AN ABSTRACT OF THE THESIS OF

Larry John Wilhelm for the degree of
Master of Science in Microbiology
presented on March 23, 2007.

Title: Conserved Properties in the Metagenome of a Large Bacterioplankton Population.

Abstract Approved:

_____
Stephen J. Giovannoni

Candidatus Pelagibacter ubique is the first cultured representative of the SAR11 clade, a clade that is found throughout the oceans and accounts for approximately 25% of all bacterial cells [1]. It has a streamlined genome that is the smallest of any known free-living organism. In this study the complete genome sequence of *Candidatus* Pelagibacter ubique (strains HTCC1062 and HTCC1002) is used to explore the genomic variability of this organism by taking advantage of the large amount of DNA sequence data collected by Venter et.al. [2] from the Sargasso Sea.

Pelagibacter gene homologues in the metagenomic data were identified by tblastn and screened by a reciprocal best blastx test against the NCBI database to identify fragments of probable SAR11 origin. Fragments passing both tests covered 97.8% of the HTCC1062 genome. A subset of fragments spanning two or more SAR11 genes was used to study the conservation of gene order between the Oregon coast isolates

and the Sargasso Sea SAR11 population. Boundaries between genes matched the gene order of the HTCC1062 genome in 96% of the cases (> 85,000 observations), although the average amino acid similarity of the genes encoded was only 71%. Alternate gene orders observed in the remaining 3432 fragments indicated that gene rearrangements within the Sargasso Sea population are more likely at boundaries between operons than within operons. Comparisons of the genomes of strains HTCC1062 and HTCC1002, and analysis of the metagenomic data, indicated four regions of genome variability, including a 48 kb cassette between the 23S rRNA gene and the 5S rRNA gene that encodes genes determining cell surface properties.

These findings indicate that the temperate gyre population of SAR11 is divergent in nucleotide and amino acid sequence from the coastal isolates, but shares similar gene order and composition in core regions of the genome. The methodology of binning environmental DNA fragments by using sequence similarity and gene-order conservation with known query genomes is validated and shows promise as a general technique to decipher the flood of metagenomic data that is accumulating.

Conserved Properties in the Metagenome of a Large Bacterioplankton Population.

by
Larry John Wilhelm

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented March 23, 2007
Commencement June 2007

Master of Science thesis of Larry John Wilhelm presented on March 23, 2007.

APPROVED:

_____

Major Professor, representing Microbiology

_____

Chair of the Department of Microbiology

_____

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

_____

Larry John Wilhelm, Author

ACKNOWLEDGEMENTS

The author expresses sincere appreciation to the following people for their help and support through the course of this study.

I owe gratitude to Scott Givan and Daniel Smith for their expert technical assistance. Jim Tripp has contributed a great deal in the development of various ideas in this thesis and having a cohort of similar age and humor has been crucial. Kevin Vergin has been an invaluable source of information and guidance, as well as the rest of the members of the Giovannoni laboratory, particularly Ulrich Stingl. Thanks to Robert Burton for the statistical analysis in Figure 10 and Dee Denver for his advice on collecting the evolutionary parameters in Table 4. Many thanks to Stephen Giovannoni for providing an environment where I could gain the skills that I was interested in obtaining in pursuit of this degree and for his encouragement and enthusiasm towards this project. I thank the Microbiology department at OSU and the Pernot fellowship. Though not the source of my direct funding, the Gordon and Betty Moore foundation have helped fund the larger team required to do work of this nature and their philanthropy is greatly appreciated.

AUTHOR CONTRIBUTIONS

H. James Tripp made major contributions in the early conceptual development of the methodology described here. He also did a great deal of work in decribing the gene content of the hypervariable regions and in the genomic comparison of strains HTCC1062 and HTCC1002. His contribution to numerous discussions on this project can’t be overestimated.

Daniel P. Smith improved the computer code for assembling syntigs and the syntig viewer. He also entirely produced the genome rearrangements figure.

Scott A. Givan provided the computing infrastructure and bioinformatics expertise without which none of this would have been possible. He also made large contributions to the editing and critiquing of the manuscript.

Larry J. Wilhelm developed all of the original algorithms and software and performed all of the analyses on the data. He also managed the manuscript preparation, review, and submittal process.

TABLE OF CONTENTS

Chapter Page

1 Introduction

A brief history of metagenomics…….. 1

The impact of metagenomics on the concept of bacterial speciation…… 6

2 Conserved properties in the metagenome of a large bacterioplankton population……. 10

Abstract……. 11

Introduction……... 12

Results and Discussion……. 15

Homologue detection…… 15

Syntigs…….. 20

Genome rearrangements…….. 27

Divergence among Sargasso Sea SAR11 populations…….. 32

Hypervariable regions……. 39

Genes conserved in the coastal isolates but not found in the Sargasso Sea……. 54

Conclusions……. 57

Materials and Methods…… 60

Homologue search…….. 60

Syntig detection…….. 61

Testing the syntig concept with different query genomes…… 62

TABLE OF CONTENTS (Continued)

Chapter		Page
	Fragments with SAR11 ribosomal RNA genes……	62
	Genome rearrangements……...	62
	Genes conserved in the coastal isolates but not found in the Sargasso Sea SAR11 populations……	63
	Searching for genes conserved in Sargasso Sea SAR11 populations and not found in the genomes of the coastal isolates…….	64
	Calculating a synteny index……..	64
	Tests of selection……..	65
	Accession Numbers of strains used in this study….	65
	Acknowledgements…….	65
3	General Conclusions……	67
Bibliography……		70

LIST OF FIGURES

Figure		Page
1	Environmental fragment detection and display……...	16
2	Overlay of homologue coverage, GC content, and syntig coverage for template genome HTCC1062…….	17
3	Expect score distribution among syntig genes…….	22
4	Syntig plots of three representative organisms……	23
5	Syntig plot of *Escherichia coli*……	26
6	GC content of syntigs……..	28
7	Syntig plot showing fragments carrying at least 3 genes……	29
8	Comparison of syntig analyses performed on unassembled reads, and sequence data containing assemblies…….	30
9	Rearrangements in the order of SAR11 genes in the Sargasso Sea metagenome, relative to the HTCC1062 genome….	31
10	Statistical analysis of observed genome rearrangements……	35
11	Homologous fragments with synteny in region of proteorhodopsin gene……..	41
12	Enlargement of HVR2…….	42
13	Enlargement of HVR3 and HVR4……...	43
14	Enlargement of HVR1…….	44
15	Deletion of duplicate genes in the HVR1 region of strains HTCC1002 and HTCC1062……	53

LIST OF TABLES

Table		Page
1	BLAST and synteny analysis of the HTCC1062 genome against the Sargasso Sea data set…….	19
2	Summary of syntig analysis of 7 marine microbial species and *E. coli*……	24
3	Genes from the coastal SAR11 strains that had homologs in the Sargasso Sea data, but were never found on syntigs……..	34
4	Non-synonymous (Ka) and synonymous (Ks) substitution rates for HTCC1062 genes measured with syntig data……...	38
5	Gene content of HVR2 (523–568), by functional category……	46
6	HVR2 Genes……...	47
7	HVR gene content summary……..	48
8	HVR1 Genes……..	49
9	HVR3 Genes…….	50
10	HVR4 Genes…….	51
11	Evidence for genes specific to the coastal variant of SAR11…..	56

DEDICATION

This thesis is dedicated to the most important people in my life, my wife Claire Careaga, my mother Marilyn Wilhelm, and the memory of my father Richard Wilhelm.

Conserved Properties in the Metagenome of a Large Bacterioplankton Population.

CHAPTER 1.
GENERAL INTRODUCTION

A brief history of metagenomics

Ever since Robert Koch established his pioneering postulates on the microbial nature of disease [3], the field of microbiology has centered on the process of cultivating individual microbial species. Growing an organism in pure culture has been the critical first step towards understanding the properties of a given microbe and has proven invaluable in understanding and controlling the many diseases and industrial processes that microbes govern. This culture-dependent, reductionist approach has produced the many impressive successes of microbiology in the 20th century. However, the limitations of culture-dependent studies were first recognized by what has become known as the 'great plate count anomaly' [4]. The standard methodology of retrieving samples from the environment and quantitating the number of active bacteria by counting the number of colony forming units on a petri plate was often in disagreement with the number of cells observed by direct microscopic examination [5]. The magnitude of the error was large (several factors of 10) regardless of whether the sample was aquatic or terrestial [6-8]. The diversity, distribution, and relative abundance of these uncultivated organisms was largely unknown until the advent of molecular techniques that utilize structural RNA molecules as molecular clocks [9]. The remarkably powerful concept of using the 16S rRNA molecule as an evolutionary chronometer has not only become the basis of culture-independent studies, but indeed

led Carl Woese to redraw the tree of life in 1987 [10]. The 16S rRNA molecule is the most commonly employed molecule in phylogenetic and microbial diversity studies because it posseses the key properties of a good molecular chronometer: a) changes to it occur in a clock-like, random fashion, b) changes occur at a rate that allows a wide range of evolutionary distance to be measured and, c) it is large enough (sequence length is sufficient) to capture enough data to make it a smooth-running clock. The relatively small number of independent domains (cloverleaf structure) of the 16S rRNA molecule makes it a near-perfect clock because nonrandom changes in one domain do not appreciably effect the other domains. These structural features coupled with the ubiquity of the molecule – it is present in every known bacterial cell – make it the preferred choice for inferring prokaryotic evolutionary relationships. Any protein-coding gene, even highly conserved genes for proteins like cytochrome *c*, can't compete as molecular clocks because the redundancy of the genetic code and the range of allowable amino acid substitutions in a protein (without altering function) allows any protein coding gene to wander over sequence space to a much greater extent than a structural RNA molecule.

The use of the 16S rRNA molecule to probe the uncultivated microbial world has uncovered remarkable microbial diversity [11] and provided a means to assess the relative abundance of microbes from various environments that elude detection by other methods, let alone any attempts at cultivation. The popularity of the method is evidenced by the fact that a PubMed search on the term '16S rRNA phylogeny' retrieves some 6700 references at the time of this writing. However, the degree to

which divergence at the 16S rRNA locus reflects the overall genetic and phenotypic divergence of two organisms remains controversial. Phage-mediated lateral transfer of environmentally important genes may cause two organisms to behave quite differently under competitive pressure for resources even though they share a high degree of 16S rRNA sequence similarity. Understanding the ecology of microbes in their natural environment requires an understanding of a) the members involved, and b) the activities of those members. Cloning 16S rRNA molecules directly from the environment has added much to our understanding of the diversity of the individual players and their relative abundance but tells us little about their ecological role in the community. The need to ascertain the physiology, and thus ecological role, of the individual as-of-yet uncultured members of a microbial community has spawned the field of metagenomics.

Metagenomics has been defined as the functional and sequence-based analysis of the collective microbial genomes in an environmental sample [12]. The process begins with the collection of raw DNA from the environment then proceeds to the cloning of the DNA by various means into a suitable host to produce a metagenomic library. The idea started with Pace [13] and was first realized by Schmidt [14] with seawater samples and phage lambda as a vector. Once a metagenomic library has been collected it is mined for information, either the presence of particular genes of interest by their selective amplification via polymerase chain reaction (PCR), or the direct expression of enzyme activity. One of the main limitations of metagenomic libraries is that each clone represents only a small fragment of the genome from which it derived.

Advances in cloning technologies, particulary the use of F1-origin based cosmid (fosmid) vectors, and bacterial artificial chromosomes (BACs), has increased the size of the cloned DNA fragment from the original 20-40Kb reported by Schmidt to >300Kb in a fosmid library [15]. Such increases in average insert size increase the likelihood of finding a phylogenetic anchor, such as a 16S rRNA gene, on the insert and thus increases the probability that an investigator can assign the function associated with the genes on an insert to a specific member of the community. Another approach is to attempt the reconstruction of individual genomes from the metagenomic data, or at least bin the environmental sequences based on GC content and sequence coverage as Tyson *et.al*. [16] did with a microbial community from an acid mine drainage. Tyson was able to associate a 16S rRNA phylogenetic anchor with each of his sequence bins and thus associate a collection of genes with a particular member of the community.

A relatively new approach to metagenomics is the construction of shotgun metagenomic sequence libraries. Random shotgun sequencing involves the random mechanical shearing of DNA into small pieces (2KBP – 10 KBP) and subsequent cloning to produce a small insert clone library. The inserts are then sequenced and assembled with a computer. The technique was originally developed by Claire M. Fraser and J. Craig Venter to sequence the *Haemophilus influenzea* genome [17,18]. At the time the assembly programs had been written to build consensus cDNAs from tens of thousands of expressed sequence tag (EST) sequences. Fraser and Venter realized that they could leverage this software to assemble the small insert clones of a

H. influenzae random shotgun library, which they did successfully. They proceeded to sequence *Mycoplasma genitalium* to further prove the concept and then most famously used the technique to sequence the human genome [19]. The technique has become the method of choice for genome sequencing projects of all types. There are currently over 450 complete microbial genomes listed at the National Center for Biotechnology Information (NCBI) that have been sequenced with this technology.

The first reports of the application of high-throughput random shotgun sequencing to the field of metagenomics appeared in 2004 when J. Craig Venter applied it to Sargasso Sea samples [2] and Gene Tyson[16] to an acid-mine drainage community. Up to that point the technique had been applied to DNA from clonal sources – an organism was grown in pure culture, the DNA extracted and purified, then subjected to the mechanical shearing process. The only difference from a technological standpoint in applying the technique to environmental samples was the source of the DNA. Venter filtered seawater between 1µm and 8µm filters (to remove most eukaryotes) then simply prepared the DNA. Tyson prepared his DNA from a biofilm sample collected from an acid-mine drainage. While Tyson was able to partially assemble a few genomes from his samples, probably due to the fact that he started with a relatively simple community, Venter was highly unsuccessful in assembling genomes from his plethora of sequence data, but then he was dealing with a much more complex and ancient community.

Metagenomics holds great potential for discovering novel genes and elucidating the structure of microbial communities where individual members are recalcitrant towards

cultivation. While the modern high-throughput shotgun methodology provides remarkable amounts of sequence data it is hampered by the fact that it remains impossible to assemble individual genomes when there exists a great deal of neutral sequence variation in the gene pool (a problem not encountered when the technique is applied to clonal populations). The inability to associate a good phylogenetic anchor to a given stretch of sequence seriously hinders the interpretation of the data from the perspective of a microbial ecologist. For this reason the more traditional approaches that clone larger pieces into fosmids or BACs provide an important complementary approach. The work presented in this thesis addresses this problem of assigning an individual sequence read from a metagenomic library to a given organism from the community in which it derived.

