

Methodology article

Open Access

An application of statistics to comparative metagenomics

Beltran Rodriguez-Brito¹, Forest Rohwer^{2,3} and Robert A Edwards*^{1,2,3,4}

Address: ¹Computational Science Research Center, San Diego State University, San Diego, USA, ²Center for Microbial Sciences, San Diego State University, San Diego, USA, ³Department of Biology, San Diego State University, San Diego, USA and ⁴Fellowship for Interpretation of Genomes, Burr Ridge, USA

Email: Beltran Rodriguez-Brito - beltran.rodriguezbrito@gmail.com; Forest Rohwer - frohwer@sunstroke.sdsu.edu; Robert A Edwards* - redwards@salmonella.org

* Corresponding author

Published: 20 March 2006

Received: 24 August 2005

BMC Bioinformatics 2006, **7**:162 doi:10.1186/1471-2105-7-162

Accepted: 20 March 2006

This article is available from: <http://www.biomedcentral.com/1471-2105/7/162>

© 2006 Rodriguez-Brito et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Metagenomics, sequence analyses of genomic DNA isolated directly from the environments, can be used to identify organisms and model community dynamics of a particular ecosystem. Metagenomics also has the potential to identify significantly different metabolic potential in different environments.

Results: Here we use a statistical method to compare curated subsystems, to predict the physiology, metabolism, and ecology from metagenomes. This approach can be used to identify those subsystems that are significantly different between metagenome sequences. Subsystems that were overrepresented in the Sargasso Sea and Acid Mine Drainage metagenome when compared to non-redundant databases were identified.

Conclusion: The methodology described herein applies statistics to the comparisons of metabolic potential in metagenomes. This analysis reveals those subsystems that are more, or less, represented in the different environments that are compared. These differences in metabolic potential lead to several testable hypotheses about physiology and metabolism of microbes from these ecosystems.

Background

Metagenomics describes the functional and sequence-based analysis of DNA isolated from environmental sample without first culturing the associated microbes [1]. Four viral and four prokaryotic shotgun metagenome datasets have been published [2-8]. The acid mine drainage (AMD) metagenome data set was taken from a low complexity environment and includes slightly more than 10 Mb of sequence in 2,455 contiguous sequences (contigs) and ~8,000 predicted protein sequences [8]. The Sargasso Sea metagenome data set is from a more complex environment and includes 788 Mb of sequences in

809,112 contigs, and approximately a million predicted protein sequences [7].

Another conceptual "metagenomic" library can be constructed from the combined sequence data collective generated over the last 30-plus years of DNA and protein sequencing and deposited in the international databases. In contrast to the environmental metagenome libraries, the collective metagenome was built by the incremental addition of sequences from many different sources. The SEED database, developed by the Fellowship for Interpretation of Genomes (FIG), is an annotated non-redundant database, compiled from several sources including Gen-

Bank (including GenBank's non-redundant and refseq databases), Swiss-Prot, KEGG, and from genome sequencing centers. At the time of analysis, the SEED database contained 1,731,649 proteins (not including the environmental samples; Table 1). Although the sampling of sequences that constitute this library is not random, the SEED essentially represents essentially all known genomic complexity.

FIG pioneered the use of subsystems to annotate both complete and partial genome sequences [9]. Subsystems are biochemical pathway, fragments of pathways, clusters of genes that function together, or any group of genes that any annotator considers to be related. The subsystems are annotated across genomes by the annotators, providing the most reliable and consistent annotations within and between genomes. The subsystems-based annotations are ongoing and at a given point in time the subsystems represent the snapshot of the best available annotation of the SEED database.

Comparing metagenome samples could lead to the identification of signature functions associated with each metagenome sample, however this analysis requires reliable statistical techniques that are not only robust but are rapid to perform with hundreds of thousands or millions of data points per sample.

Here the Sargasso Sea and AMD metagenomes were compared with the SEED database to identify statistically significant differences in subsystems composition. We hypothesized that there were few barriers to the transfer of subsystems between environments and therefore certain subsystems were enriched by selection in those environments where they were important. We used a difference of medians analysis to identify those subsystems that have a statistically significant presence in each of the metagenomes. These analyses provide a framework for the statistical comparison of metagenomes.

