secretion/ type IV pilus

pathway. A cluster of four carbohydrate ABC transporter genes are present in the HIMB4 genome at less than 300x coverage with a strong bias towards coastal samples. Like the other three strains, HIMB4 has a proteorhodopsin protein, and the most similar fragments mainly correspond to the 3' end of the gene.

HIMB5. HIMB5 contains the smallest genome, fewest genes, and lowest GC content of all studied strains. This genome of HIMB5 is the only one examined to recruit open ocean fragments more frequently than coastal fragments, which this strain does from 80-95% similarity. Above 95% similarity, the genome recruits fragments in coastal regions 25% more frequently than it recruits open ocean fragments (Table 3).

The gene coding for proteorhodopsin and surrounding genes (iron-sulfur binding protein, pyridine nucleotide-disulfide oxidoreductase, glutathione S-transferase, multidrug resistance protein, and propionyl-CoA synthetase) recruit fragments from the coastal and open ocean databases in a very similar pattern as HIMB4.

The genome of HIMB5 contains a creatinase/prolidase enzyme and a set of four molybdenum cofactor biosynthesis proteins that have very high recruitment from the coastal database relative to the open ocean database and, while none of these proteins are unique to HIMB5, their recruitment profile is. The genome of HIMB5 also contains a small (4500bp) region of low recruitment encompassing two 1-deoxy-D-xylulose 5-phosphate reductoisomerases, two membrane transport proteins, and a cytidylyltransferase. None of these proteins are unique to HIMB5, but their low recruitment is unusual. There is a phosphate transport system regulatory protein that is present at very low coverage in the coastal database, but is over three times more abundant in the open ocean database peaking at over 500x compared to 150x coastal

coverage where ‘x’ for a given genomic nucleotide is the number of aligned fragments passing over that nucleotide. The genome of HIMB5 contains a phage integrase gene with a recruitment profile comparable with the phage integrase of HIMB4 containing low, mainly coastal recruitment. There is also a Pyrrolo-quinoline quinone cofactor unique to HIMB5 that has only one recruited fragment longer than 100bp.

The HVR1 of HIMB5 contains 46 proteins and is 52 KB in length. Over a quarter of those proteins are components of the type II secretion system, and the majority of the remainder have unknown functions (Table 4). The region of low recruitment which composes HVR2 spans 39kbp, and contains proteins coding for a variety of different functions (Table 5). Four genes involved in N-methylglutamate synthesis recruited fragments mostly from the coastal database at high coverage in the similarity score range of 79-85%. A predicted Zn-dependent hydrolase gene has a normal recruitment profile at a low similarity score, but in the 85-92% range coastal recruitment exceeds open ocean recruitment by a large margin. A cluster of five ABC transporter proteins that span a 5kbp region of the HIMB5 genome have a similar recruitment profile to the four ABC transporters in HIMB4 with low, mainly coastal coverage.

HIMB83. HIMB83 has some characteristics that are very different than the other strains of SAR11 examined in this study. In addition to having by far the fewest unique genes, HIMB83 also boasts the smallest number of ‘missing’ genes: genes which are present in the other three strains, but are not found in this strain (Fig. 6). The dispersed nature of the HIMB83 recruitment plot made it difficult to identify genes of interest, so very few were selected from this genome.

The fragment recruitment plot of HIMB83 shows much higher recruitment in the 90% similarity and above range compared to the other three strains. At the ≥95% similarity score range, HIMB83 recruited over 1% of the coastal fragment database, which is more than 250% of HIMB4, the next most abundant strain (Table 3). The fragment recruitment plots for HIMB4 and HIMB5 both show high recruitment in the low percentile ranges, and decreasing recruitment in the higher similarity range. In these strains the highest recruitment contains mostly gene fragments from the coastal dataset. The plot for HIMB140 is similar, but contains more fragments in the higher percent identity range. HIMB83, however has a bimodal distribution of fragment recruitment. The majority of the recruited fragments are at a low percent identity range, but there are also a large number of fragments that have been recruited with a higher degree of sequence similarity. In between the high and low range, there is a small gap where fewer fragments than expected are recruited. The similarity score for these ranges shift from gene to gene, but the bimodal distribution remains constant. Because the similarity range of the gap changes along the length of the genome, it is difficult to quantify the distribution.

There are four prominent regions of low recruitment in the HIMB83 genome. HVR2 is comprised of 19 genes, making it the smallest among the four strains (Table 5). The main components of this region are nine cell wall and membrane biogenesis proteins and three total transport and metabolism enzymes. HVR1 in contrast is fairly large, containing 40 genes, none of which are unique to the strain (Table 4). The main components of this gene cluster are the 12 proteins involved in the type II secretion/type IV pilus construction pathway. The third region of low recruitment contains 19 genes, 13 of which are unique to HIMB83. Four cell wall or membrane biogenesis proteins and

three carbohydrate transport and metabolism proteins constitute the region. The fourth region of low recruitment contains 14 genes, 12 of them unique to HIMB83. The three large genes that make up the backbone of this recruitment gap are involved in the type I restriction-modification system, and of the remaining 11 genes, ten have no known function.

HIMB140. At 1.44mbp, HIMB140 has the largest genome of the strains examined in this study (Table 3). HIMB4 has one, HIMB83 has six, and HIMB5 has no unique ABC transporter-coding genes. HIMB140 has 21 unique proteins that code for ABC transporter components responsible for the transport of simple sugars, amino acids, phosphate and trivalent iron. The genome of HIMB140 recruited genes at a higher frequency than HIMB4 or HIMB5, but a lower frequency than HIMB83.

The genome of HIMB140 contains a mitomycin antibiotic gene that has a similar recruitment profile as the same gene in HIMB4. A cluster of phosphate ABC transporter genes are present that have a much higher recruitment in open ocean samples than coastal samples, which is atypical at low coverage. None of these genes are unique to HIMB140. A cluster of 5 genes involved in amino acid ABC transport recruited fragments in the 80-85% similarity range and 90-94% similarity range with a small gap in the middle. Both ranges recruited fragments from both the coastal and open ocean databases indiscriminately.

Although the recruitment profile of the proteorhodopsin gene in HIMB140 looks very similar to that of HIMB4 or HIMB5, some of the proteins surrounding it have changed or moved.

There are 4 regions of low recruitment in the HIMB140 genome. The first region spans only 8kb and contains a transcriptional regulator and eight proteins responsible for the transport and metabolism of mannitol and/or sorbitol. Four of these genes are not found in any of the other strains. The second region spans 21kbp, and contains 22 protein-coding genes, 15 have no homologues in the other three strains. Fourteen genes, the majority of those that constitute the region have no known function. Of the remainder, three are involved in amino acid transport and metabolism. The third region is the region encompassing HVR1 which contains 30 genes making it the smallest HVR1 of the four studied genomes (Table 4). None of the genes in that region are unique to HIMB140. The last region of low recruitment is HVR2, which is larger, and contains more genes than the corresponding region in the other three strains (Table 5). The HVR2 of HIMB140 also contains a larger portion of carbohydrate and amino acid transport/metabolism genes, and a larger portion of unique genes than the other strains.