The impact of genomics on the concept of a prokaryotic species

Taxonomic classification of the bacterial world has always challenged microbiologists and controversy around the subject continues to this day. The asexual nature of bacterial reproduction is such a fundamental difference in biology compared with sexually reproducing metazoans and plants that the definition of a species itself as applied to higher organsisms is hardly applicable to bacteria. An acceptable hierarchical classification system to describe the breadth of diversity and evolutionary history of the microbial world requires a firm theoretically-based definition of a bacterial species. Before the advent of molecular techniques bacteria were classified by all of the properties that current technology made observable, namely morphology,

physiology, and metabolism. The problem with classification systems based on these phenotypic properties is that they fail to reveal the evolutionary relationships among members and can incorrectly group organisms based on an emergent phenotype – such as the ability to oxidize a given carbohydrate. This problem works in the other direction as well, as some closely related bacteria may exhibit significantly different morphologies and physiologies based on the presence of just a few horizontally transferred genes and be incorrectly placed in separate groups on this basis. The inherent weakness of phenotype-based systematics has been ameloriated as much as possible by employing polyphasic classification schemes that group organisms by as many different characteristics as possible. As technology progressed, particularly the chemical analysis of nucleic acids, direct genotypic information was available to help classify microbes. With purified DNA and simple chemical analyses the mol% G+C content of an organism's genome could be easily determined and quickly became a necessary datum to collect when describing a new species. Two organisms with highly different mol% G+C content were clearly not related. DNA-DNA hybridization has also became a standard measurement to make when comparing the relatedeness of two taxonomic entities. The principle rests on the notion that the more similar two DNA molecules are the more they can hybridize with each other by classic Watson-Crick base pairing. Hybridization techniques proved able to better delineate clusters of strains than simple mol% G+C or phenotypic properties [20]. Both of these molecular methods are still crude approximations of the genotypic differences between organisms. The advent of 16S rRNA molecular phylogenies had great impact on

microbial systematics because it provided information on the evolutionary relationships among the organisms under a proposed classification scheme. Being based on sequence analysis it is far less crude than mol% G+C or degree of hybridization. The molecular-clock like nature of the 16S rRNA molecule (discussed in previous section) make it an ideal measure of the relatedeness of individual microbes, and 97% similarity at the 16S rRNA locus has been largely accepted as a degree of separation that corresponds well to the delineation of species in prokaryotes. As useful as the 16S rRNA molecule has turned out to be it is not without its inherent limitations. The degree to which an organism's entire genome can change within the time it takes for a single mutation to occur at the 16S rRNA locus is not well understood but it is thought to be significant. Bacterial genomes are thought to consist of a core genome shared by all members, and a 'pan-genome' consisting of genes that are variably present in different members of the species. Core genes that are vertically-inherited should produce phylogenies that are consistent with the 16S rRNA locus and these phylogenies will reflect any speciation event that resulted from evolutionary pressure acting on the products of core genes. Speciation events that result from the lateral-transfer of genes will obfuscate the 16S rRNA and core-gene phylogenies. If the aim is to develop a theoretically-based classification system for microorganisms that groups species based on shared evolutionary history then it is critical that the impact of laterally transferred genes on speciation events is known. The age of genomics has brought answers to questions of this nature within the realm of possiblity. Classifying organisms based on the content of their genomes is the

obvious extention of measuring general properties such as mol% G+C and hybridization levels, and circumsribes the inherent limitation of comparing organisms based on just one (16S rRNA), or a few (core-genes) loci. Genomic studies provide the information on gene content that is used to identify key phylogenetic markers for a given group of organisms in a proposed classification system based on the Genomic-Phylogenetic Species Concept (GPSC) [21]. The GPSC considers expression data in addition to gene content and phylogeny. There are several key advantages to this system: a) It is a methodology based on sequence data and thus not overly biased to interpretation. b) An organism's ecology is encompassed because niche-adaptative changes are reflected in the sequence data of functional proteins. c). Non-adaptive changes are encompassed as well because genetic differences caused solely by drift are reflected in the content and sequence of core genes.

The thesis presented here utilizes existing complete genome sequences of cultured strains and the currently available metagenomic data to gather information on the core and pan genome of the SAR11 clade. By studying not just a handful of genomes but a whole population of them (~775) information on the breadth of sequence, gene-content, and gene-order variability is gained. Such information is the foundation of modern approaches to microbial systematics that encompass both genomics and phylogenetics.

CHAPTER 2.

Conserved Properties in the Metagenome of a Large Bacterioplankton population

Larry J. Wilhelm[1], H. James Tripp[1], Scott A. Givan[2], Daniel P. Smith[1], and Stephen J. Giovannoni[1]

[1]Department of Microbiology.

[2]Center for Genome Research and Bioinformatics

Oregon State University, Corvallis, OR 97331

Proceeding of the National Academy of Sciences
500 Fifth Street, NW
Washington, DC 20001
<In Review>

Abstract

Genome sequences from coastal isolates of the alphaproteobacterium *Candidatus* Pelagibacter ubique (strains HTCC1062 and HTCC1002) were used to identify related SAR11 DNA fragments in Venter's environmental shotgun sequence data set from the Sargasso Sea. Pelagibacter gene homologues in the metagenomic data were identified by tblastn and screened by a reciprocal best blastx test against the NCBI database to identify fragments of probable SAR11 origin. Fragments passing both tests covered 97.8% of the HTCC1062 genome. A subset of fragments spanning two or more SAR11 genes was used to study the conservation of gene order between the Oregon coast isolates and the Sargasso Sea SAR11 population. Boundaries between genes matched the gene order of the HTCC1062 genome in 96% of the cases (> 85,000 observations), although the average amino acid similarity of the genes encoded was only 71%. Alternate gene orders observed in the remaining 3432 fragments indicated that gene rearrangements within the Sargasso Sea population are more likely at boundaries between operons than within operons. Comparisons of the genomes of strains HTCC1062 and HTCC1002, and analysis of the metagenomic data, indicated four regions of genome variability, including a 48 kb cassette between the 23S rRNA gene and the 5S rRNA gene that encodes genes determining cell surface properties. These findings indicate that the temperate gyre population of SAR11 is divergent in nucleotide and amino acid sequence from the coastal isolates, but shares similar gene order and composition in core regions of the genome.

Introduction

A particularly vexing aspect of microbial genomics is the common observation of high genome variability among strains of a species [22-24]. Such observations have raised significant questions about the validity of the microbial species concept, and the value of single genome sequences for comparisons between taxa [25]. To reconcile this dilemma, it has been suggested that bacterial species have a "core-genome" consisting of genes that are always present, and a "pan-genome" of genes that are variably present [24]. Metagenomics, the study of genome sequence retrieved from mixed assemblages of organisms collected from nature, is providing high coverage of genome sequence variation from some natural microbial communities [16].

The Sargasso Sea metagenomic data consists of 1.6 G base pairs of unique environmental genomic DNA shotgun sequence. The SAR11 clade accounts for 380 of the 1412 16S rRNA genes in the Sargasso Sea data (27%), suggesting that it includes enough SAR11 genome sequence data to encode the equivalent of about 775 SAR11 strain HTCC1062 genomes [2]. The Sargasso Sea is an oligotrophic subtropical gyre where average surface temperatures are about 23 C, and rarely drop below 20 C [26]. SAR11 strain HTCC1062 was isolated from the cold, nutrient rich Oregon coast [27,28]. HTCC1062 and other closely related coastal Oregon isolates (HTCC1002) belong to a 23S-ITS-16S phylogenetic cluster that is distinct from SAR11 23S-ITS-16S sequences reported by Venter from the Sargasso Sea [29]. The data suggest that they represent a genetically distinct, allopatric population.

Despite the abundance of SAR11 genome sequences in the Sargasso Sea data, the assembly of SAR11 genomes failed when traditional DNA assembly methods were applied [2]. The largest SAR11 rRNA-anchored scaffold reconstructed with the Celera Assembler was relatively small (ca. 21,000 bp), and assembly depth-of-coverage was low (0.94 – 2.2 fold) [2]. This observation led to speculation that the SAR11 clade might be a diverse assemblage of many species, each with low coverage in the shotgun sequence library [2]. However, ecological data suggests that the SAR11 clade consists of a few ecotypes, which can be differentiated either phylogenetically [30], or by their appearance in the environment at different depths and seasons [31]. To reconcile these observations, we hypothesize that the large population sizes of SAR11, and the age of these clades, allow them to accumulate very extensive neutral sequence variation, but that genome properties subject to selection will be conserved [23].

Here we describe using the complete genome sequence of HTCC1062 as a query to identify metagenomic DNA fragments that originated from the SAR11 clade. Our approach is similar to that employed by Hallam et.al. [32], who used a composite genome sequence of *Crenarchaeum symbiosum* to study the diversity of Marine Group I archaeal genomes in metagenomic databases using BLAST score ratios [33] to identify conserved genes. Our study relies on a reciprocal best-hit BLAST test that compares predicted amino acid sequences to identify fragments encoding SAR11

homologues. We quantified gene-to-gene boundaries to assess gene insertions, deletions and rearrangements, and the occurrence of non-SAR11 genes on fragments encoding SAR11 homologues. We measured the conservation of synteny and display the relationship between synteny and amino acid identity in a novel way that allows observation of small-scale (<5) gene insertions. The results suggest that extraordinarily high allelic variation and genome rearrangements mask the conservation of many genome properties in native SAR11 populations.

Results and Discussion

Homologue Detection. The expect score of $1 \times e^{-10}$ employed in the initial homologue detection step (Fig. 1A) is a relatively permissive cutoff that ensures the inclusion of homologues, including those from distant taxa, such as other alphaproteobacteria. For convenience, we hereafter refer to this set of fragments as 'homologous fragments'. The log of the number of homologous fragments for each HTCC1062 ORF is shown as a function of gene position in Fig. 2A. We found that homologous fragments from the Sargasso Sea cover 97% of the HTCC1062 genome and account for 43% of the fragments in the complete dataset (349,742 of 811,372, Table 1). This number is somewhat high when compared to predictions from SSU rRNA data - SAR11 genes account for 27% of the total SSU rRNA genes found among the fragments. This observation is not unexpected, considering that the set of homologous fragments contains homologues from non-SAR11 species as well as genes originating from SAR11. Of the 1354 ORFs in the HTCC1062 genome, 32 are found on at least 3000 environmental fragments (the maximum number returned from our homologue search), and 31 are not found at all. Fig. 2A reveals three regions of the HTCC1062 genome where coverage is low. These hypervariable regions are labeled HVR1 through HVR3. The longest of these, HVR2, spans almost 50 ORFs.

Figure 1. Environmental fragment detection and display. Template organism genes are represented on x-axis. A) Homologue detection step. Amino acid sequence of template gene returns $1 \times e^{-10}$ or better expect score to environmental fragment nucleotide sequence (tblastn). B) Fragments must contain homologues in same gene order as template. C). Reciprocal best-hit test. Nucleotide sequence of environmental fragment gene yields corresponding template gene as best hit in blastx comparison to NCBI nr (non-redundant) database. D) Fragment is drawn on vertical axis corresponding to average amino acid identity score of all genes on fragment, from tblastn search at step A.

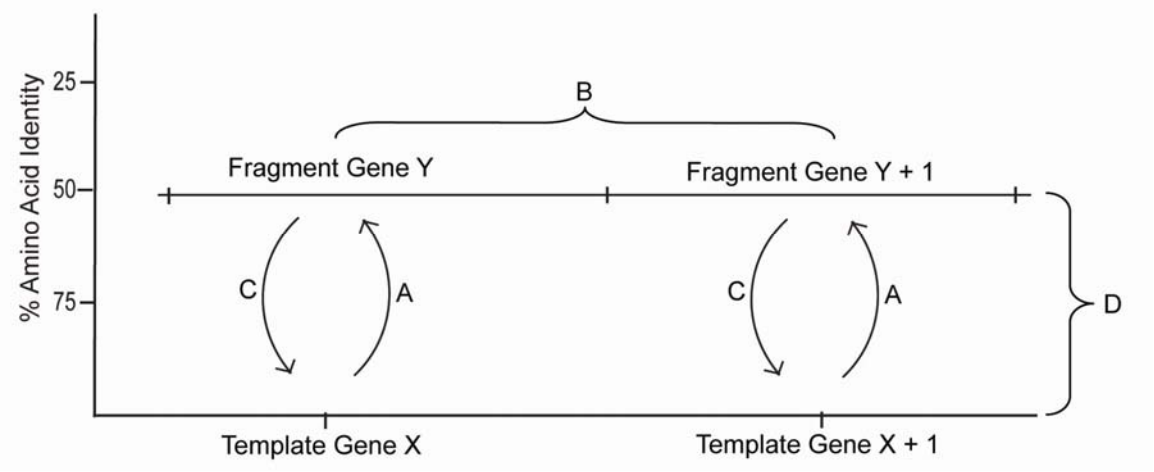


Figure 2. Overlay of homologue coverage, GC content, and syntig coverage for template genome HTCC1062. A). The number of environmental fragments that contain homologues to each HTCC1062 gene, plotted by position in the HTCC1062 genome. B) GC content of HTCC1062 genome, using a window size of 300 nucleotides. C) The distribution of environmental fragments with synteny to HTCC1062 genome (syntig plot). See Fig. 1 for explanation of syntig plot. Regions of blue on the fragments indicate gaps. Syntigs were allowed to be missing as many as five intervening genes (gaps) between the syntenous genes. Genes that encode ribosomal proteins are indicated in black.