Table 1: Number of genomes and protein encoding genes in the SEED database at the time of analysis. The two environmental samples are the Sargasso Sea and Acid Mine Drainage metagenomes.

Domain	Number of Genomes	Proteins	Percent of all proteins
Archaea	37	61,709	2
Bacteria	550	1,187,180	44
Eukarya	556	482,760	18
Environmental Samples	(2)	968,149	36
Total	1,145	2,699,798	100

Results

Determination of statistically significantly different subsystems

A difference between medians calculation was applied to rapidly identify statistically significant differences between metagenomes. This technique has several advantages over other possible statistical methods that could have been applied. For example, the difference between medians is extremely rapid for the calculation of differences between subsystems from different samples, and the method does not depend on the distribution of samples. The source code and step-by-step description of the method are provided as part of the supplemental material [see Additional Files 1 and 2].

Number of samples needed to identifying significant differences between metagenomes

Fig. 1 plots the sample size (S) of proteins that were sampled against the number of significantly different subsystems when the Sargasso data was compared to the SEED data at different confidence levels. At the most stringent level (99% confidence that the difference does not occur by chance) there are about 75 unique phylosubsystems in the Sargasso sample (solid red line) and about 150 unique phylosubsystems in the SEED. In either case at least the sample size needs to approach 300,000 proteins to achieve statistical significance. In contrast, at lower confidence levels, there are more significantly different phylosubsystems, and the samples need to be smaller to detect them. Thus with a 90% confidence that the difference do not occur by error there are approximately 100 unique subsystems in the Sargasso data, and a sample size of about 150,000 proteins is required to identify them all.

The AMD dataset has many fewer phylosubsystems than the SEED dataset, and only those subsystems present in the both samples are used in the comparison to identify significantly different phylosubsystems. This limits the analysis to 284 of the 523 different phylosubsystems. Of those that are present in the AMD and the SEED, only 19 are significantly more abundant in the AMD sample (Table 2) and 58 are more abundant in the SEED dataset. Saturation was reached at ~145,000 samples.

Subsystem differences between the SEED and Sargasso Sea metagenomes

Statistically different phylosubsystems between the SEED and Sargasso Sea metagenomes are shown in Fig. 2A. The effect of sample size is apparent. For example, statistically differences between RNA metabolism, oxidative phosphorylation, and membrane transporters can be detected at small sample sizes (e.g., 100,000 proteins). However, the differences between carbohydrate and amino acid metabolism, as well as most cofactors, vitamins, and pigments are not statistically different until >150,000 pro-