Discussion:

Abundance of each strain. Each strain recruited a higher percentage of coastal fragments than open ocean fragments, suggesting that each of these strains, originally isolated from a coastal area, are likely specialized for coastal environments, and were not simply open ocean strains that drifted in to Kaneohe Bay. This is an important observation due to the relatively short residence time of seawater in the bay. The possible exception is strain HIMB5, which recruited with a slight bias towards the open ocean database in the higher similarity score ranges, but in the highest category recruited mainly from the coastal database. This could mean that representatives of HIMB5 are more

prominent in coastal regions, but there may be some closely-related strains that have a slight preference for the open ocean. The unusually large number of recruited fragments in the >90% identity range of the HIMB83 genome compared to the other three strains indicate that it may be one of the more prominent genotypes of SAR11 subclade 1a that dominate coastal environments globally. HIMB83 also shows a strong preference to coastal waters, so it is possible that its genome is more specialized than the other strains for success in coastal regions. Following HIMB83, HIMB140 is likely to represent the next most prominent group, while HIMB4 and HIMB5 represent groups with a much smaller global presence. Despite the fact that HIMB4 and HIMB140 have nearly identical 16S rRNA, the strains display noticeably different recruitment profiles. Of the two strains, HIMB140 shows a higher relative coastal presence.

Hypervariable Region 2. The region adjacent to the rRNA operon in all four genomes contains a high density of unique genes. This region has been identified in previous studies as a region of low recruitment present in all sequenced SAR11 representatives, and has been found to contain unique genes which may confer strain-specific metabolism (Wilhelm *et al.,* 2007; Grote *et al.,* 2012). Although the genes are unique, the region of each strain contains similar families of genes at a similar frequency with a focus on cell wall/membrane proteins and a smaller focus on carbohydrate transport and metabolism. The consistency of the co-occurrence of these genes suggests cooperation between them. It is possible that these two seemingly separate gene categories are related, since a consistent component of the cell wall and membrane protein group is at least one glycosyltransferase, which is an enzyme that is often utilized in the production of glycoconjugates, including membrane and cell wall components

(Breton *et al.*, 2006). The size of this region is closely correlated to the total number of unique genes present in each strain, but not to the total protein-coding gene count. The implication is that this region undergoes gene transfer at a much higher rate than the other sections of the genome containing non-core genes. The frequent exchange of these genes indicates that they are most likely involved in defense against the numerous SAR11 viruses (Zhao *et al.*, 2013). Rodriguez-Valera *et al.* (2009) suggested as much, citing that many of the proteins coded in HVR2 affect regions that are known to be phage recognition sites, specifically the O chain of LPS.

Hypervariable Region 1 and type II secretion/ type IV pilus genes. The region of low recruitment present in all four strains containing type II secretion system and type IV pilus genes which is most likely HVR1 has a large number of conserved genes and a generally conserved gene arrangement. The fact that these genes are conserved between strains, but are absent from the recruitment plot indicates that there is either a very high mutation and/or recombination rate within this region, or that the genes are very narrowly distributed. Although this region was found in all four SAR11 subgroup 1a strains, it has not been identified in any SAR11 genomes outside of this subgroup (Grote *et al.*, 2012). The main gene category present in this region appears to be Type II secretion or type IV pilus components. Many of the genes involved in type IV pilus formation are homologous to genes involved in the type II secretion system, which is the reason they are combined into a single category in this study. Type IV pili have two known functions: motility and conjugation (Mattick, 2002; Filloux, 2010). Motility offered by type IV pili is generally only useful to bacteria moving across a solid surface or biofilm (Mattick, 2002), a profile that does not match any known member of SAR11. The most likely

reason that these representatives of SAR11 have conserved type IV pili/type II secretion system genes is as a means of bacterial conjugation. Cells of SAR11 representative HIMB114 have been observed to form bridges or connections in electron microscopy images, and it is possible that such bridges are used for horizontal gene transfer and/or recombination (Rappé, unpublished data). Each strain has the same 12 genes relating to this system, and many of the surrounding proteins, including a 211kDa protein, are the same as well.

Proteorhodopsin. Not only is proteorhodopsin present in all sequenced SAR11 isolates, it is also present in the genomes of up to 80% of the bacteria present in the oceanic surface water (de la Torre *et al.*, 2003; Moran and Miller, 2007). A few of the genes surrounding proteorhodopsin were also conserved between strains. The proteorhodopsin gene provides an extra source of ATP in times of starvation, which are frequent in the open ocean (Steindler *et al.* 2011). Proteorhodopsin present in all four examined strains recruits more similar fragments from the 3' end of the gene, which indicates that the 3' end is more conserved; however the active site which contains retinal consists of amino acids along the majority of the protein (Shi *et al.*, 2009; Reckel *et al.*, 2011).

Unique ABC transporters in HIMB140. HIMB140 has genes encoding far more unique ABC transporters than the rest of the strains combined, and these genes make up a large portion of its unique gene content. Each ABC transporter is potentially involved in the uptake of a new energy or nutrient source, and several of the unique ABC transporters in HIMB140 have been specifically matched to sugar and protein importers. One or more other enzymes may be required in order to modify or utilize the moiety imported by the

ABC transporters, and these enzymes may represent a much larger fraction of the unique gene content of HIMB140. Three of the sugars that are potentially taken up by some of the HIMB140 ABC transporters, ribose, xylose, and arabinose, make up a larger percentage of aldoses in the open ocean than in coastal areas (Panagiotopoulos and Sempéré, 2005). HIMB140 appears to have a smaller percentage of its population present in the open ocean than its nearest taxonomic neighbor, HIMB4, which is contrary to what would be expected given the presence of these open ocean genes. HIMB140 has the largest genome of the four studied strains, likely because of its many extra transporters. The small size and small genome is part of what makes SAR11 strains so competitive. With a small genome size, there is less DNA to replicate, repair, and transcribe, all processes that require energy. Therefore, there must be a reason HIMB140 has kept these transporters, and given that there is a stronger coastal preference, it is possible that the other transporters are moving compounds that are more often found in coastal areas, thus shifting the overall metabolic advantage back to the coastal environment.

Possible bacteriophage genes in HIMB4. The lack of recruitment in the region surrounding phage integrase and DNA polymerase I in the genome of HIMB4 is interesting because the other three strains have nothing analogous to it. HIMB4 has another DNA polymerase I gene which is similar to those of the other three strains, and there is only one other strain with a phage integrase gene. Phage integrase plays a major role in the lysogenic cycle of bacteriophages. Phage integrase is responsible for storing the viral DNA in the host genome until conditions are favorable for a virus to enter the lytic phase of infection (Groth and Calos, 2004). Additionally, SAR11 strains in general have minimal genomes consisting around 1,500 genes. It would be uncharacteristic of

HIMB4 to express two copies of such a large gene. The most reasonable explanation is that the region of low recruitment likely formed via a recent lysogenic viral infection that left those two enzymes and the surrounding unknown proteins in the SAR11 genome. The presence of viral genes in the genome of HIMB4 should not be surprising considering the overwhelming concentration of SAR11 viruses present in the world's oceans (Zhao *et al.* 2013). Although the size fraction that was used to collect these samples was small enough to capture SAR11 cells, it was still too large to catch many virus particles, which explains why all 11 genes that constitute this 9000bp segment of the genome recruited only a handful of fragments larger than 50bp in length. These few recruited fragments most likely came from other infected cells.