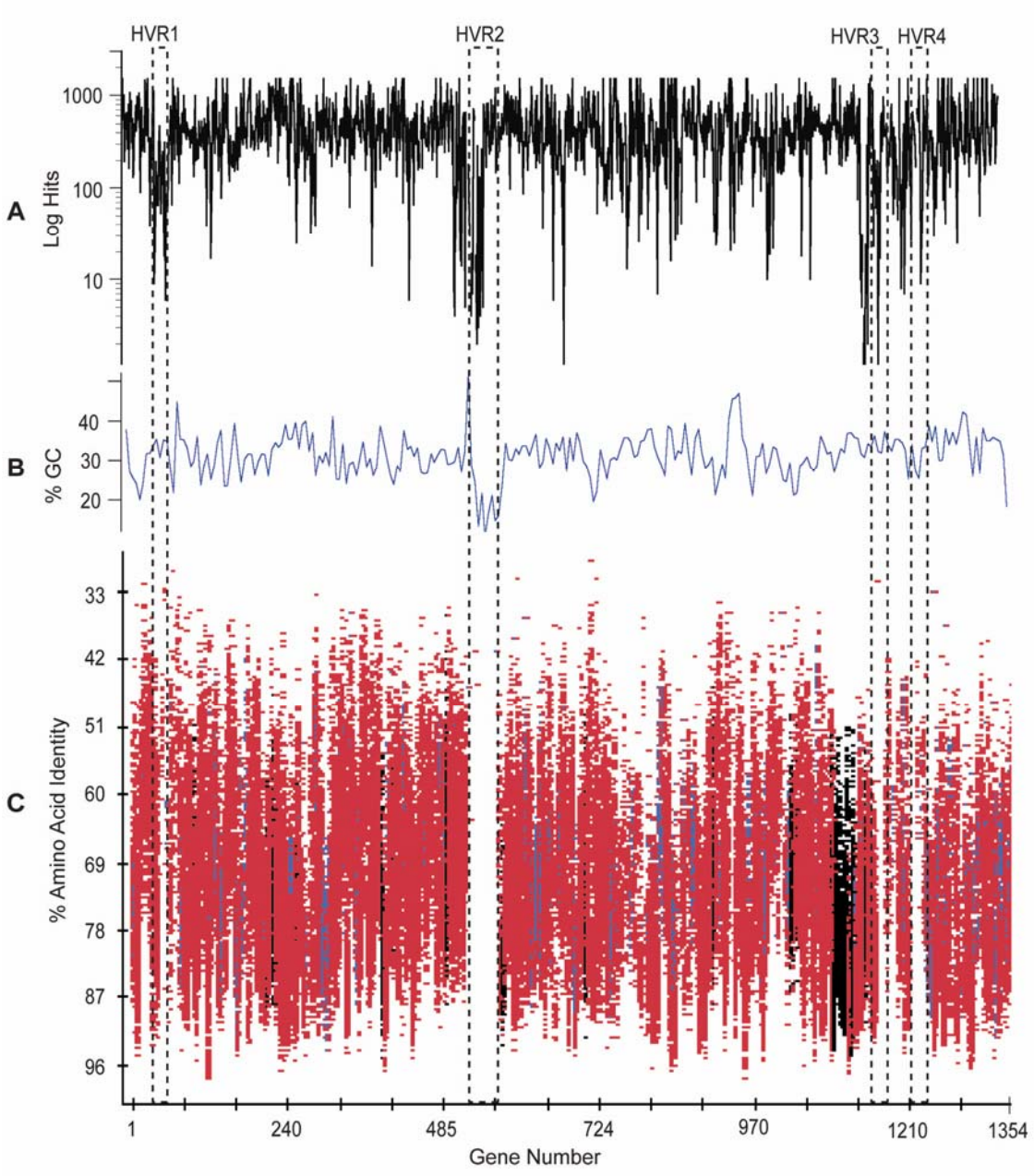

HVR1
HVR2
HVR3
HVR4
A
Log Hits
1000
100
10
B
% GC
40
30
20
C
% Amino Acid Identity
33
42
51
60
69
78
87
96
1
240
485
724
970
1210
1354
Gene Number

Figure 2.

Table 1. BLAST and synteny analysis of the HTCC1062 genome against the Sargasso Sea data set.

Class	Number	Genome coverage	Avg. AA Identity
ORFs in HTCC 1062	1354	--	--
Fragments In Dataset	811,372	--	--
Homologous Fragments	349,742	97%	--
Homologous Fragments With Synteny	111,332	97%	64.0%
Syntigs*	71,696	91%	71.0%

* For all genes on the fragment, HTCC1062 genes are the best BLAST hits.

Syntigs. Syntigs are homologous fragments that pass the more stringent synteny and best-hit tests (Fig. 1B,C). Of the 349,742 fragments shown in Fig. 2A, 111,332 share synteny with HTCC1062. A large proportion of these (71,696) are syntigs, passing the best-hit test (Fig. 2C). The numbers of fragments at each stage of the process are shown in Table 1. The distribution of expect scores for the genes on syntigs has a maximum at $1 \times e^{-33}$ and declines sharply as it approaches $1 \times e^{-10}$ suggesting that the cutoff used in the homologue detection step was appropriate (Fig. 3). In other words, the decline in number of syntigs returned at higher expect scores indicates that we did not exclude fragments in the initial homologue detection step that were likely to pass the subsequent criteria necessary to become a syntig.

To visualize variation among the syntigs, the data were plotted as a function of gene position in the HTCC1062 genome and amino acid identity (Fig. 1D). The vertical axis is inverted (low amino acid identity scores at the top) to emphasize the high-scoring syntig's relationship to the template genome. The syntigs shown in Fig. 2C vary in average amino-acid identity score from 30% to 98%, with an average of 71.0%. The ranges spanned on the vertical axis vary between genes because the amino acid sequences of some genes are more conserved than others. For example, genes for ribosomal proteins appear relatively low in the plots (Fig. 2C, regions colored black).

To validate the syntig process, we compared syntig plots of HTCC1062 and six other marine microbial genomes (Fig. 4 and Table 2) and found that the plots are characteristic for each group, and the numbers of syntigs in the plots correlate with the abundance of each organism's 16S rRNA genes in the metagenomic data (Fig. 4). For each selected organism, the relative abundance of genomic DNA within the

Figure 3: Expect score distribution among syntig genes.

Figure 4. Syntig plots of three representative organisms. The bar chart in the upper right corner indicates the number of fragments containing the target organism's 16S rRNA, at the indicated degree of similarity. The horizontal line indicates the average syntig score.

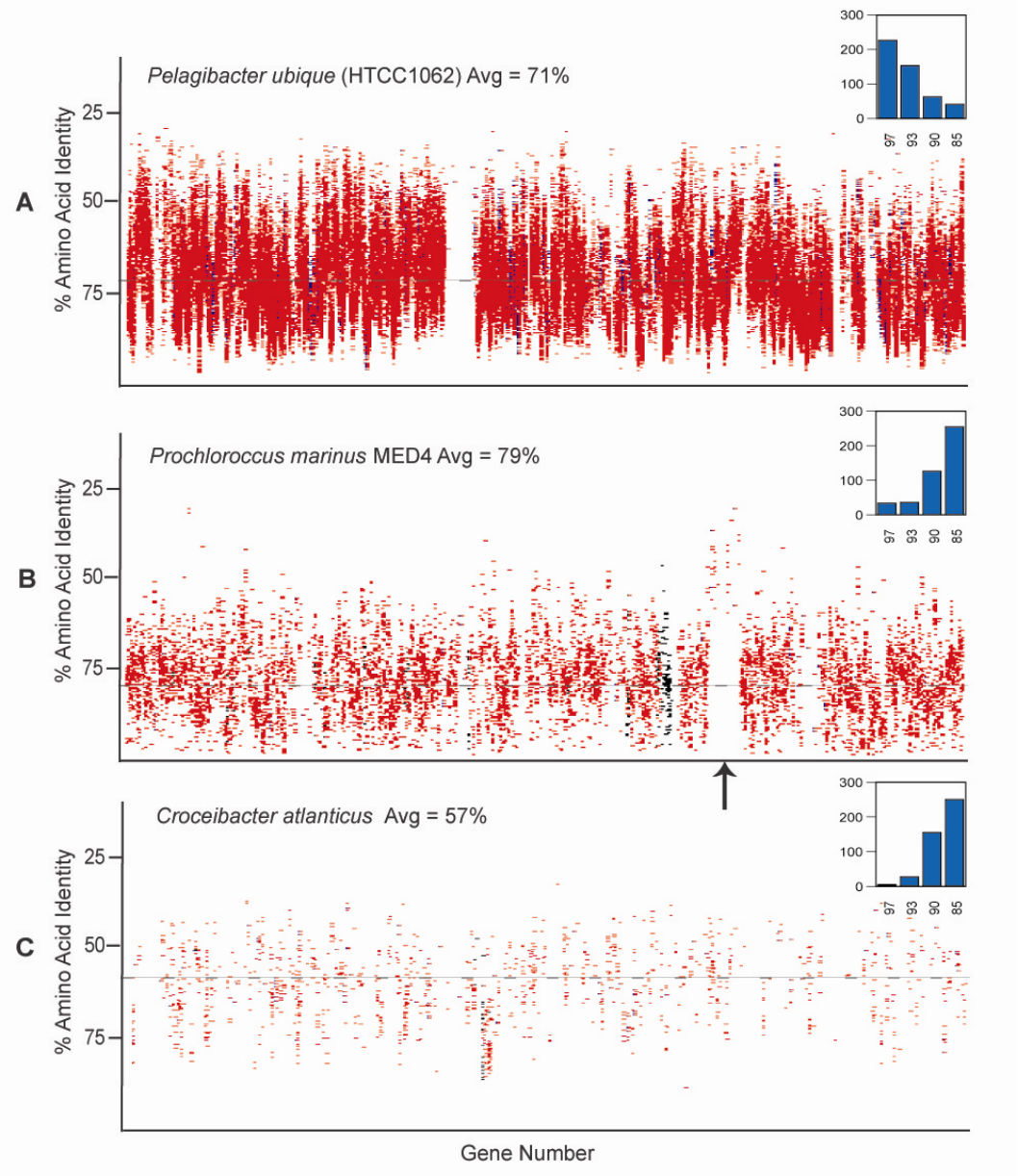


Table 2. Summary of syntig analysis of 7 marine microbial species and *E. coli*.

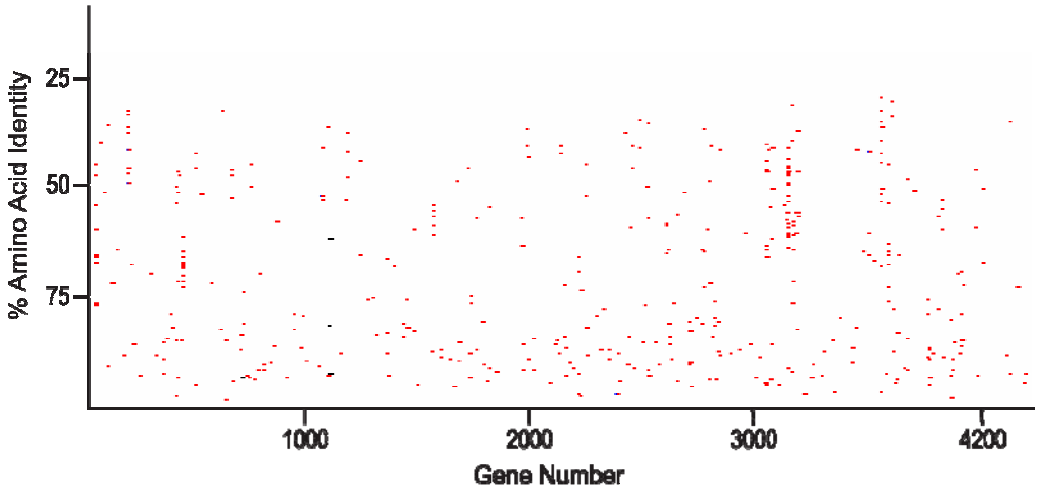
Species	Orfs	Fragments with Homology	Fragments with Synteny	Fragments Passing Best-Hit	% Unique to group
Pelagibacter ubique	1332	349742	99457	71696	98.8
Prochlorococcus marinus (MED4)	1713	226594	21304	8398	99.9
Croceibacter atlanticus	2633	256645	15089	1258	99.0
Oceanicola batsensis	4614	354793	49278	532	60.5
Oceanicaulis alexandrii	3365	330225	36845	500	31.0
Escherichia coli	4289	306293	33306	406	92.9
Parvularcula bermudensis	2824	310044	30809	239	31.8
Janibacter sp.	4367	240360	18570	23	91.3

Sargasso Sea metagenome is estimated by the number of 16S rRNA fragments satisfying identity thresholds of 97%, 93%, and 90%. These values are shown as insets in figure 4. The number of syntigs recovered for each organism correlates with the number of similar 16S rRNA genes. The most abundant organism, SAR11, produced by far the most syntigs (71,696) while *P. marinus* returned an intermediate number (8,398). Other organisms (e.g. *C. atlanticus* in Fig. 4C, Table 2) that are virtually undetectable by 16S rRNA analysis returned a number of syntigs (1258) similar to our negative control *E. coli* (406). Figure. 5 shows the distribution of syntigs recovered for *E. coli*.

If the syntig detection process described here accurately recovers environmental fragments arising from a given template organism, then the list of syntigs should be unique for each organism. Thus, to estimate the selectivity of our method we compared the syntig lists of all organisms tested. The percentage of syntigs unique to each organism (Table 2) is 98.8 % for HTCC1062 and 99.0 % for *P. marinus*. Less-abundant organisms have as few as 30% unique syntigs, but not in all cases. In numbers, the low abundance organisms are not distinguishable from each other or *E. coli*, though other marine microbes show a narrower, more distinct syntig pattern than *E. coli* , similar to Fig. 4C (data not shown).

The average GC content of the syntigs, 29.1 percent, is nearly identical to that of the HTCC1062 genome (29.7%) and the plotted values approximate a normal distribution

Figure 5. Syntig plot of *Escherichia coli.*



(Fig. 6). The results are consistent with the interpretation that the syntigs originate from a discrete pool of genomic DNA. Though 86% of the syntigs carry just two genes, the distribution of syntigs across the HTCC1062 genome is very similar when syntigs carrying 3 or more genes are plotted (Fig. 7).

The Sargasso Sea metagenome data includes assemblies that are known to contain some errors resulting from unrelated fragments being joined incorrectly [34]. However, syntig analyses from unassembled sequence reads yielded similar plots that supported the same conclusions obtained from assembled reads. We chose to include assemblies in our analyses because they provide more information about the linear order of genes as compared to the shorter, single reads (Fig. 8).