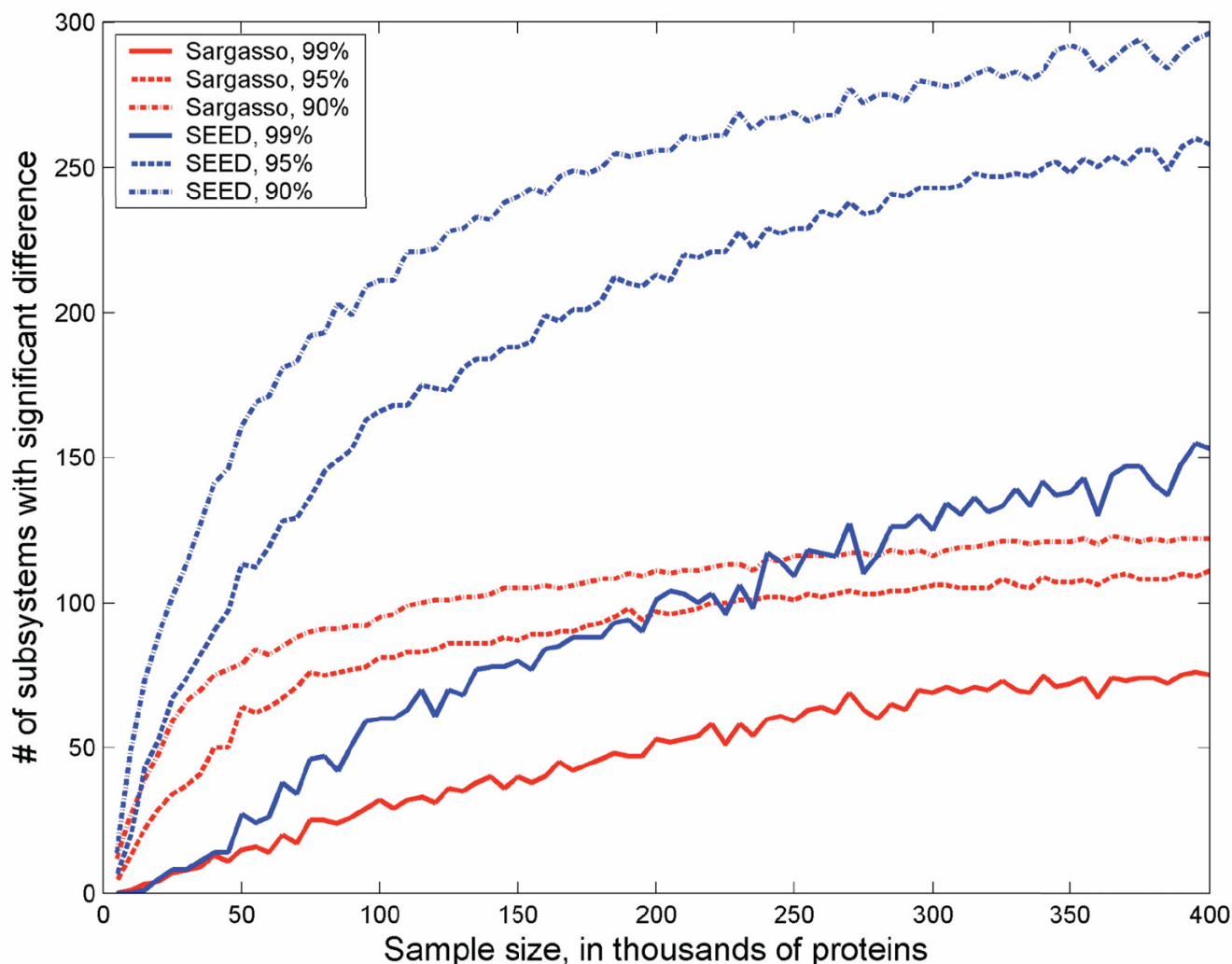


Figure 1
Effect of sample size on identifying differences between phylosubsystems. The red lines reflect the number of phylosubsystems overrepresented in the Sargasso Sea dataset. The blue lines represented the number of phylosubsystems overrepresented in the SEED dataset. Three different confidence levels (90, 95, and 99%) are plotted.

teins have been sampled. Out of the 77 phylosubsystems that are significantly overrepresented in the Sargasso Sea, 69 are from the Bacteria, 5 from the Archaea, and 3 from the Eukarya.

Fig. 2B highlights some of the differences for phylosubsystems. A more detailed description of all of the subsystems with statistically significant differences in occurrence between environmental data sets is given in the supplemental material [see Additional file 3]. Some ecologically important differences between the Sargasso Sea and the SEED database are discussed below with data extracted from the Supplementary Table.

Potential osmoregulation by amino acids in the Sargasso Sea

The phylosubsystems involved in the synthesis of serine (S), threonine (T), and glycine (G) are overrepresented in the Sargasso Sea metagenome (Table 3). For example, there are 503 proteins per million proteins in the Sargasso Sea database that are similar to bacterial glycine synthesis proteins, and only 390 proteins per million proteins in the SEED database that are similar to bacterial glycine synthesis proteins. Exactly these three amino acids are also the most abundant amino acids found in the Sargasso Sea [10,11]. However proteins in the Sargasso Sea metagenome do not contain significantly more S, T, or G (Fig. 3),

Table 2: Phylosubsystems that are overrepresented in AMD dataset versus SEED dataset with 99% confidence at a sample size of 145,000 proteins.