Suggestions for future research. HIMB140 has several unique ABC transporters that have no specifically-matched substrate, all of which were carefully chosen through the natural selection process on a very conservative genome. As suggested in this study, these transporters likely transport moieties that are particularly useful to HIMB140, and likely account for its strong coastal preference relative to HIMB4. Further examination of these genes is sure to reveal metabolic pathways that are more helpful in coastal environments.

The putatively viral genes present in the HIMB4 genome present a unique opportunity to study the associated virus. Comparing ocean metagenomic samples from a viral size fraction to the low recruitment region containing the viral genes would verify whether or not the genes were truly viral.

Lastly, although there is noticeable variation between coastal and open ocean recruitment, methods used in this study make it difficult to draw conclusions

pertaining to relative abundance. The information used to generate recruitment data was taken from 32 discrete points and times, many of which were close to another, and three were taken from the same location at the same time. Because all these data were blended into two sample groups, local anomalies were incorporated into the larger dataset. Such occurrences could have created perceived global trends that were in reality either spatially or temporally isolated events. If a similar study was undertaken using a larger number of sample sites, local variance would be smoothed out by the number of samples. Such a study would provide data from which global extrapolations could more realistically be drawn. Alternatively, the same study could be done by sample site. Such a study would locate sites with relatively large or small populations of individual strains. This information could be used to determine the global variance in strain abundance, and identify regions for which certain strains are best suited.

References:

Béjà O, Spudich EN, Spudich JL, Leclerc M, DeLong EF (2001) Proteorhodopsin phototrophy in the ocean. Nature 411:786–789.

Breton C, Šnajdrová L, Jeanneau C, Koča J, Imberty A (2006) Structures and mechanisms of glycosyltransferases. Glycobiology. 16 (2): 29R-37R

Coleman ML, Sullivan BS, Martiny AC, Steglich C, Barry K, DeLong EF, Chrisholm SW (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311:1768–1770.

Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann I, Barbe V, *et al.* (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc. Natl. Acad. Sci. U. S. A. 100:10020–10025.

Eiler A, Hayakawa DH, Church MJ, Karl DM, Rappé MS (2009) Dynamics of the SAR11 bacterioplankton lineage in relation to environmental conditions in the oligotrophic North Pacific subtropical gyre. Environmental Microbiology 11:2291-2300.

Filloux, A. (2010) A variety of bacterial pili involved in horizontal gene transfer. Journal of Bacteriology 192(13): 3243-3245.

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403.

Giovannoni SJ, Bibbs L, Cho JC, Stapels MD, Desiderio R, Vergin KL, Rappé MS, Laney S, Wilhelm L, Tripp HJ, Mathur EJ, Barofski DF (2005a) Proteorhodopsin in the ubiquitous marine bacterium SAR11. Nature 438 (7064): 82–85.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur EJ (2005b) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309:1242–1245.

Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, Rappé MS (2012) Streamlining and core genome conservation among highly divergent members of the SAR11 clade. mBio 3(5): e00252.

Groth AC, Calos MP (2004) Phage integrases: biology and applications. Journal of Molecular Biology 335(3): 667-678.

Markowitz VM, Mavromatis K, Ivanova NN, Chen IMA, Chu K, Kyrpides NC (2009) IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics 25:2271-2278.

Man D, Wang W, Sabehi G, Aravind L, Post AF, Massana R, Spudich JL, Béja O (2003) Diversification and spectral tuning in marine proteorhodopsins. EMBO J. 22:1725–1731.

Mattick JS (2002) Type IV pili and twitching motility. Annu. Rev. Microbiol. 56(1): 289–314.

Merhej V, Raoult D (2011) Rickettsial evolution in the light of comparative genomics. Biol. Rev. Camb. Philos. Soc. 86:379–405.

Moran MA, Miller WL (2007) Resourceful heterotrophs make the most of light in the coastal ocean. Nat Rev Microbiol 5: 792–800.

Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, Giovannoni SJ (2002) SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420: 806–810.

Niu B, Zhu Z, Fu L, Wu S, Li W (2011) FR-HIT, a Very Fast Program to Recruit Metagenomic Reads to Homologous Reference Genomes. Bioinformatics 27(12): 1704-1705.

Panagiotopoulos C, Sempéré R (2005) The molecular distribution of combined aldoses in sinking particles in various oceanic conditions. Marine chemistry 95(1): 31-49.

Rappé MS, Grote J, Huggett MJ (unpublished). Comparative genomics within a coastal SAR11 population. Poster session.

Rappé MS, Connon SA, Vergin KL, Giovannoni SJ (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418:630–633.

Reckel S, Gottstein D, Stehle J, Löhr F, Verhoefen MK, Takeda M & Dötsch V (2011) Solution NMR structure of proteorhodopsin. Angewandte Chemie International Edition 50(50): 11942-11946.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424: 1042–1047.

Rodriguez-Valera F, Martin-Cuadrado AB, Beltran Rodriguez-Brito, LPC, Thingstad, TF, Forest Rohwer AM (2009) Explaining microbial population genomics through phage predation. Nat. Rev. Microbiol. 7: 828–836.

Rodríguez-Ezpeleta N, Embley TM (2012). The SAR11 group of alphaproteobacteria is not related to the origin of mitochondria. PLoS ONE 7: e30520.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, *et al.* (2007) The *Sorcerer II* Global Ocean Sampling expedition: Northwest Atlantic through easterntropical Pacific. PLoS Biol 5(3): e77.

Sabehi G, Kirkup BC, Rozenberg M, Stambler N, Polz MF, Béjà O (2007) Adaptation and spectral tuning in divergent marine proteorhodopsins from the eastern Mediterranean and the Sargasso Seas. ISME J. 1: 48–55.

Shi L, Lake EM, Ahmed MA, Brown LS & Ladizhansky V (2009) Solid-state NMR study of proteorhodopsin in the lipid environment: secondary structure and dynamics. Biochimica et Biophysica Acta (BBA)-Biomembranes 1788(12): 2563-2574.

Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera sp*. APS. Nature 407:81–86.

Steindler L, Schwalbach MS, Smith DP, Chan F, Giovannoni SJ (2011) Energy starved *Candidatus Pelagibacter ubique* substitutes light-mediated ATP production for endogenous carbon respiration. PLoS ONE 6: e19725.

Sun S, Chen j, Li W, Altinatas I, Lin A, Peltier S; Stocks K, Allen EE, Ellisman M, Grethe J, Wooley J. (2010) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. Nucleic Acids Research 39(1): D546-D551.

Tamames J (2001) Evolution of gene order conservation in prokaryotes.Genome Biol. 2:research0020-research0020.11.

Teira E, Martinez-Garcia S, Lønborg C, Alvarez-Salgado XA (2009) Growth rates of different phylogenetic bacterioplankton groups in a coastal upwelling system. Environmental Microbiology Report 1: 545–554.

Thrash, JC, Boyd, A, Huggett, MJ, Grote, J, Carini, P, Yoder, RJ, Robertse B, Spatafora JW, Rappé MS & Giovannoni, SJ (2011) Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Scientific reports 1.

de la Torre JR, Christianson L, Béjà O, Suzuki M. T, Karl D, *et al.* (2003) Proteorhodopsin genes are widely distributed among divergent bacterial taxa. Proc. Natl. Acad. Sci. U. S. A. 100: 12830–12835.