Genome Rearrangements. Syntigs cover most of the HTCC1062 genome (Fig. 2C), indicating that gene order is relatively conserved between the Pacific coastal strains and the Sargasso Sea SAR11 population. An analysis of gene-to-gene boundaries revealed that gene order in 96% of the Sargasso Sea SAR11 homologous fragments matches the gene order of the HTCC1062 genome (see synteny index, in Methods). While the aforementioned facts are striking, they nonetheless allow considerable latitude for genome rearrangements. To map genome rearrangements we plotted the gene positions found on non-syntenous fragments (Fig. 9). These are fragments containing two or more adjacent SAR11 genes that differ from HTCC1062 in gene order. Fragment inclusion in this set follows the same rules as those in Fig. 1, except

Figure 6. GC content of syntigs.

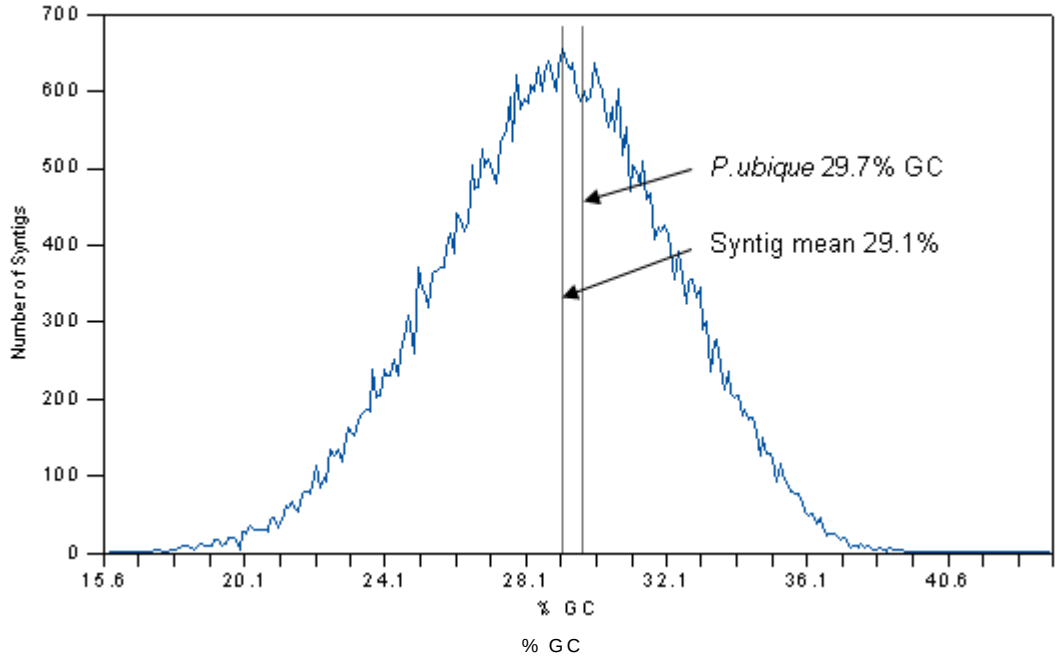


Figure 7. Syntig plot showing fragments carrying at least 3 genes.

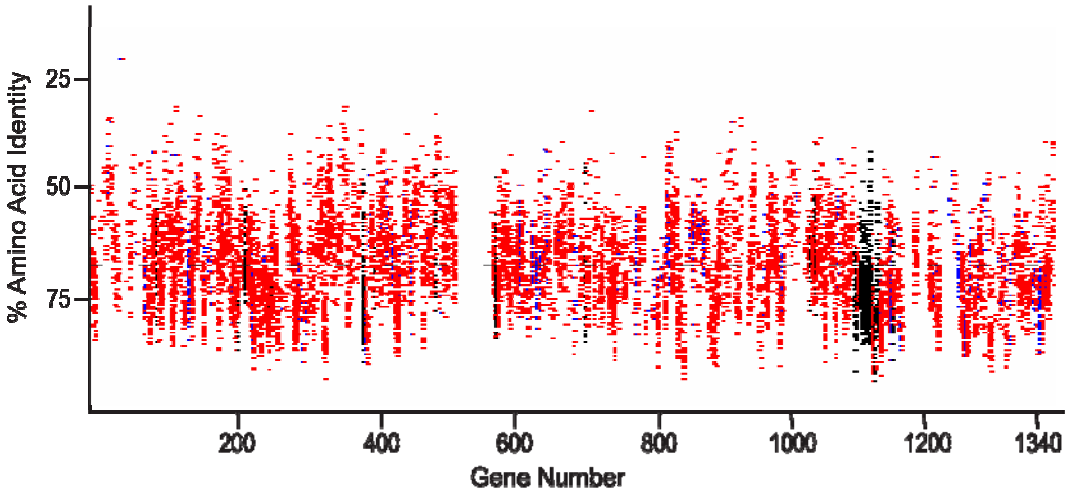
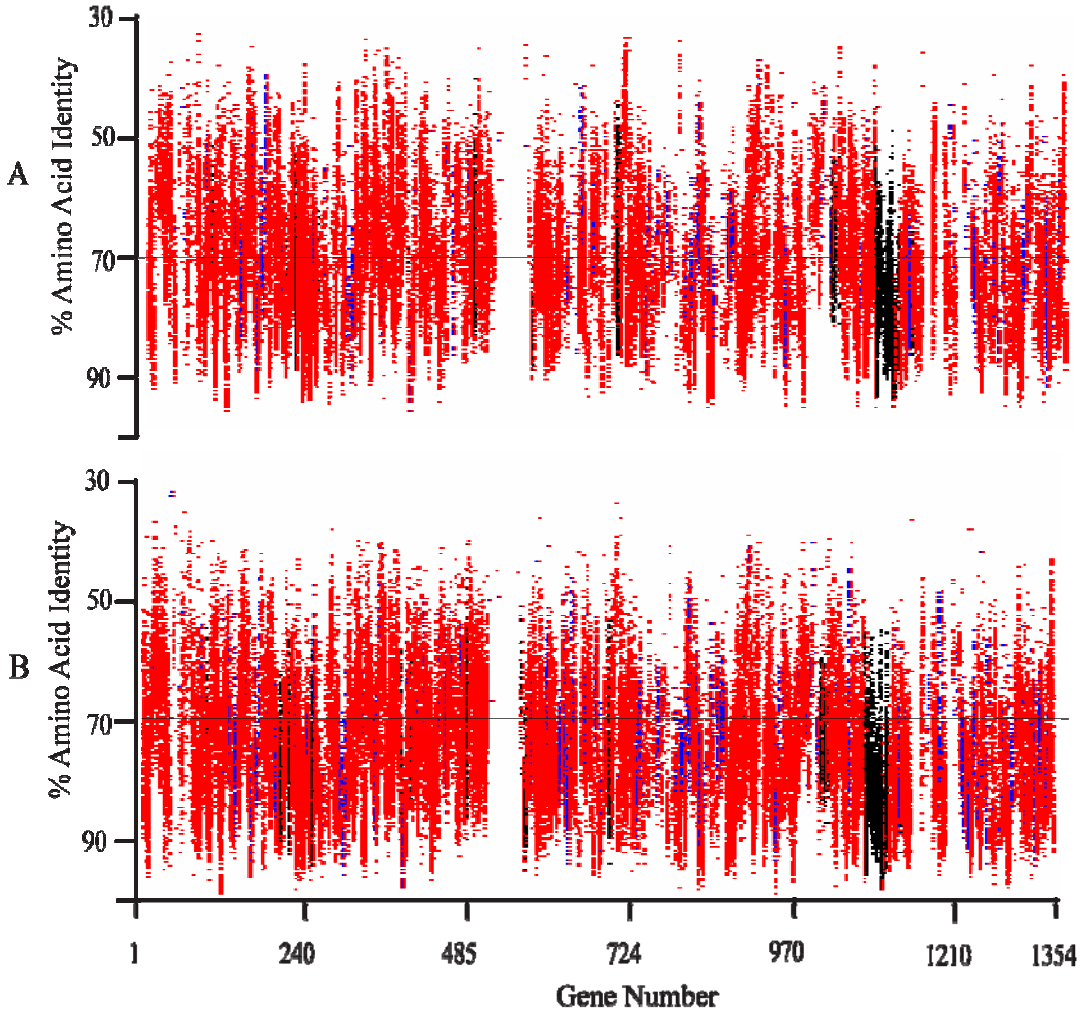


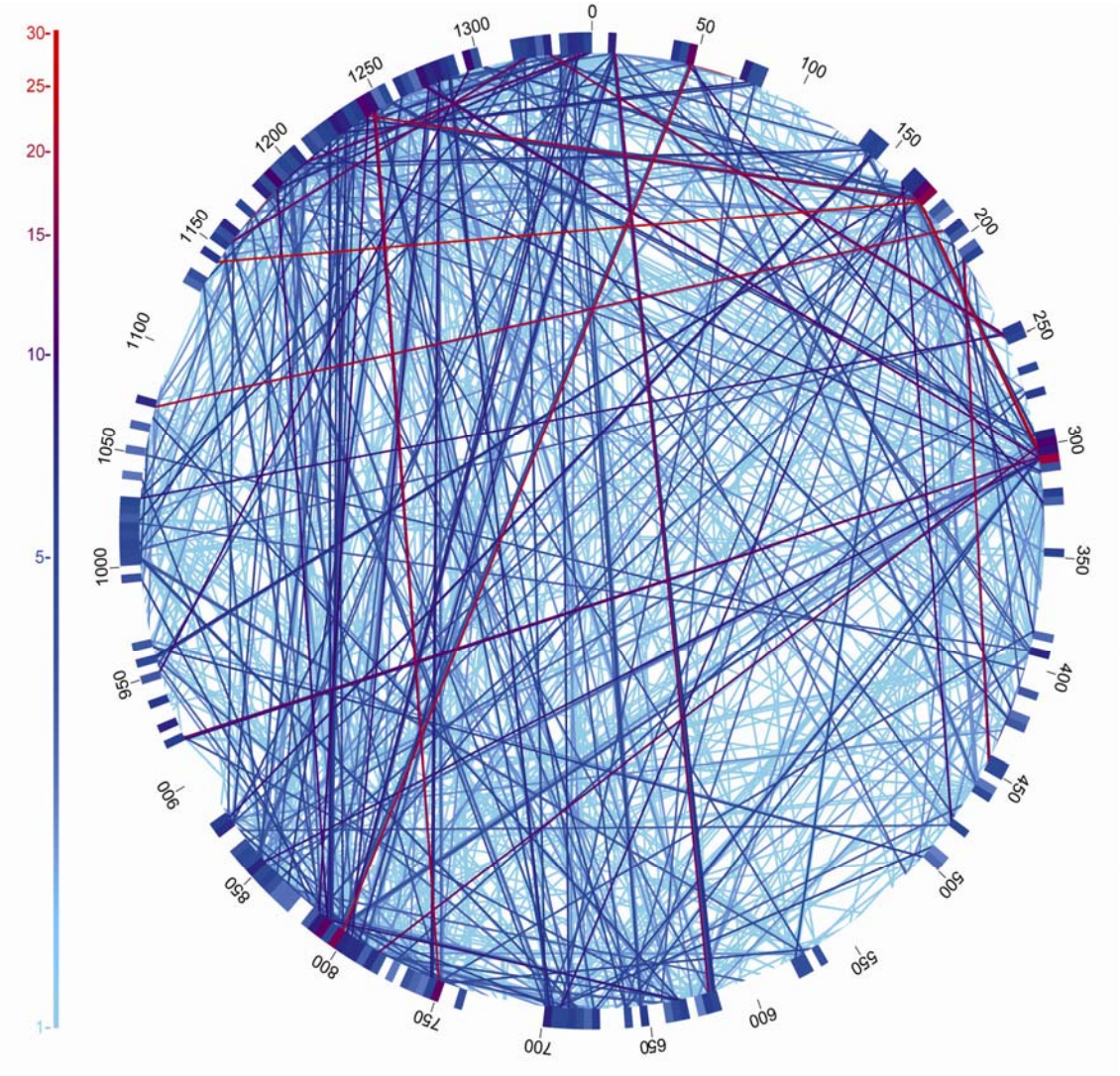
Figure 8. A) Syntig analysis performed on unassembled reads, and B) syntig analysis on sequence data containing assemblies as well as unassembled reads.



	database	sequences	hfws*	syntigs
A	unassembled reads	1.9×10^6	185,103	61,626
B	assemblies + singleton reads	811,372	111,332	74,005 (66% singletons)

* homologous fragments with synteny

Figure 9. Rearrangements in the order of SAR11 genes in the Sargasso Sea metagenome, relative to the HTCC1062 genome. The genome is represented by the outer circle. Internal lines (chords) indicate SAR11 gene rearrangements found on environmental sequence fragments. The number of occurrences of each gene rearrangement is indicated by the color scale.



no synteny requirement was imposed. The high incidence of synteny (Fig. 2C) can be reconciled with the many gene rearrangements shown in Fig. 9 by considering that a few very frequent rearrangements account for most of the cases of rearranged genes. 3,350 fragments had at least two genes that passed the reciprocal best BLAST test but displayed an alternate gene order. Of the 627 different genes found in non-syntenous positions in Fig. 9, seventeen can be tracked to genes that are not found on any syntigs (Table 3) These are likely to represent examples of gene re-arrangements that are conserved in the Sargasso Sea SAR11 population relative to the Oregon coast population.

Genome rearrangements were not random, but were concentrated at boundaries between operons. We compared the evidence for gene re-arrangements to the distribution of predicted operon boundaries [35]. The average number of rearrangements detected per gene boundary was 2.58, but the average at boundaries between operons was 3.21, and the average within operons was 1.94. An analysis of variance indicated that these differences are highly significant (Fig. 10). Perhaps not surprisingly, this finding suggests that selection allows rearrangements between operons more frequently than re-arrangements within operons.

Divergence among Sargasso Sea SAR11 populations. The high conservation of gene synteny is especially noteworthy considering that the amino acid sequence

identity between the syntig genes and the HTCC1062 genome ranges from high (98%) to low (30%) and averages only 71%. For comparison, the amino acid identity

Table 3. Genes from the coastal SAR11 strains that had homologs in the Sargasso Sea data, but were never found on syntigs.