#	Subsystem	Classification	Domain
1	Arginine degradation	Amino Acids and Derivatives	Archaea
2	Chorismate Synthesis	Amino Acids and Derivatives	Archaea
3	Histidine Degradation	Amino Acids and Derivatives	Archaea
4	Leucine Biosynthesis	Amino Acids and Derivatives	Archaea
5	Calvin-Benson cycle	Carbohydrates	Archaea
6	Embden-Meyerhof and Gluconeogenesis	Carbohydrates	Archaea
7	Methylcitrate cycle	Carbohydrates	Archaea
8	Riboflavin metabolism	Cofactors, Vitamins, Prosthetic Groups, Pigments	Archaea
9	Conserved tRNAs	Experimental Subsystems	Archaea
10	Fatty acid metabolism	Fatty Acids and Lipids	Archaea
11	Fatty acid oxidation pathway	Fatty Acids and Lipids	Archaea
12	<i>de-novo</i> Purine Biosynthesis	Nucleosides and Nucleotides	Archaea
13	<i>de-novo</i> Pyrimidine Biosynthesis	Nucleosides and Nucleotides	Archaea
14	Pyrimidine conversions	Nucleosides and Nucleotides	Archaea
15	Ribosome LSU (eukaryotic and archaeal)	Protein Metabolism	Archaea
16	Ribosome SSU (eukaryotic and archaeal)	Protein Metabolism	Archaea
17	Translation initiation factors (eukaryotic and archaeal)	Protein Metabolism	Archaea
18	tRNA aminoacylation	RNA metabolism	Bacteria
19	TTSS transporters	Virulence	Bacteria

suggesting that the observed bias in amino acid synthesis subsystems and in the water samples is not related to protein synthesis. The primary organic osmolytes used by marine bacteria are small organic, uncharged, solutes that have little effect on the intracellular biochemistry, such as glycine betaine [12-15]. The subsystems for the production of betaine are also overrepresented in the Sargasso Sea compared to the SEED database (Fig. 2B). Similarly, serine and threonine are small, polar, uncharged amino acids that are perfect compatible solutes for balancing osmotic pressure. Serine has also been previously shown to be as effective as glycine betaine at protecting enzymes from the effects of increased osmolality [15]. In contrast, genes encoding the osmoprotectants proline, sucrose, and trehalose [13,14,16,17] were underrepresented in these samples. Therefore, we predict that marine microbes are synthesizing glycine, serine, threonine, and betaine as osmolytes, an hypothesis that can be tested experimentally.

Photosynthesis in the Sargasso Sea

As previously observed [6,7], there was a strong bias towards subsystems involved in photosynthesis in the Sargasso Sea metagenome. This bias includes subsystems for the Calvin-Benson cycle, chlorophyll biosynthesis genes, the cytochrome B6-F complex, Photosystem I, Photosystem II, isoprenoid biosynthesis, and carotenoid biosynthesis.

Some phylosubsystems involved in one-carbon metabolism, including the synthesis and degradation of carbohydrates, cell walls, and capsules are more abundant in the

Sargasso Sea. In contrast, the genes for the utilization of complex carbon sources including lactose, arabinose, fructose, mannose, galcitol and inositol are all underrepresented in the marine environment suggesting these are not significant sources of carbon in this environment.

Nucleic acid and phosphate metabolism in the Sargasso Sea

Phylosubsystems involved in purine and pyrimidine *de novo* synthesis and scavenging pathways, as well as ribonucleotide reduction (scavenging ribonucleotides for DNA synthesis) are more abundant in the Sargasso Sea. Similarly, the subsystems involved in the capture of phosphate *via* the conversion of ADP to ATP coupled to oxidative phosphorylation are also overrepresented in the Sargasso sample. In contrast, nitrogen metabolism phylosubsystems are less abundant in the Sargasso than the SEED, with the sole exception of ammonia assimilation that is marginally overrepresented in the Sargasso sample at larger sample sizes. The Sargasso Sea has previously been reported to be phosphate limited. The concentration of dissolved inorganic phosphate is approximately 4 nM in the Sargasso Sea. By comparison, the North Pacific and typical bacterial minimal media have phosphate concentrations of approximately 100 nM [18,19]. Together, these results suggest that phosphate acquisition is critical for microbial growth the Sargasso Sea environment.