Van Leuven JT, McCutcheon JP (2012) An AT mutational bias in the tiny GC-rich endosymbiont genome of Hodgkinia. Genome Biology and Evolution, 4(1): 24-27.

Vergin KL, Tripp HJ, Willhelm LJ, Denver DR, Rappé MS, Giovannoni SJ (2007) High intraspecific recombination rate in a native population of *Candidatus Pelagibacter ubique* (SAR11). Environmental Microbiology 9: 2430 –2440.

Vos M, Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. ISME J. 3:199–208.

Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ (2007) Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. Biol. Direct 2:27.

Williams KP, Sobral BW, Dickerman AW (2007) A robust species tree for the alphaproteobacteria. J Bacteriol 189:4578–4586.

Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC, Giovannoni SJ (2013) Abundant SAR11 viruses in the ocean. Nature 494(7437): 357-360.

Zhu Z, Niu B, Chen J, Wu S, Sun S and Li W (2012) MGAviewer:
A desktop visualization tool for analysis of metagenomics alignment data. Bioinformatics 29(1): 122-123.

FIGURES AND TABLES:

Figure 1. Transmission electron micrograph of Pelagibacteraceae (SAR11) strain HIMB114. Photo credit to Kehau Manoi and Michael Rappé

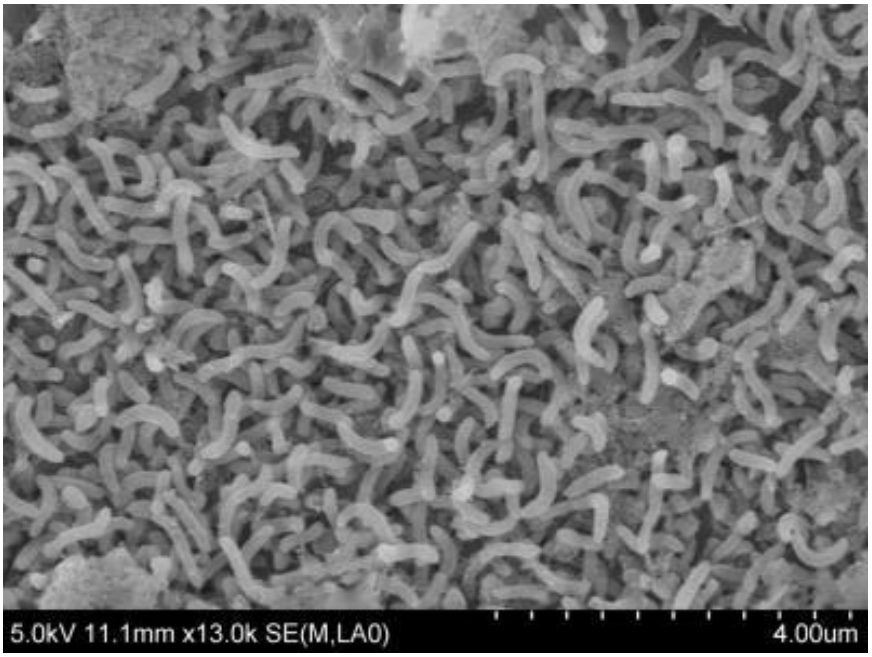


Figure 2. 16S rRNA gene phylogenetic tree of the SAR11 clade displaying the 5 subgroups. A subset of isolated strains with genomes sequenced are shown in red. Bootstrap values are displayed at the nodes (from Grote *et al.* 2012)

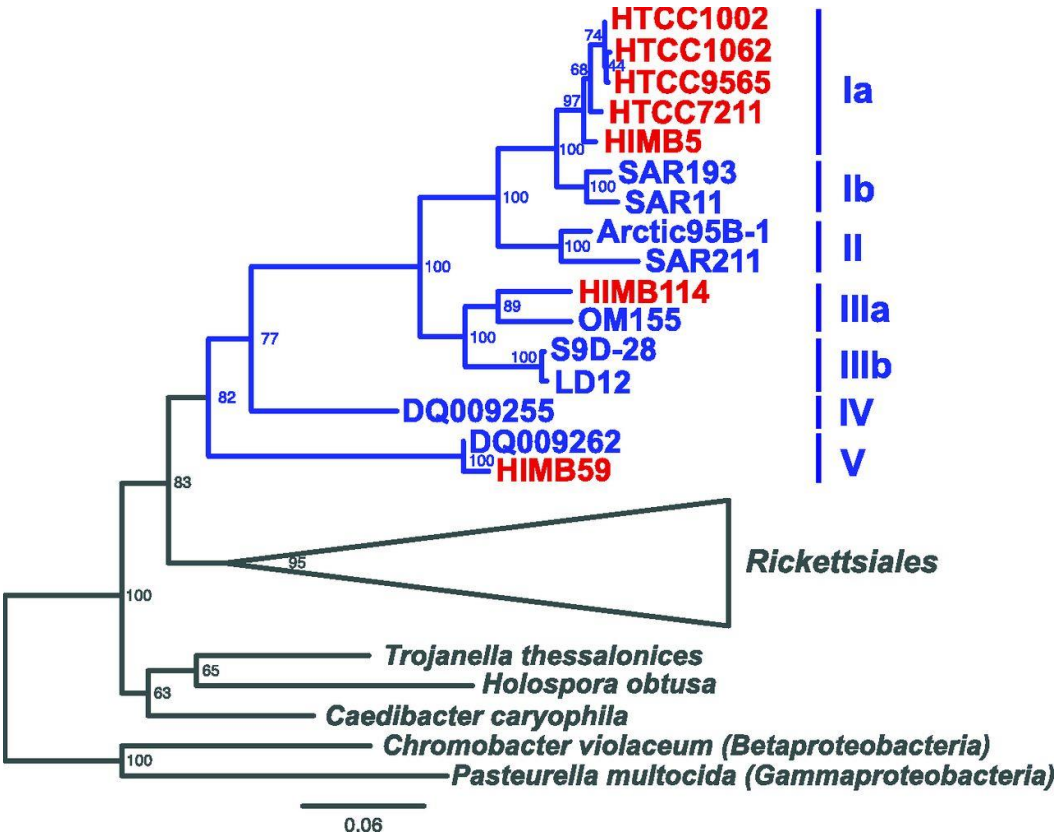


Figure 3. Workflow used in this study to create fragment recruitment plots for the four SAR11 strains using the GOS metagenomic database.