Gene number	Gene product	Number of Homologs found	Occurrences in alternate gene order
SAR11_0079	hypothetical protein ytoQ	66	7
SAR11_0171	Rhodanese-related sulfurtransferase (ion tranport)	185	3
SAR11_0312 *	Unknown (potential Sulfotransferase domain)	343	19
SAR11_0393	SAM-dependent methyltransferase	117	42
SAR11_0461	Unknown	66	4
SAR11_0642	Trypsin-like serine protease	354	5
SAR11_0691	Unknown	22	4
SAR11_0796	aldehyde dehydrogenase	1200	16
SAR11_0815	Carbonic anhydrase	52	29
SAR11_0845	steroid monooxygenase (ion transport)	307	17
SAR11_0852	homoserine dehydrogenase	102	8
SAR11_0959	Unknown	21	5
SAR11_1071	GCN5-related N-acetyltransferase	18	5
SAR11_1144 *	cyclopropane-fatty-acyl-phospholipid synthase	754	48
SAR11_1227	ADP-ribosylglycohydrolase	80	29
SAR11_1248	Winged helix DNA-binding	160	3
SAR11_1347	Unknown	15	3

* found on just one syntig, or just a few syntigs with low scores

Figure 10. Statistical analysis of observed genome rearrangements.

Query								
Example	669	= non-allowed pairs						
	1332	= possible pairs						
	0.502252	= proportion of nonallowed pairs						
	0.497748	= proportion of allowed pairs						
	1708.27		theoretical mean number of allowable pairs					
	857.9826		theoretical variance					
	29.29134		theoretical standard deviation					
		1796.14			422.73			
		1620.4			14.4319			
Observation								
	3432	= total number of observations not sequencially ordered						
	1301	= total number of non-allowed pairs observed						
	2131	= total number of allowed rearranged pairs						
Maximum Likelyhood Estimator =		p(hat) =						
	0.620921	= sample proportion of allowed pairs				= 2131 / 3432		
Null Hypothesis: randomness alone is responsible for this observation.								

	P[X > 2131] = is less than the probability that a standard normal distribution is							
		more than 14 standard deviations above the mean.						
		< 10^(-86)						
Conclusion:	The small p-value provides strong evidence that the null hypothesis is incorrect.							
Conclusion:	The hypothesis that rearrangements are more likely							
	at boundaries between operons is conclusive.							
Observation:	Even if the true probability were 0.6 (or less)							
	the probability of observing 2131 allowed pairs							
	would be less than	0.0004						

between *E. coli* and *Salmonella* for GroEL is 98% and within *Burkholderia* species is 75%, whereas the lowest SAR11 GroEL syntigs have an amino acid identity of 77% (avg= 0.87). For RecA the numbers are 96% similarity for *E. coli* and *Salmonella*, 92% within *Burkholderia,* and 63% (avg= 0.81) for the lowest *Pelagibacter* syntigs. Hallam et.al observed a similar average amino acid identity (65%) among 4000 Sargasso Sea fragments related to *C. symbosium,* and a similar high degree of gene-order conservation as well [32]. Coleman et.al [36] observed a protein sequence identity of 80% among 1574 genes between two strains of the abundant marine cyanobacterium *Prochlorococcus* that are 99.2% similar at the 16S rRNA locus. Syntig analysis predicts a very similar average sequence identity in the *Prochlorococcus* metagenome (79%, Fig. 4B).

The ratio of non-synonymous to synonymous substitution rates for a selection of 19 genes from the syntig data ranges from 0.04 to 0.23, indicating purifying selection (Table 4). The implication of these observations is that the divergence of amino acid sequences in the Sargasso Sea SAR11 populations is largely neutral variation in proteins that serve important functions. The accumulation of neutral sequence variation likely explains why poor results are often obtained using traditional, DNA-based, methods to assemble fragments from large bacterioplankton populations.

Scrutinized in detail, the syntig plots reveal many conserved patterns in the SAR11 metagenome. For example, the syntig plots illustrate that the Sargasso Sea SAR11 genomes include proteorhodopsin genes similar to the proteorhodopsin gene found in

Table 4. Non-synonymous (Ka) and synonymous (Ks) substitution rates, and nucleotide diversity at synonymous (Pi(s)) and non-synonymous (Pi(a)) sites in HTCC1062 genes. Six highly conserved genes are listed first. The remaining thirteen were sampled from the 22 HTCC1062 genes missing from closely related strain HTCC1002 that show variable syntig coverage. Standard error is shown in parentheses.

Gene number	Gene product, function	n*	Ks	Ka	Ka/Ks	Pi(s)	Pi(a)
SAR11_0162	groEL, chaperonin	18	1.0378 (0.2577)	0.0524 (0.0284)	0.0504 (0.0268)	0.575	0.038
SAR11_0426	suv3, ATP dependent helicase	6	1.4618 (0.2414)	0.1686 (0.0415)	0.1153 (0.0155)	0.618	0.085
SAR11_0428	thlA, acetyl-coa transferase	6	0.6576 (0.6807)	0.0933 (0.0966)	0.1419 (0.1419)	0.617	0.284
SAR11_0641	recA, recombinase	15	0.8581 (0.2554)	0.0604 (0.0277)	0.0704 (0.0223)	0.424	0.036
SAR11_0906	dnaE, DNA polymerase	13	1.5300 (1.2792)	0.2160 (0.1809)	0.1412 (0.1198)	0.592	0.228
SAR11_1122	rpoC, RNA polymerase	5	1.2869 (0.8312)	0.0535 (0.0356)	0.0415 (0.0269)	0.625	0.070
SAR11_0078	Epimerase	38	0.6909 (0.3092)	0.1468 (0.0698)	0.2124 (0.0979)	0.525	0.131
SAR11_0267	CHO transport	14	0.9431 (0.2749)	0.0992 (0.0323)	0.1052 (0.0204)	0.584	0.088
SAR11_0268	CHO transport	17	0.9077 (0.4752)	0.1090 (0.0612)	0.1200 (0.0743)	0.545	0.113
SAR11_0273	CHO transport	2	2.3162 (0.000)	0.2294 (0.000)	0.0991 (0.000)	0.365	0.110
SAR11_0274	CHO transport	3	0.9938 (0.8414)	0.1177 (0.1019)	0.1184 (0.0838)	0.415	0.124
SAR11_0655	AA transport	6	0.8693 (0.4377)	0.0370 (0.0163)	0.0425 (0.0274)	0.497	0.042
SAR11_0660	AA transport	8	0.9990 (0.4437)	0.0776 (0.0394)	0.0777 (0.0365)	0.513	0.076
SAR11_0677	Unknown	5	1.5142 (0.7423)	0.1599 (0.0940)	0.1056 (0.0477)	0.531	0.531
SAR11_0764	Conserved hypothetical	5	1.5142 (0.7423)	0.1599 (0.0940)	0.1056 (0.0477)	0.638	0.142
SAR11_1012	Glycosyl transferase	27	1.4290 (1.2799)	0.3259 (0.2998)	0.2280 (0.2110)	0.552	0.387
SAR11_1174	Phosphate regulation	29	1.2303 (0.5857)	0.2008 (0.1037)	0.1632 (0.0850)	0.598	0.180
SAR11_1175	Phosphate regulation	20	0.8141 (0.5575)	0.1070 (0.1140)	0.1314 (0.0950)	0.484	0.101
SAR11_1288	CHO metabolism	3	1.5345 (0.1053)	0.1548 (0.0377)	0.1009 (0.0190)	0.568	0.158

* number of sequences tested.

the HTCC1062 genome (Fig. 11). Proteorhodopsins are light-dependent proton pumps that are hypothesized to provide an alternative energy source for bacterioplankton cells [37], although their metabolic role remains uncertain [38]. Many of the genes surrounding the Sargasso Sea SAR11 proteorhodopsins are syntenic with the same regions of the HTCC1062 genome, but a nearby gene on the same strand, the MOSC binding protein (gene 629) is found consistently elsewhere in the SAR11 metagenome.

Hypervariable Regions. Interesting relationships between the HTCC1062 genome and the Sargasso Sea SAR11 populations emerge from the syntig plot (Fig. 2C). Although the distributions of homologous fragments and syntigs suggest a relatively conserved SAR11 “core” genome, they also reveal distinct hypervariable regions (Fig.2A,C). The hypervariable regions observed in the homologue-coverage plot (Fig 2A) are almost entirely devoid of syntigs. An additional hypervariable region (HVR4) that is not apparent in the homologue-coverage plot is evident in the syntig plot. Similar “islands” of genome variability have been found in many microbial genome comparisons [36,39] . These islands have been shown to include genes potentially involved in pathogenicity [40] and lipopolysaccharide (LPS)- associated variability [41]. Most evidence for large microbial pan-genomes comes from variable genomic islands.

Both gene order (synteny) and sequence similarity drop dramatically in the SAR11 hypervariable regions, causing them to stand out prominently as gaps in the plots of homologues and syntigs (Fig.2 A,C, and Figs. 12 – 14). The largest of these, HVR2,

Figure 11. Homologous fragments with synteny in region of proteorhodopsin gene; only fragments containing 3 or more genes are shown. (SMRP) small multi-drug resistance protein, (ACAS) acyl-coenzyme A synthetase, (FD) ferrodoxin, (TD) thioredoxin disulfide reductase, (GST) glutathione S-transferase, (DKS) DnaK suppressor protein.

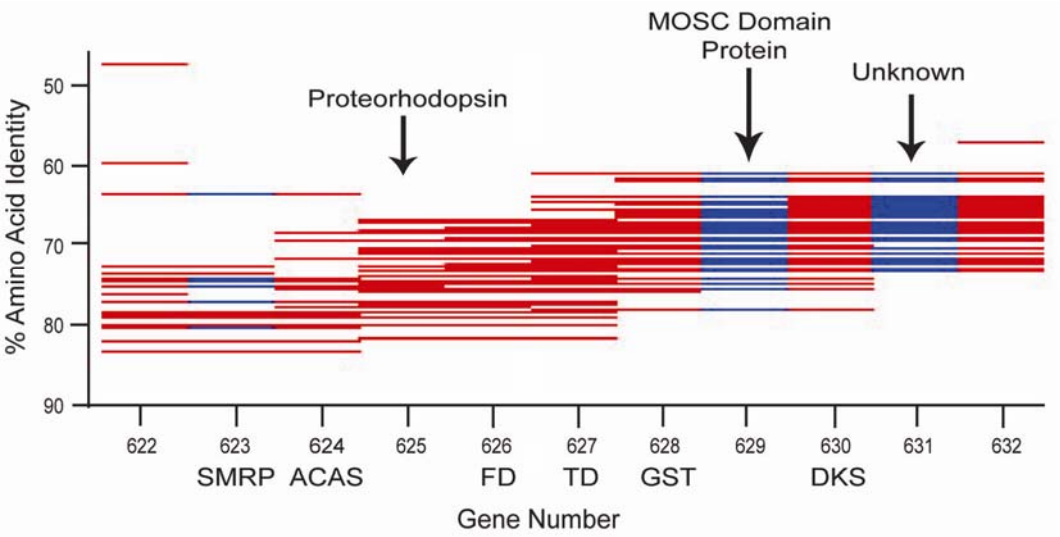


Figure 12. Enlargement of HVR2.