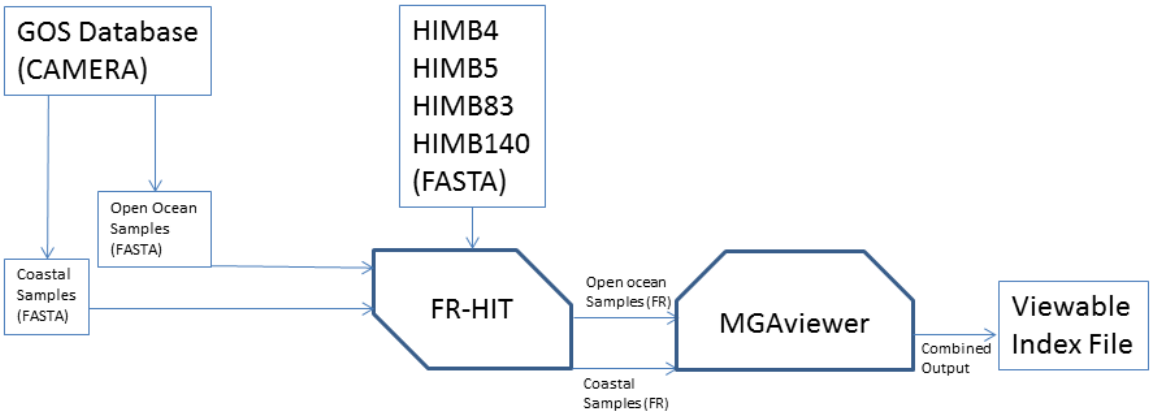


Figure 4. Map of GOS sample sites color coded by type. Coastal samples are shown as red, and open ocean samples are shown as green. Sites represented by a black or grey dot were not used in this study. The inset contains samples 27 through 36 taken from the coastal area surrounding the Galapagos Islands. This image is a modification of Figure 1 from Rusch *et al.* (2007)

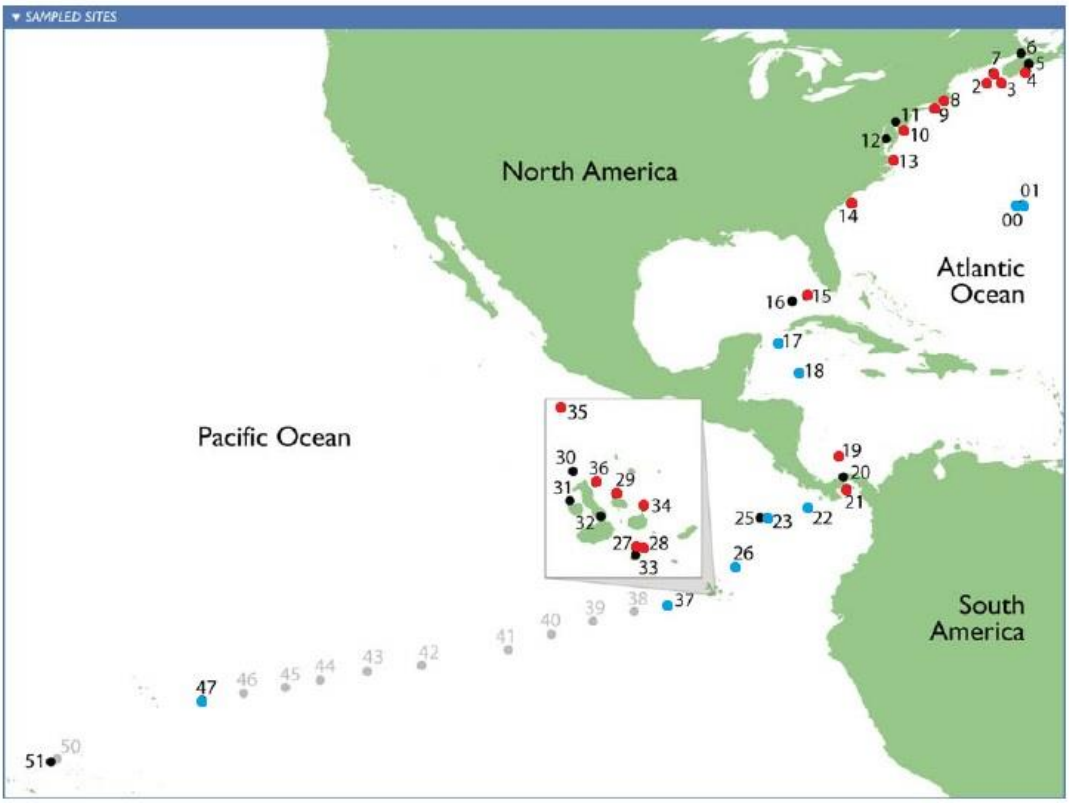


Figure 5. Phylogeny of cultivated SAR11 strains based on 16S rRNA gene sequence comparisons. Strains in red were isolated in Hawaii, and subgroup 1a is highlighted within the box (Rappé *et al.*, unpublished).

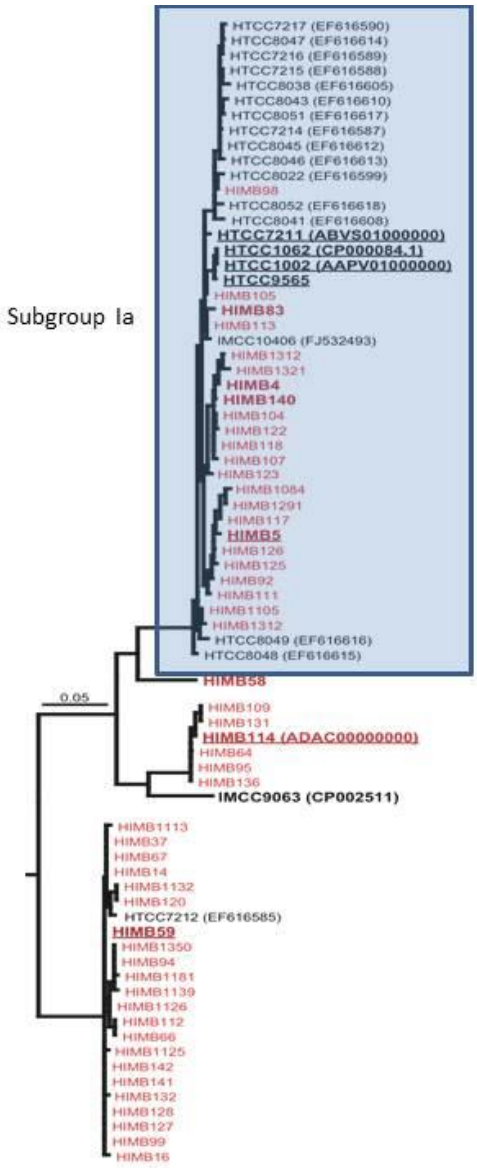


Figure 6. Venn diagram depicting the shared gene fraction and number of unique genes for the four HIMB subclade Ia strains. Each number represents the shared number of homologs specific to the intersecting strains.

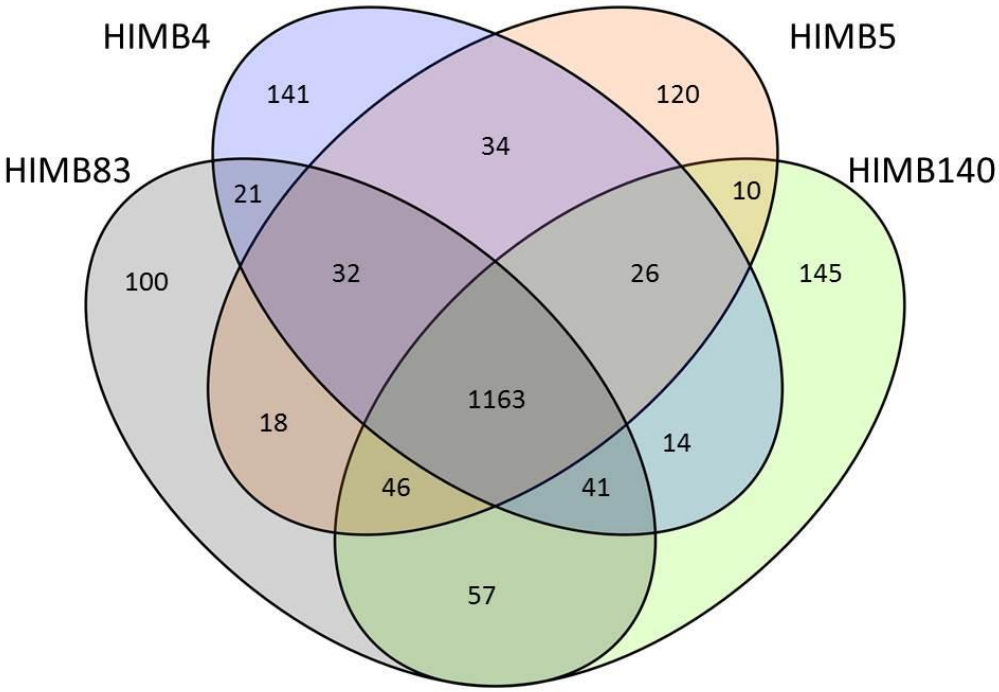
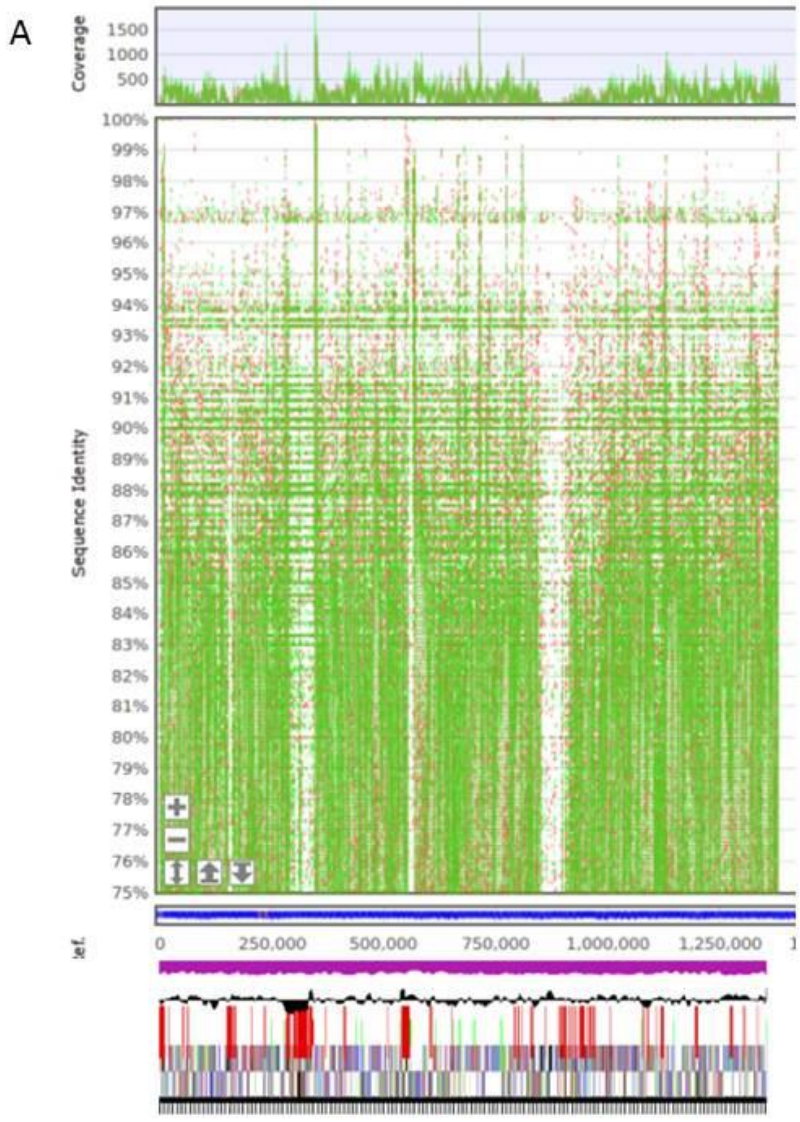
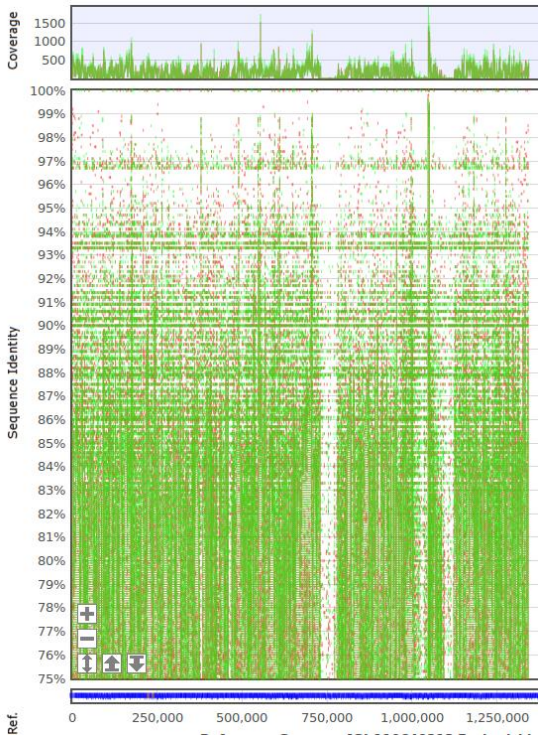
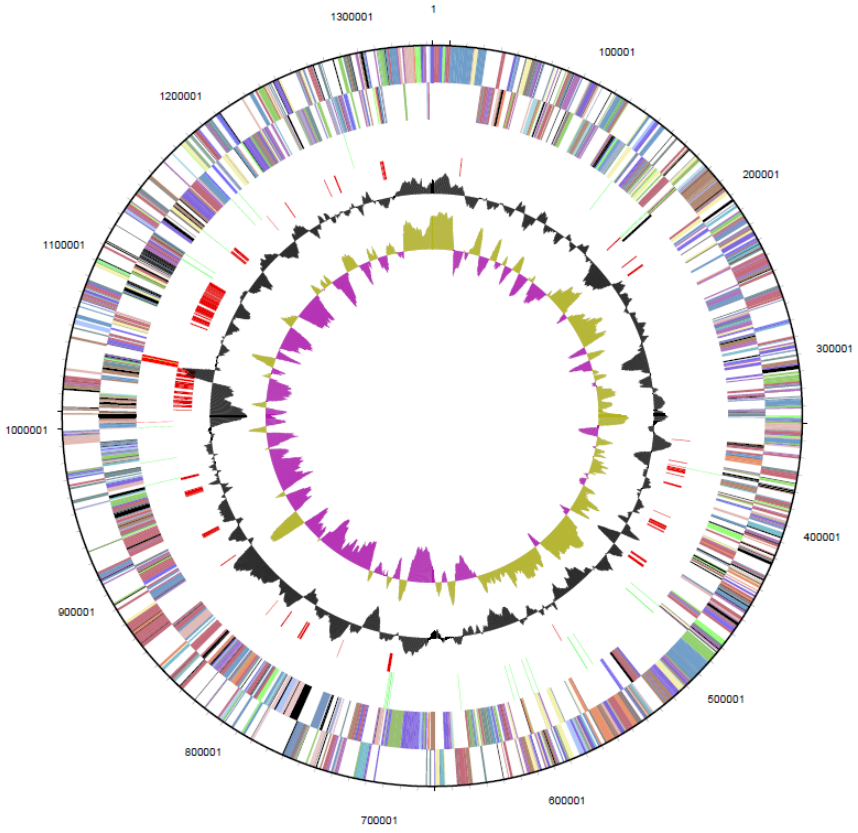


Figure 7. Fragment recruitment plots of SAR11 group 1a strains A. HIMB4 B. HIMB5 C.HIMB83 and D. HIMB140 generated using MGAviewer (Zhu *et al.* 2012). The red bars below the recruitment plot indicate locations of unique genes (continues on next two pages).