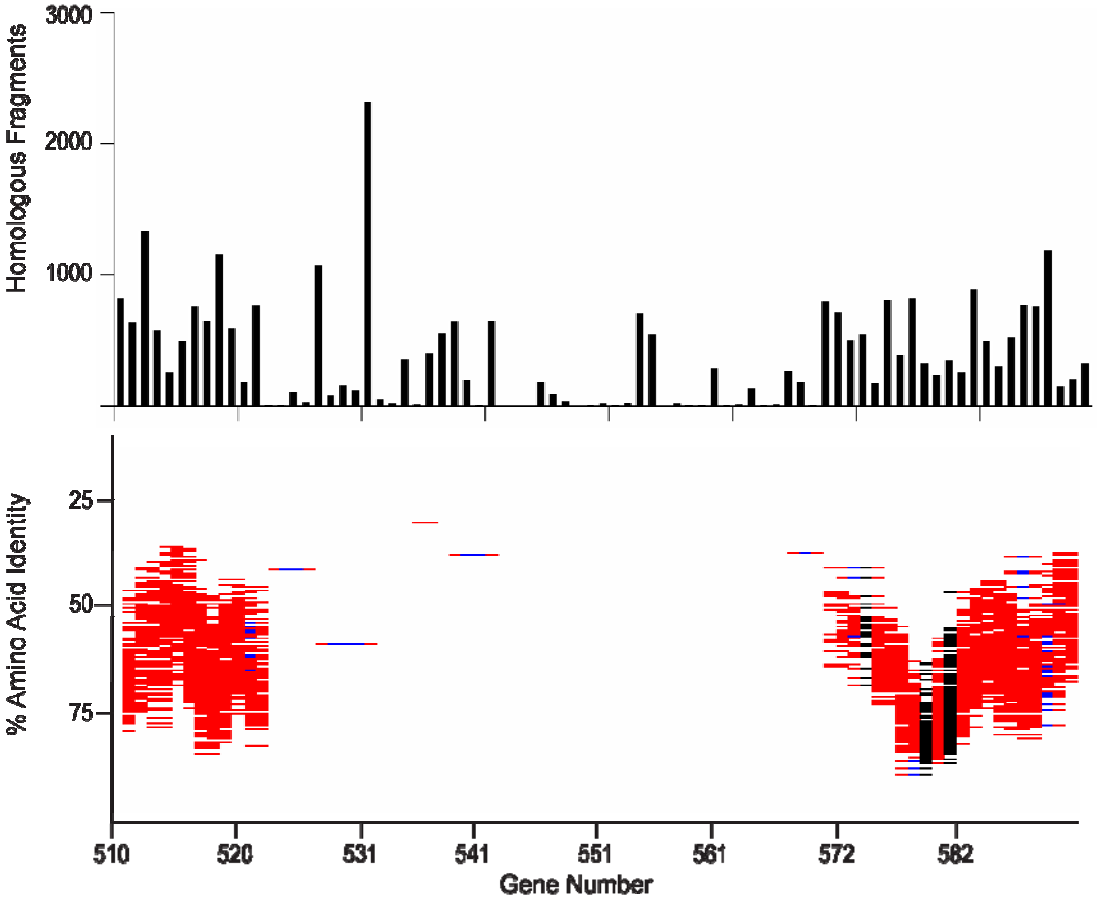
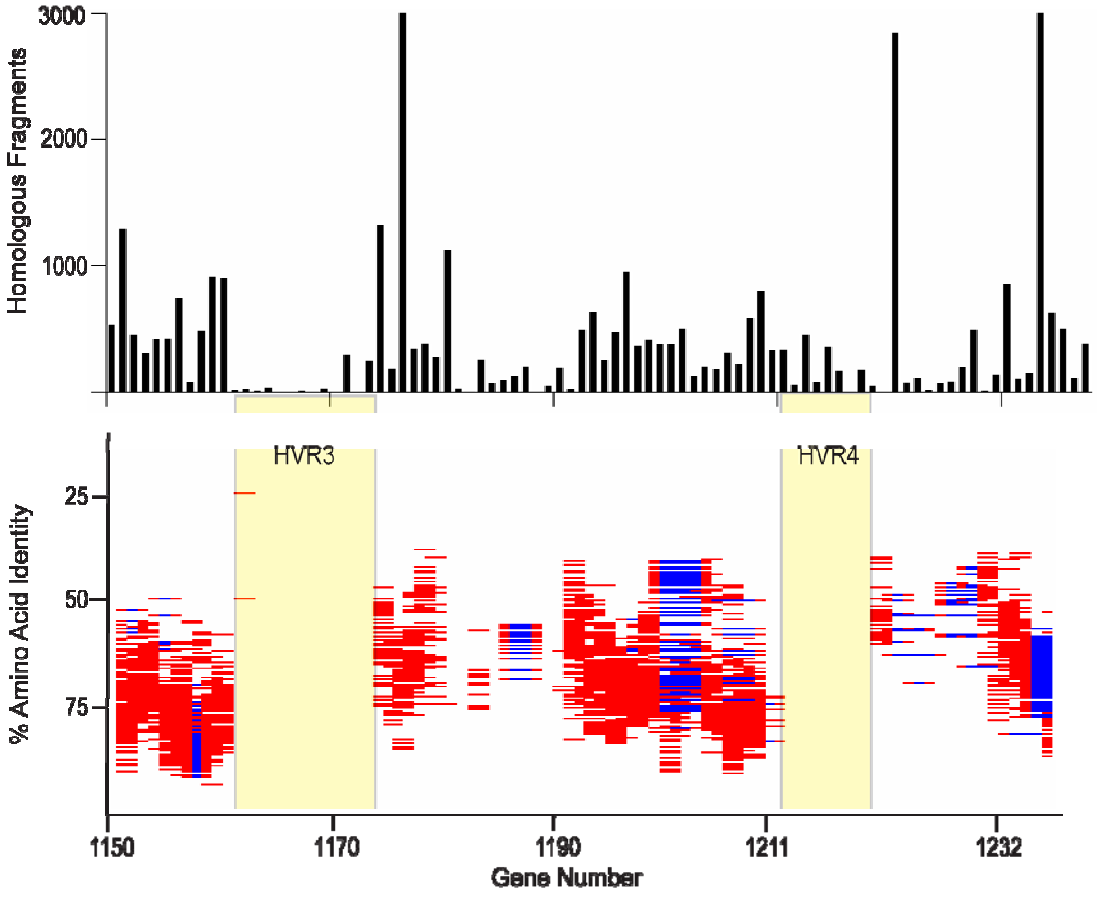
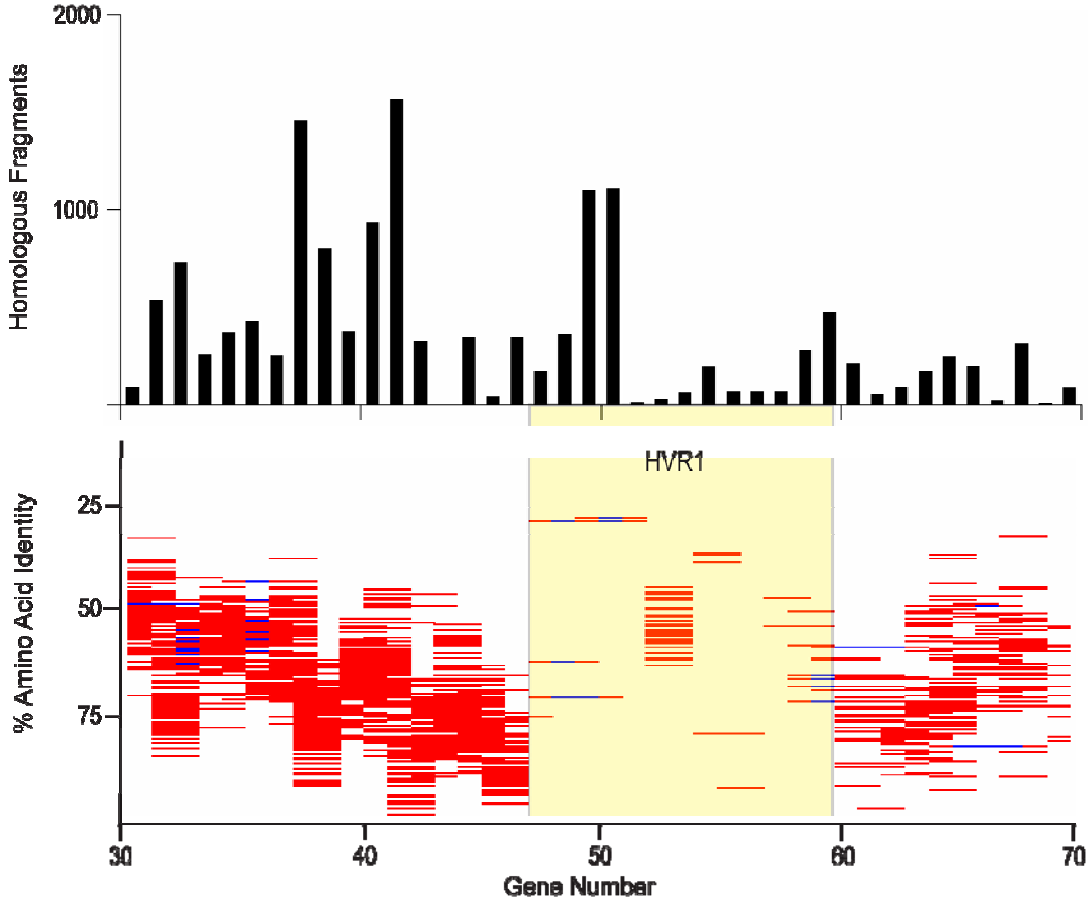


Figure 13. Enlargement of HVR3 and HVR4.



SI Figure 14. Enlargement of HVR1.



is a 48 kb region flanked by the sole 5S and 23S genes in the HTCC1062 genome. HVR2 mainly contains proposed lipopolysaccharide (LPS) biosynthesis genes (Table 5), and appears to be analogous to previously observed regions encoding cell surface properties. Based on current annotations, all but one of the enzymes involved in the biosynthetic pathways for the inner and outer core of lipopolysaccharide (LPS) are present in HVR2, while the enzymes involved in synthesis of the unexposed regions of the LPS are found elsewhere in the genome (Table 6). Brisk evolutionary change in the *Prochlorococcus* and SAR11 LPS cassettes may be related to selection pressure to avoid viral predation, which is a major source of mortality in oceanic bacterioplankton populations [42]. It has been postulated that similar variability in the *C. jejuni* LPS cassette is an adaptive response to selection pressure to evade host acquired immune responses [43]. The remaining SAR11 HVR's appear to be related to transport and secretion (HVR1&4) or unknowns (HVR3) (Tables 7 - 10). The transport and secretion functionality associated with HVRs 1 and 4 is consistent with the assertion of Coleman [36] that these islands may play a role in niche adaptation by differential nutrient acquisition capabilities.

Site-specific recombination mediated by integrases has been shown to cause rapid change in some islands of genomic variability [39]. However, SAR11, as with most examples thus analyzed [44], failed to display clear signatures of the integron model - attC sites were not found associated with any of the HVRs. HVR2 includes an integrase gene, and HVR4 is flanked by tRNA genes, which have been shown to serve

Table 5. Gene content of HVR2 (523–568), by functional category.

Functional Category	# genes
Cell envelope biogenesis, outer membrane	23
Amino Acid, Carbohydrate, or Nucleotide Transport	7
Defense Mechanisms	2
Export of O-antigen and teichoic acid	2
Signal Transduction	1
Unknown	3
Other	12

Table 6. HVR2 Genes. Bold font indicates genes involved in inner and outer core lipopolysaccharide (LPS) biosynthesis. Genes involved in the biosynthesis of the unexposed (lipid-A portion) of LPS are not present in HVR2.

Gene number	Gene product	Functional category	TM helices	Non-syntenous hits
SAR11_0524	Unknown	Unknown		0
SAR11_0525	nucleotide sugar epimerase	CHO metabolism		1
SAR11_0526	nucleotide sugar epimerase	CHO metabolism		3
SAR11_0527	acetolactate synthase	AA metabolism		5
SAR11_0528	methyl transferase	AA metabolism		22
SAR11_0529	Alcohol dehydrogenase	CHO metabolism		11
SAR11_0530	phospholipid synthase	OM		6
SAR11_0531	Short-chain dehyrogenase	CHO metabolism		4
SAR11_0532	Oxidoreductase	Unknown specificity	1	0
SAR11_0533	methyltransferase	Unknown specificity		2
SAR11_0534	Aminotransferase	Unknown	2	3
SAR11_0535	Unknown	Unknown	1	0
SAR11_0536	tktC	pentose-phosphate		5
SAR11_0537	tktN	pentose-phosphate		11
SAR11_0538	nucleotide sugar epimerase	CHO metabolism		34
SAR11_0539	Carbohydrate kinase	CHO metabolism		47
SAR11_0540	glycosyl transferase	CHO metabolism	2	2
SAR11_0541	nucleotide sugar dehydratase	NT-CHO metabolism		11
SAR11_0542	Unknown	Unknown	9	0
SAR11_0543	Unknown	Unknown	3	0
SAR11_0544	Unknown	Unknown	4	0
SAR11_0545	Phosphoheptose isomerase	CHO metabolism		10
SAR11_0546	sugar phoshatase	OM		4
SAR11_0547	amino-acid oxidase	AA metabolism		15
SAR11_0548	Unknown	Unknown	10	0
SAR11_0549	Unknown	Unknown	10	0
SAR11_0550	glycosyl transferase	OM		9
SAR11_0551	probable phage integrase	Unknown		3
SAR11_0552	glycosyl tranferase	OM		2
SAR11_0553	Amino transferase	OM		5
SAR11_0554	Carbamoyl transferase	Antibiotic synthesis		57
SAR11_0555	phospholipid synthase	OM		0
SAR11_0556	glycosyl tranferase	OM		2
SAR11_0557	glycosyl tranferase	OM	3	2
SAR11_0558	phospholipid synthase	OM		0
SAR11_0559	trehalose phosphate synthase	OM		0
SAR11_0560	methyl transferase	Unknown		0

Table 7. HVR gene content summary

HVR	Dominant Category	Genes	TM Spanning
1	Transport / Secretion	13	10
2	Outer Membrane / CHO Metabolism	44	12
3	Unknowns	12	2
4	Transport / Secretion	8	1

Table 8. HVR1 Genes

Gene number	Gene product	Functional category	TM helices	Non-syntenous hits
SAR11_0042	Autotransporter	Secretion	1	14
SAR11_0043	Unknown	Unknown		0
SAR11_0044	Autotransporter	Secretion	1	10
SAR11_0045	Unknown	Unknown	1	17
SAR11_0046	Autotransporter	Secretion	1	25
SAR11_0047	LexA / Cap	Transcription Regulation		7
SAR11_0048	Sodium Symporter	Ion Transport	8	218
SAR11_0049	Ammonium Transporter	Ion Transport	11	3
SAR11_0050	Ammonium Transporter	Ion Transport	11	3
SAR11_0051	Metal Ion Tranporter	Ion Transport	1	0
SAR11_0052	Unknown	Unknown		10
SAR11_0053	Putative pseudo-pilin PulG	Pilin / Secretion	1	28
SAR11_0054	pilin	Pilin	1	0

Table 9. HVR3 Genes

Gene number	Gene product	Functional category	TM helices	Non-syntenous hits
SAR11_1162	Unknown	Unknown		0
SAR11_1163	Unknown	Unknown		3
SAR11_1164	Unknown	Unknown	1	24
SAR11_1165	Unknown	Unknown		0
SAR11_1166	Unknown	Unknown		0
SAR11_1167	Unknown	Unknown		2
SAR11_1168	Unknown	Unknown		0
SAR11_1169	Unknown	RecB-like		17
SAR11_1170	Unknown	Unknown	2	0
SAR11_1171	Oxidoreductase	C metabolism		27
SAR11_1172	OsmC – like	Osmotically induced		1
SAR11_1173	Betaine-homocysteine methyltransferase	AA metabolism		103