B


Coverage
1500
1000
500
Sequence Identity
100%
99%
98%
97%
96%
95%
94%
93%
92%
91%
90%
89%
88%
87%
86%
85%
84%
83%
82%
81%
80%
79%
78%
77%
76%
75%
Ref.
0
250,000
500,000
750,000
1,000,000
1,250,000


1
100001
200001
300001
400001
500001
600001
700001
800001
900001
1000001
1100001
1200001
1300001

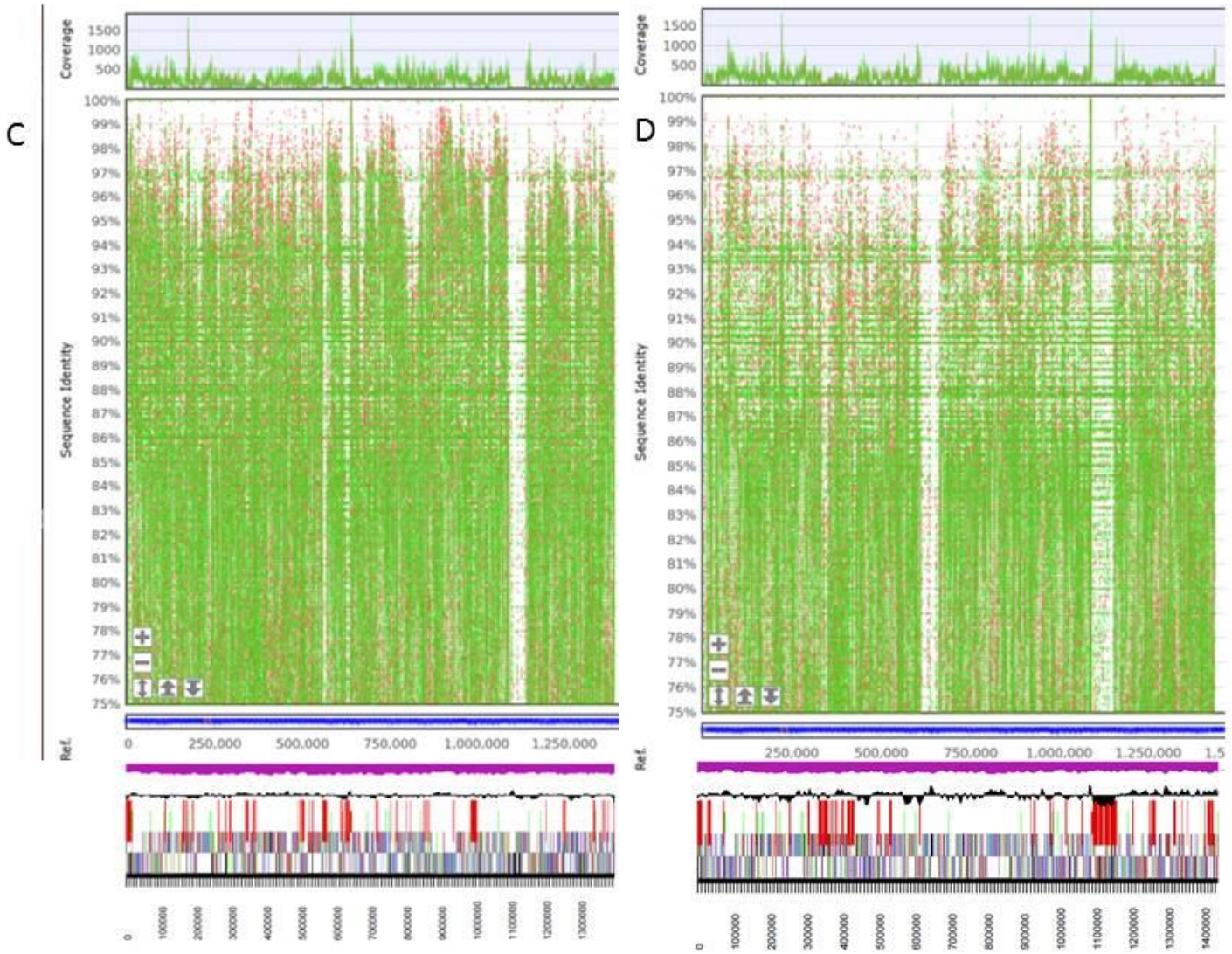

C
D
Coverage
1500
1000
500
Sequence Identity
100%
99%
98%
97%
96%
95%
94%
93%
92%
91%
90%
89%
88%
87%
86%
85%
84%
83%
82%
81%
80%
79%
78%
77%
76%
75%
Ref.
0
250,000
500,000
750,000
1,000,000
1,250,000
1.5
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
1100000
1200000
1300000
1400000

Figure 8. HIMB4 chromosome from coordinates 540000 to 580000 highlighting the region where DNA polymerase I, phage integrase, and other putative viral genes are present. The putative phage spans from approximately coordinates 560,000 to 567,000.

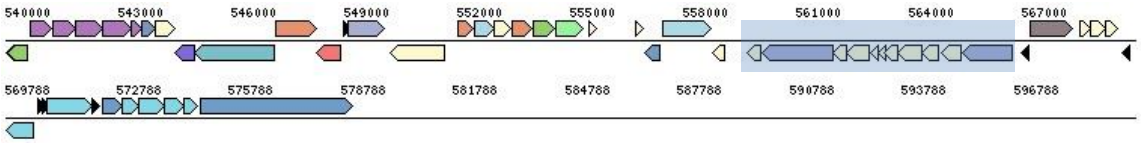


Table 1. Characteristics of metagenome samples taken from the *Sorcerer II* GOS expedition [information extracted from Rusch *et al*. (2007)].

ID	Location	Date	Time	Sample Depth (m)	Water Depth (m)	Size fraction (µm)	Habitat type	Good sequences
GS00a	Sargasso Stations 13 and 11	2/26/2003	3:00	5	>4,200	0.1-0.8	Open Ocean	644,551
GS00b	Sargasso Stations 13 and 11	2/26/2003	3:35	5	>4,200	0.22-0.8	Open Ocean	317,180
GS00c	Sargasso Station 3	2/25/2003	13:00	5	>4,200	0.22-0.8	Open Ocean	368,835
GS00d	Sargasso Station 13	2/25/2003	17:00	5	>4,200	0.22-0.8	Open Ocean	332,240
GS01a	Hydrostation S	5/15/2003	11:40	5	>4,200	3.0-20.0	Open Ocean	142,352
GS01b	Hydrostation S	5/15/2003	11:40	5	>4,200	0.8-3.0	Open Ocean	90,905
GS01C	Hydrostation S	5/15/2003	11:40	5	>4,200	0.1-0.8	Open Ocean	92,351
GS17	Yucatan Channel	1/9/2004	13:47	2	4,513	0.1-0.8	Open Ocean	257,581
GS18	Rosario Bank	1/10/2004	8:12	2	4,470	0.1-0.8	Open Ocean	142,743
GS22	250 Miles from Panama City	1/20/2004	16:39	2	2,431	0.1-0.8	Open Ocean	121,662
GS23	30 Miles from Cocos Island	1/21/2004	15:00	2	1,139	0.1-0.8	Open Ocean	133,051
GS26	134 Miles NE of Galapagos	2/1/2004	16:16	2	2,376	0.1-0.8	Open Ocean	102,708
GS37	Equatorial Pacific TAO Buoy	3/17/2004	16:38	2	3,334	0.1-0.8	Open Ocean	65,670
GS47	201 miles from French Polynesia	3/28/2004	15:25	30	2,400	0.1-0.8	Open Ocean	66,023
GS02	Gulf of Maine	8/21/2003	6:32	1	106	0.1-0.8	Coastal	121,590
GS03	Browns Bank, Gulf of Maine	8/21/2003	11:50	1	119	0.1-0.8	Coastal	61,605
GS04	Outside Halifax, Nova Scotia	8/22/2003	5:25	2	142	0.1-0.8	Coastal	52,959
GS07	Northern Gulf of Maine	8/25/2003	8:25	1	139	0.1-0.8	Coastal	50,980
GS08	Newport Harbor, RI	11/16/2003	16:45	1	12	0.1-0.8	Coastal	129,655
GS09	Block Island, NY	11/17/2003	10:30	1	32	0.1-0.8	Coastal	79,303
GS10	Cape May, NJ	11/18/2003	4:30	1	10	0.1-0.8	Coastal	78,304
GS13	Off Nags Head, NC	12/19/2003	6:28	1	20	0.1-0.8	Coastal	138,033
GS14	South of Charleston,SC	12/20/2003	17:12	1	31	0.1-0.8	Coastal	128,885
GS15	Off Key West, FL	1/8/2003	6:25	2	47	0.1-0.8	Coastal	127,362
GS19	Northeast of Colón	1/12/2004	9:03	2	3,336	0.1-0.8	Coastal	135,325
GS21	Gulf of Panama	1/19/2004	16:48	2	76	0.1-0.8	Coastal	131,798
GS27	Devil's Crown, Floreana	2/4/2004	11:41	2	2.3	0.1-0.8	Coastal	222,080
GS28	Coastal Floreana	2/4/2004	15:47	2	156	0.1-0.8	Coastal	189,052
GS29	North James Bay, Santiago	2/8/2004	18:03	2	12	0.1-0.8	Coastal	131,529
GS34	North Seamore	2/19/2004	17:06	2	35	0.1-0.8	Coastal	134,347
GS35	Wolf Island	3/1/2004	16:44	2	71	0.1-0.8	Coastal	140,814
GS36	Cabo Marshall, Isabella	3/2/2004	12:52	2	67	0.1-0.8	Coastal	77,538

Table 2. Average nucleotide identity across the genome (upper-right) and 16S rRNA gene sequence similarity (lower-left) of the four SAR11 strains used in this study.

	HIMB83	HIMB4	HIMB140	HIMB5
HIMB83	-	77.2	82.0	76.5
HIMB4	98.8	-	77.1	76.9
HIMB140	98.8	99.7	-	76.6
HIMB5	98.5	98.9	98.9	-

Table 3. General characteristics of the SAR11 genomes studied here, and summary of the recruitment profile of each strain. In order to make comparisons possible between coastal and open ocean data, recruited fragments are measured as a percentage of the total fragments present in their respective databases. Recruitment ratios favoring open ocean recruitment are highlighted in red. Ratios favoring coastal recruitment by no more than 10% are shown in yellow. Ratios favoring coastal recruitment by 10-30% are highlighted in light green and above 30% are highlighted in light blue.

	HIMB4	HIMB5	HIMB83	HIMB140
IMG Taxon ID	2503754001	2503982040	2510461079	2503754000
Genome Size (Mbp)	1.38	1.34	1.40	1.44
Protein-coding Genes	1453	1437	1468	1504
GC Content	28.95	28.63	29.15	29.35
Unique genes	141	120	100	145
Unique genes as a Percent of Genome	9.70%	8.35%	6.81%	9.64%
Percent of Coastal Recruitment ≥75% Identity	18.61%	18.23%	20.02%	19.96%
Percent of Open Ocean Recruitment ≥75% Identity	17.02%	16.90%	18.34%	18.14%
Percent of Coastal Recruitment ≥80% Identity	12.24%	11.62%	14.99%	14.05%
Percent of Open Ocean Recruitment ≥80% Identity	11.81%	11.75%	13.83%	12.97%
Percent of Coastal Recruitment ≥85% Identity	4.19%	3.66%	7.14%	5.26%
Percent of Open Ocean Recruitment ≥85% Identity	4.08%	3.88%	5.81%	4.67%
Percent of Coastal Recruitment ≥90% Identity	1.43%	1.17%	3.51%	1.75%
Percent of Open Ocean Recruitment ≥90% Identity	1.37%	1.21%	2.42%	1.50%
Percent of Coastal Recruitment ≥95% Identity	0.40%	0.26%	1.01%	0.39%
Percent of Open Ocean Recruitment ≥95% Identity	0.31%	0.21%	0.53%	0.27%
Coastal/Open Ratio of Recruitment ≥75% Identity	1.09	1.08	1.09	1.10
Coastal/Open Ratio of Recruitment ≥80% Identity	1.04	0.99	1.08	1.08
Coastal/Open Ratio of Recruitment ≥85% Identity	1.03	0.94	1.23	1.13
Coastal/Open Ratio of Recruitment ≥90% Identity	1.04	0.96	1.45	1.17
Coastal/Open Ratio of Recruitment ≥95% Identity	1.28	1.25	1.88	1.47

Table 4. Characteristics of HVR1 in the four studied strains. Unique genes are genes present in one strain only, and values for each gene category denote the number of genes in each region matched to those categories in the IMG-ER database. Several genes marked under the cell motility category are involved in the type II secretion/type IV pilus pathway.

	HIMB4	HIMB5	HIMB83	HIMB140
Size (kbp)	67	52	61	46
Genes	59	35	40	30
Unique Genes	4	1	0	0
Cell Motility	12	10	10	11
Signal Transduction Mechanisms	3	2	2	2
Amino Acid Transport/Metabolism	3	1	1	0
Intracellular Trafficking	1	1	1	1
Cell Cycle Control	2	1	1	1
Unknown	25	15	16	10
Other	9	4	9	5

Table 5. Characteristics of HVR2 in the four studied strains. Unique genes are genes present in one strain only, and values for each gene category denote the number of genes in each region matched to those categories in the IMG-ER database.

	HIMB4	HIMB5	HIMB83	HIMB140
Size (kbp)	61	39	19	63
Genes	58	42	19	61
Unique Genes	33	25	8	43
Cell Wall/Membrane/Envelope Biogenesis	20	14	9	19
Carbohydrate Transport and Metabolism	7	4	2	11
Amino Acid Transport and Metabolism	3	3	1	6
Lipid Transport and Metabolism	1	1	0	3
Energy Production and Conversion	0	3	0	1
Unknown Function	20	10	3	12
Other	7	7	4	9