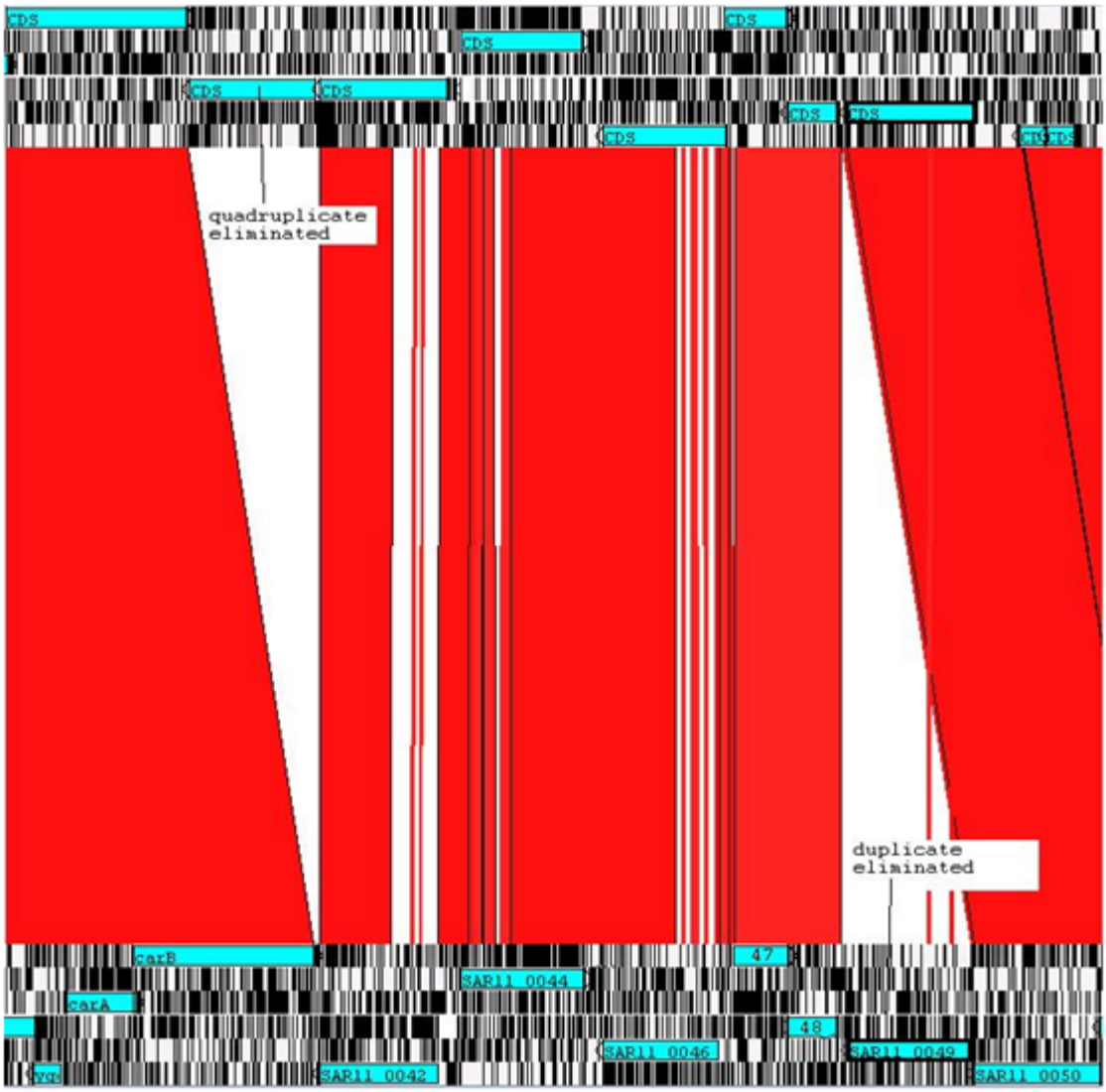
Table 10. HVR4 Genes

Gene number	Gene product	Functional category	TM helices	Non-syntenous hits
SAR11_1214	Possible Type III Secretion	Secretion		28
SAR11_1215	Sulfotransferase	Unknown		12
SAR11_1216	Pilin Precursor	Pilin		1
SAR11_1217	Bacterial-like Globin	Unknown		0
SAR11_1218	Phosphatase	sigB regulator		0
SAR11_1219	Pilin	Pilin		0
SAR11_1220	Unknown	Unknown	1	0
SAR11_1221	Sarcosine dehydrogenase	AA Metabolism		103

as site-specific recombination sites for temperate phages and transmissible plasmids [45].

Comparison of the genomes of strains HTCC1062 and HTCC1002 provides evidence that the 23S and 5S rRNA genes flanking HVR2 are sites of homologous recombination that allow novel variations of the LPS region to spread rapidly within populations. The HVR2 regions in these two genomes are 99.96% similar in nucleotide sequence, compared to 97.4% similarity for the genomes overall. In addition to a few point mutations, the two HVR2 sequences differ by a deletion of 13 nucleotides that removes one from a set of four tandem repeats within an ORFan gene. But, the HVR1 regions of HTCC1062 and HTCC1002 reveal the loss and gain of divergent, tandem duplicated genes. One gene is deleted from a set of four tandem, divergent gene duplications of Type V secretion proteins in strain HTCC1062 (Fig. 15). In strain HTCC1002, a single ammonium transporter gene is deleted from two, tandem duplicated genes. A high proportion of novel (ORFan) genes, such as we found in the SAR11 hypervariable regions, is a general feature of genome islands. While phages are suspected to be the reservoir for this novel gene pool [44], direct evidence for this hypothesis remains elusive. Daubin and Ochman reported that *E. coli* ORFan genes are short, AT rich, and most likely originate from phage [46]. The evidence for rapid gene evolution by duplication, divergence, domain rearrangements and deletion observed in HVR1 could also be explained by the alternative hypothesis that genetic processes intrinsic to the cell cause at least some of the rapid change in the SAR11 HVRs.

Figure 15. Deletion of duplicate genes in the HVR1 region of strains HTCC1002 and HTCC1062. Strain HTCC1002 appears at the top of the display and HTCC1062 at the bottom. One of four homologous Type V Autotransporters is deleted in HTCC1062 relative to HTCC1002, and one of two homologous ammonium transporters is deleted in HTCC1002 relative to HTCC1062.



Further information about SAR11 genome evolution comes from comparison of the genomes of HTCC1062 and HTCC1002, which were isolated from the same seawater sample [27]. The 16S rRNAs of these strains differ by one nucleotide, and in protein coding regions they are 97.4% similar in nucleotide sequence. The genome of HTCC1002 is 12,298 nucleotides larger than the genome of HTCC1062. Most of the length difference is due to 31 genes inserted in HVR3 of HTCC1002, supporting the conclusion that this hypervariable region is a hotspot for the acquisition of foreign DNA by horizontal gene transfer (HGT).

Genes conserved in the coastal isolates but not found in the Sargasso Sea SAR11 populations. Our analysis included an average of 118 fragments that covered each end of each gene and provided evidence for the identity of the adjacent gene. The number of Sargasso Sea fragments in SAR11 syntig plots declines sharply at sequence similarities above 90%, indicating that the coastal isolates, HTCC1002 and HTCC1062, have genetically diverged from their counterparts in the Sargasso Sea. However, there are no genes unique to the coastal genomes that suggest significant physiological differences between the coastal strains and Sargasso Sea SAR11 populations. Only 19 genes from the HTCC1062 genome are not represented in the SAR11 metagenome (syntigs or fragments passing a reciprocal best BLAST hit test). Of these, nine are from the hypervariable regions (Table 11), and six are classified as ORFans (returning no hits to NCBI databases with expect score less than $1 \times e^{-10}$).

Two members of this group (ORFs 542 and 555) are suspected to be involved in outer membrane biosynthesis and ORF 1217 is a bacterial-like globin (Table 11). With

Table 11. Evidence for genes specific to the coastal variant of SAR11. This list includes all genes for which no homologs were found passing the expect score criteria of less than or equal to $1 \times e^{-10}$. There were no examples of genes with known significant physiological functions in this category. The best tblastn expect score against the Sargasso Sea dataset (SSD) and the NCBI non-redundant nucleotide database (NCBI), and the hypervariable region (HVR) in which the gene is found are listed.

Gene number	Gene Description	SSD	NCBI	HVR
SAR11_0043	Unknown	4	None	1
SAR11_0163	Unknown	$1e^{-10}$	1.6	
SAR11_0414	Unknown	$1e^{-4}$	None	
SAR11_0471	Unknown	$3e^{-9}$	2.1	
SAR11_0542	Unknown Memb - biogenesis of cell wall	$3e^{-10}$	5e-11	2
SAR11_0544	Unknown Memb	$3e^{-6}$	3e-4	2
SAR11_0548	Unknown Memb	$8e^{-9}$	6e-5	2
SAR11_0555	CMAS Family - biogenesis of cell wall	$3e^{-8}$	6e-4	2
SAR11_0631	Unknown	1.1	None	
SAR11_0788	Unknown Memb	$9e^{-5}$	None	
SAR11_0875	Unknown	$1e^{-10}$	None	
SAR11_0930	Unknown Memb	0.47	1.2	
SAR11_0989	Unknown	0.61	None	
SAR11_1165	Unknown, possible exonuclease	$4e^{-7}$	1.3	3
SAR11_1170	Unknown	1.8	2.8	3
SAR11_1182	Unknown	$3e^{-10}$	2.2	
SAR11_1217	Bacterial-like globin	$4e^{-10}$	2e-8	4
SAR11_1220	Unknown	$2e^{-4}$	6.2	4
SAR11_1249	Unknown	$2e^{-6}$	3.7	

only one exception, we found closer homologues in the Sargasso Sea dataset than in NCBI databases.

Conclusions

The very high coverage of SAR11 genomes in the Sargasso Sea metagenome database allowed us to ask the question, what properties of the SAR11 genome are conserved between coastal Oregon SAR11 strains and SAR11 populations from an oligotrophic gyre? The amino acid sequence divergence between these populations exceeds the divergence between some microbial genera, suggesting that genomic properties have had ample time to diverge in response to selection.

Previous reports have shown that the conservation of gene order between prokaryotic genomes dissipates faster than protein sequence identity or gene complement [47,48]. Synteny is regarded as a rapidly evolving property of genomes, second only to DNA with regulatory functions [49]. Huynen and coworkers compared orthologs from an evolutionarily diverse set of 9 genomes to show that gene order becomes nearly random before protein identity decays below 50% [49]. We are not aware of reports comparable to ours that show the conservation of synteny within and between populations. We speculate that the seemingly high conservation of synteny observed in the SAR11 populations may be an example of selection acting to preserve local gene order.

The new analytical approaches we describe here reveal elements of conserved gene order, and genes that are inserted or deleted relative to a reference genome. Perhaps more importantly, the plotted data graphically convey some of the complexity of genome evolution. These approaches are robust for some conclusions; for example, the identification of genomic regions that are missing or highly diverged from the query sequence, and regions, such as the proteorhodopsin gene locus, where genes and gene order vary in conserved patterns. There are also caveats. For example, it is likely that only the termini of large cassettes of inserted genes in the target genomes can be observed, and then, only where they abut regions of conserved gene order. The absence of any observations suggesting conserved insertions of novel genes in the metagenomic data suggests that the Sargasso Sea SAR11 variants are very similar to their coastal counterparts, aside from the hypervariable regions observed in the syntig plots.

Our observations are consistent with the interpretation that natural selection has concentrated genes that encode cell surface properties into HVR2, and that this region is subject to unusually rapid rates of sequence divergence and re-arrangements of gene order. Viral predation on microbial cells is intense in the ocean water column and is likely to provide a keen source of selective pressure that favors microbial populations with diverse, rapidly evolving surface properties. An analogous variable genome region containing genes for cell surface components (LPS cassettes) has been

observed in *Prochlorococcus sp.* [41], and is evident in our syntig plot for this organism (arrow in SI Fig 3B). We propose that the structural RNA genes flanking the LPS cassettes provide zones of conserved DNA sequence that promote horizontal exchange of the cassettes by homologous recombination. Multi-locus sequence typing has shown that rates of intraspecific recombination are high within the coastal SAR11 population [50]. Alternatively this variable genome region could be explained by horizontal gene transfer from another species, a hypothesis that is consistent with the observation that the AT content of HVR2 is anomalously high (73%).

Our findings indicate that SAR11 genomes from different oceanic provinces share many conserved features despite dynamic processes of genome change that are at work in nature. The Sargasso Sea SAR11 populations are conserved in local gene order, and gene complement, with respect to populations that live in richer, colder coastal water, but diverge dramatically in amino acid sequence similarity. A broad implication is that large microbial populations such as bacterioplankton accumulate high diversity in some genome properties, while remaining constrained in others [23]. Protein evolution provides an analogy. Protein families can encompass wide variation in amino acid sequences while retaining the key elements of three-dimensional structure that confer function [51]. Similarly, in old clades that comprise large populations, microbial genomes may wander over sequence space, giving an illusion of variability, while remaining highly constrained in features that govern cellular structure and function.

Materials and Methods

The process used to identify SAR11 fragments in metagenomic data is illustrated in Fig. 1. Metagenomic fragments carrying genes with high similarity to HTCC1062 genes at the protein level were identified using tblastn [52] with an expect score cutoff of $1 \times e^{-10}$ (Fig. 1A). From this set of fragments, a subset of fragments was identified that shared gene order with the HTCC1062 genome (Fig. 1B). Finally, only those fragments carrying exclusively genes that returned the appropriate HTCC1062 gene as the best hit in queries to the NCBI non-redundant proteins database (nr) were retained (Fig. 1C). We refer to these fragments as 'syntigs'. A second set of fragments that pass the reciprocal best-hit test (Fig. 1A,C), but fail the synteny requirement (non-syntigs) were used to explore genome rearrangements, and to search for genes present in the Sargasso Sea SAR11 metagenomic DNA that have no homologues in the SAR11 HTCC1062 genome.

We use the term "fragment" to refer to DNA sequences from the Sargasso Sea environmental data set, whether they are single reads or contigs assembled from multiple reads. Although some of the contigs may be mis-assemblies, the inferences made about SAR11 genomes are based on multiple fragments carrying SAR11 homologues, as described below, in known SAR11 gene order, and therefore are unlikely to be impacted by mis-assemblies (Fig. 8).

Homologue search. Fragments carrying genes with high similarity to HTCC1062

genes are identified at the protein level with tblastn [52], using the amino acid sequence as input, a $1 \times e^{-10}$ expect score cutoff, and complexity filtering off. The results are limited to the first 3000 hits. Command line: blastall –i sar11_proteins.fa –d venter_nt -p tblastn –e '$1e^{-10}$' –F F –v 3000 –b 3000, where sar11_proteins.fa is a fasta file of HTCC1062 proteins, venter_nt is the Sargasso Sea fragment data in blast format. Default values were used for all unnamed parameters, blastall version was 2.2.12. For convenience, the set of fragments identified in this fashion are hereafter referred to as "homologous fragments".

Syntig detection. A subset of the homologous fragments that shared synteny with the HTCC1062 genome was identified by finding fragments that were common to the lists from adjacent HTCC1062 genes, and verifying that the genes are arranged in tandem on the fragment. Each gene on these syntenous fragments was subjected to the reciprocal best-hit test (Fig. 1C). The fragment nucleotide sequence of the high-scoring sequence pair (HSP) from the tblastn search for homologues was searched against the NCBI non-redundant proteins database using blastx. The accession number of the best hit was compared to the accession number of the predicted HTCC1062 gene to confirm the identity of the reciprocal best hit. The term "syntig" thus designates fragments containing at least two best-hitting HTCC1062 genes in the proper order, but does not itself indicate anything about the rest of the fragment, for instance the presence of genes without homologues in HTCC1062. To visualize the data, syntig plots were constructed where the fragment is plotted horizontally by gene

position and vertically by the average amino-acid identity of all genes present on the fragment (Fig. 1D).

Testing the syntig concept with different query genomes. To assess the selectivity of syntigs we studied a set of organisms of varying relevance to ocean surface ecology. Of these, *Prochlorococcus marinus* MED4 provides an example from a clade that is relatively abundant in the Sargasso Sea but forms a shallow cluster by 16S rRNA gene sequence analysis [53]. *Escherichia coli* was chosen as an organism that is unlikely to appear often in the Sargasso Sea. Five additional cultured marine strains were also used as query genomes. They are listed in Table 2.

Fragments with SAR11 ribosomal RNA genes. Only two fragments containing the HTCC1062 16S rRNA gene are found among the 349,742 homologous fragments. The HTCC1062 5S rRNA gene is found on 36 fragments. Nineteen of these fragments carry homologues to HTCC1062 ORF 570 upstream of the 5S rRNA gene, as found in the HTCC1062 genome, indicating the presence of a split ribosomal operon in the metagenomic data.

Genome Rearrangements. To find fragments containing HTCC1062 genes in altered gene orders, the genes on the homologous fragments that did not show synteny to HTCC1062 were subjected to the reciprocal blast analysis described above for the syntigs. The number of occurrences of a given best-hitting HTCC1062 gene adjacent

to a non-syntenous best-hitting HTCC1062 gene was determined, scoring the number of times any unique pair of genes occurred together. To visualize the data, a circular plot was developed to represent the relative occurrence of non-syntenous gene-gene pairs (Fig. 3). The outer circle represents the genome of HTCC1062 and the internal lines connect non-syntenous boundaries. The color bar indicates relative occurrence of the linkage, with those occurring most frequently shaded red. To determine if genome rearrangements were concentrated at operon boundaries a statistical analysis was performed comparing the number of gene pairs found that violate an operon boundary (disallowed pairs - Fig. 10), to the number found that preserve operon boundaries (allowed pairs – Fig. 10).

Genes conserved in the coastal isolates but not found in the Sargasso Sea SAR11 populations. HTCC1062 genes with no homologues in the Sargasso Sea dataset are considered likely candidates for genes specific to the coastal variant. To determine the extent to which these genes are represented both in the Sargasso Sea dataset and the NCBI non-redundant database, their translation products were queried against both databases with tblastn using an expect score cutoff of 10. The highest scores found for these 19 genes are listed in SI Table 3. Three genes (414, 788 and 875) failed to hit the NCBI databases with an expect score of less than $1 \times e^{-4}$ and are classified as ORFans in the HTCC1062 genome. It is likely that these genes are conserved hypotheticals within the SAR11 group.

Searching for genes conserved in Sargasso Sea SAR11 populations and not found in the genomes of the coastal isolates. Non-homologous genes found alongside genes that had best hits to HTCC1062 genes were regarded as candidates for genes specific to the Sargasso Sea SAR11 populations. The amino acid sequence of each ORF (the determination of open reading frames on the environmental fragments was taken from the conserved domain feature tags of the NCBI GenBank record) on every homologous fragment with sufficient length beyond that accounted for by HTCC1062 homologues was used as a blast query sequence against the NCBI non-redundant proteins database (BLASTP, expect score cutoff $1 \times e^{-6}$) and the NCBI Conserved Domain Database (CDD) [54]. We examined the data for the frequency of any specific non-HTCC1062 gene occurring next to a given HTCC1062 gene, using the gene descriptions from the NCBI database as well as the protein family identifier from the CDD as search strings for the identification of common genes.

Calculating a synteny index. We define a synteny index to be the fraction of best-hitting HTCC1062 homologues found adjacent to a best-hitting and syntenous HTCC1062 homologue. The amino acid sequence of each ORF on all homologous fragments carrying at least two genes (238,663 of 349,742 total, Table 1) was queried against the NCBI non-redundant proteins database. For every best-hitting gene on the fragment, if at least one best-hitting neighbor was present, it was counted as a syntenous observance. To calculate the synteny index the total of all syntenous observances is divided by the total observances of best hitting genes, syntenous and

non-syntenous.

Tests of selection. Sequences were analyzed for synonymous and non-synonymous substitution rates using the software program SWAAP 1.0.2 [55], set to the Li method (1993) with a window size of 90 and step size of 18. The values reported in Table 2 were created from alignments that include those portions of the HTCC1062 gene and syntig sequence defined by the HSP start and end positions taken from the tblastn results (see homologue detection step). Translated sequence was used to guide the alignments when necessary. Nucleotide divergence values were calculated with the software program DnaSP 4.0 using the 'synonymous non-synonymous substitution' option under the analysis menu with default parameters [56].

Accession Numbers of strains used in this study. HTCC1062: NC_007205 HTCC1002: AAPV00000000.

Acknowledgements

This work was supported by research grant MCB-0237713 from the NSF Microbial Observatories Program and a grant from the Gordon and Betty Moore Foundation. The authors wish to extend their thanks to Saul Kravits, Steve Ferriera, Justin Johnson, Robert Friedman, Yu-Hui Rogers, J. Craig Venter and their staff at the J. Craig Venter

Institute for genome sequencing of HTCC1002, and Kevin Vergin, Dee Denver, and Robert Burton for critical advice.

CHAPTER 3.
GENERAL CONCLUSION

The methodology described here of binning environmental DNA fragments based on the conservation of gene order and protein-coding capacity with a query organism shows promise as a tool for metagenomic investigations. Modern genome-based approaches to microbial taxonomy require knowledge of the breadth of diversity of genomes that exists within a group that is otherwise ‘an irreducible cluster of organisms diagnosibly[sic] different from other such clusters and within which there is a parental pattern of ancestory and descent’ – which is a reasonable working definition of bacterial species as defined by Craycraft [57]. At this early stage of development in the field of metagenomics a variety of binning techniques needs to be evaluated. The technique described here may offer an advantage in ecosystems where members exhibit a large amount of neutral nucleotide sequence variation where assembly attempts at the nucleotide level fail. With organisms such as *Candidatus* Pelagibacter ubique that are abundant in the ecosystem under study, the basic technique of binning fragments reveals a great deal about gene order conservation, gene-content variability, and amino-acid level variability. Logical extensions of the method that fall naturally from the data, such as genome-rearrangements, add to the overall attractiveness of the method in measuring and displaying genome-wide variability. Creating a syntig plot of an organism from environmental data can be a first step towards collecting and organizing sequence data for subsequent studies of evolutionary parameters, as was done here for a small subset of genes (Table 4). A computational system to collect the

sequence data from the syntigs for each gene and produce alignment files requiring as little hand-curation as possible is a natural next step in the analysis of this data. Such data could be input to programs such as DnaSP [55] and plots of synonymous vs. non-synonymous substitution ratios could be automatically generated spanning the entire query genome. Such data would help delineate the core and pan genome. The sequence data from the syntigs can also be used to represent allowed sequence diversity on a micro-array chip for environmental studies of gene expression. These chips could then be used to measure the degree to which different variants of a given protein are expressed in the environment potentially providing clues to ecotype differentiation. As more large-scale environmental sequencing projects are completed and released, such as the Global Ocean Sampling Expedition of Venter *et.al.* [58] the syntig methodology can be applied to data sets that go far beyond the oligotrophic gyre of the Sargasso Sea. Conserverd properties of the coastal *P. ubique* genome that were measured here can be tested with these new data sets to assess how universal the findings are and how well the analysis of one genome predicts the properties of many. Analyses of SAR11 ITS sequence clustering and FISH studies suggest there are a handful of SAR11 ecotypes. The application of syntig methodology to the right combinations of query genomes (coastal Pacific strain as done here versus an Atlantic strain) and metagenomic data sets should reveal a genomic picture of these ecotypes that can hopefully be correlated to biogeography.

The syntig methodology is intended to be just one of many in an arsenal of techniques to explore genomic diversity. As described here the method does not account for

intergenic space and is of limited use in the analysis of highly variable regions of a genome. The entire process starts with a single query genome and is thus limited to detecting the variability found in the metagenome as it relates to the query genome. For example, if another SAR11 exists that is radically different from HTCC1062 or HTCC1002 in gene order, but would still be considered a member of the same species by all other measures, then the syntig process will underestimate the total variability that exists. The small insert size of a shotgun sequence library produces fragmentary data that challenges attempts to assign a given read to a taxonomic group. A mere two syntigs were found with a phylogenetic anchor attached. The syntig process is an attempt at circumventing this problem but therefore requires validation. In the thesis presented here the process was validated much as an analytical assay – for linearity, limit of detection, and selectivity. A better approach to validate the notion that a syntig arose from a member of the same group as the query genome would be one that makes use of the power of phylogenetic inference. Constructing a tree from the syntig sequence data for a given gene and demonstrating that all syntig sequences cluster with the query genome would be perhaps the most compelling argument for the validity of the syntigs. Unfortunately the fragmentary nature of the metagenomic data stymies these efforts. All good phylogenetic trees start with well-aligned full-length sequences, something that is just not available in the current data.

Bibliography

1. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, et al. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309: 1242-1245.
2. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66-74.
3. Koch R (1880) Investigations into the etiology of traumatic infectious diseases.
4. Staley JT, Konopka A (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39: 321-346.
5. Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59: 143-169.
6. Ferguson RL, Buckley EN, Palumbo AV (1984) Response of marine bacterioplankton to differential filtration and confinement. Appl Environ Microbiol 47: 49-55.
7. Jones JG (1977) The effect of environmental factors on estimated viable and total populations of planktonic bacteria in lakes and experimental enclosures. Freshwater Biology 7: 67-91.
8. Kogure K, Simidu U, Taga N (1979) A tentative direct microscopic method for counting living marine bacteria. Can J Microbiol 25: 415-420.
9. Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial ecology and evolution: a ribosomal RNA approach. Annu Rev Microbiol 40: 337-365.
10. Woese CR (1987) Bacterial evolution. Microbiol Rev 51: 221-271.
11. Rappe MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57: 369-394.
12. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5: R245-249.
13. Pace N, Stahl DA, Lane DJ, Olsen GJ (1985) Analyzing natural microbial populations by rRNA sequences. ASM News 51: 8.
14. Schmidt TM, DeLong EF, Pace NR (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol 173: 4371-4378.
15. Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, et al. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci U S A 89: 8794-8797.
16. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37-43.
17. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269: 496-512.

18. Fraser CM, Read TD, Nelson KE (2004) Microbial Genomes. Humana Press.
19. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.
20. Krieg NR (1988) Bacterial classification: an overview. Can J Microbiol 34: 536-540.
21. Staley JT (2004) Speciation and bacterial phylospecies. Microbial diversity and bioprospecting: 40-48.
22. Welch RA, Burland V, Plunkett G, 3rd, Redford P, Roesch P, et al. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A 99: 17020-17024.
23. Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, et al. (2005) Genotypic diversity within a natural coastal bacterioplankton population. Science 307: 1311-1313.
24. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 102: 13950-13955.
25. Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 102: 2567-2572.
26. Steinberg DK, Carlson CA, Bates NR, Johnson RH, Michaels AF, et al. (2001) Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biochemistry. Deep-Sea Research II 48: 1405-1447.
27. Rappé MS, Connon SA, Vergin KL, Giovannoni SJ (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418: 630-633.
28. Smith RL, Huyer A, Fleischbein J (2001) The coastal ocean off Oregon from 1961 to 2000: is there evidence of climate change or only of Los Niños? Progress in Oceanography 53: 369-387.
29. Brown MV, Schwalbach MS, Hewson I, Fuhrman JA (2005) Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. Environ Microbiol 7: 1466-1479.
30. Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57: 369-394.
31. Morris RM, Cho JC, Rappe MS, Vergin KL, Carlson CA, et al. (2005) Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic Time-series Study site. Limnol Oceanography 50: 1687-1696.
32. Hallam SJ, Konstantinidis KT, Putnam N, Schleper C, Watanabe Y, et al. (2006) Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum. Proc Natl Acad Sci U S A 103: 18296-18301.
33. Rasko DA, Myers GS, Ravel J (2005) Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics 6: 2.

34. DeLong EF (2005) Microbial community genomics in the ocean. Nat Rev Microbiol 3: 459-469.
35. Price MN, Huang KH, Alm EJ, Arkin AP (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res 33: 880-892.
36. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, et al. (2006) Genomic islands and the ecology and evolution of Prochlorococcus. Science 311: 1768-1770.
37. Béjà O, Aravind L, Koonin EV, Suzuki M, Hadd A, et al. (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289: 1902-1906.
38. Giovannoni SJ, Bibbs L, Cho JC, Stapels MD, Desiderio R, et al. (2005) Proteorhodopsin in the ubiquitous marine bacterium SAR11. Nature 438: 82-85.
39. Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol 2: 414-424.
40. Hacker J, Carniel E (2001) Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep 2: 376-381.
41. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, et al. (2003) Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature 424: 1042-1047.
42. Proctor LM, Fuhrman JA (1990) Viral mortality of marine bacteria and cyanobacteria. . Nature 343: 60-62 350
43. Linton D, Karlyshev AV, Wren BW (2001) Deciphering Campylobacter jejuni cell surface interactions from the genome sequence. Curr Opin Microbiol 4: 35-40.
44. Holmes AJ, Gillings MR, Nield BS, Mabbutt BC, Nevalainen KM, et al. (2003) The gene cassette metagenome is a basic resource for bacterial genome evolution. Environ Microbiol 5: 383-394.
45. Reiter WD, Palm P, Yeats S (1989) Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. Nucleic Acids Res 17: 1907-1914.
46. Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res 14: 1036-1042.
47. Tamames J (2001) Evolution of gene order conservation in prokaryotes. Genome Biol 2: 11.
48. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res 11: 356-372.
49. Huynen MA, Bork P (1998) Measuring genome evolution. Proc Natl Acad Sci U S A 95: 5849-5856.
50. Vergin KL, Tripp HJ, Wilhelm LJ, Denver DR, Rappe MS, et al. (In Review) High Intraspecific Recombination in a Native Population of
Candidatus Pelagibacter ubique (SAR11).

51. Bashford D, Chothia C, Lesk AM (1987) Determinants of a protein fold. Unique features of the globin amino acid sequences. J Mol Biol 196: 199-216.
52. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
53. Rocap G, Distel DL, Waterbury JB, Chisholm SW (2002) Resolution of Prochlorococcus and Synechococcus ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl Environ Microbiol 68: 1180-1191.
54. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, et al. (2005) CDD: a Conserved Domain Database for protein classification. pp. D192-196.
55. Pride DT (2005) A tool for analyzing substitutions and similarity in multiple alignments. Distributed by the author.
56. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.
57. Craycraft J (1989) Speciation and Ontology: the empirical consequences of alternate species concepts for understanding patterns and processes of differentiation. Speciation and its consequences: 28-59.
58. Rusch DB, *et.al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. In Press.