Mobility of Sargasso Sea microbes

Estimates of the percentage of bacteria in the ocean that are motile vary from less than 5% to more than 80% [20,21], and there were far fewer genes encoding flagella

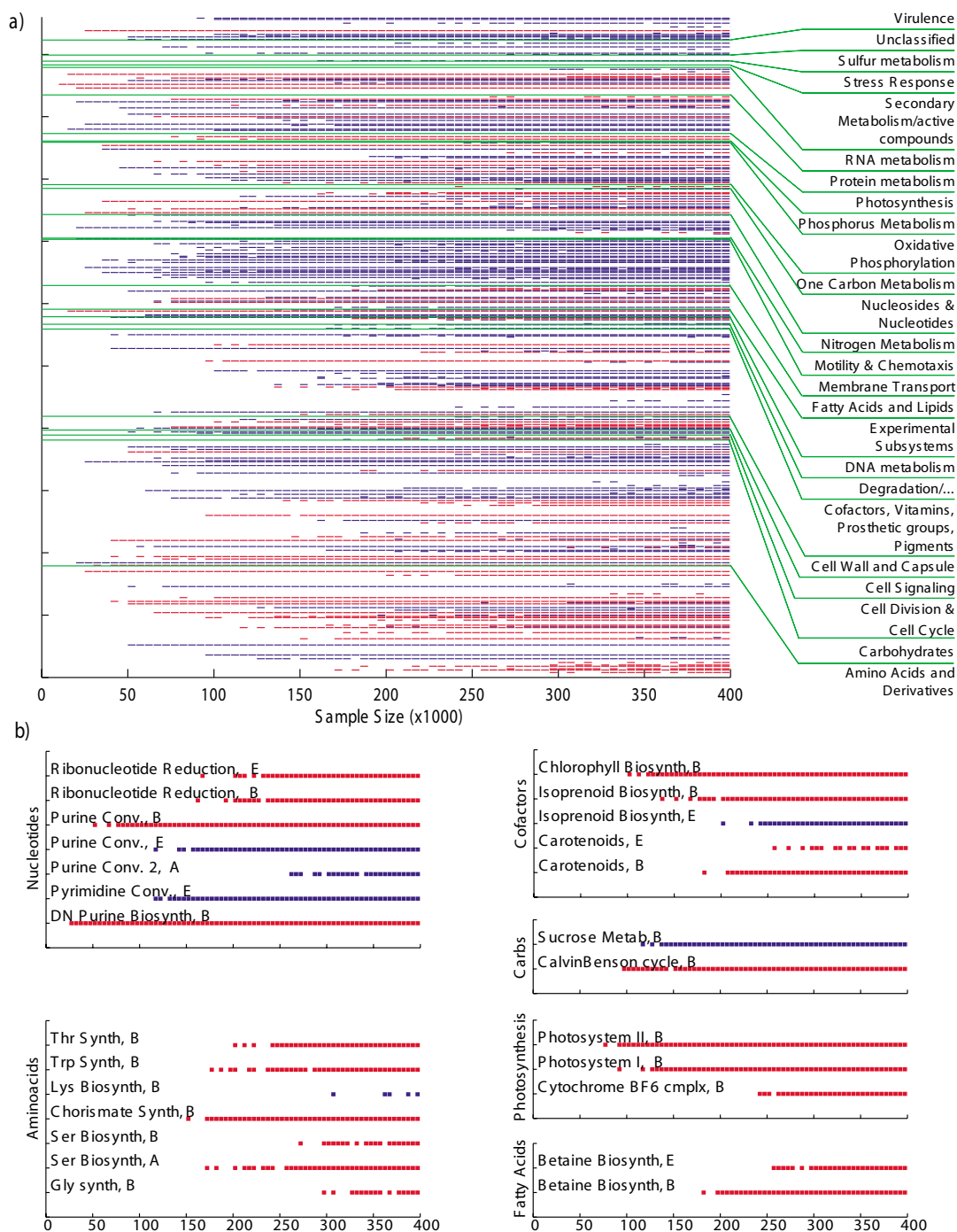


Figure 2
Significantly different subsystems between Sargasso Sea and SEED datasets. (A). Each subsystem was bootstrapped with between 10 and 400 samples per bootstrap, and subsystems that are significantly different with 99% and 2,000 bootstraps are highlighted. Those subsystems that are significantly more prevalent in the SEED database are colored blue, and those subsystems that are significantly more prevalent in the Sargasso Sea dataset are colored red. **(B).** Magnified view of several different subsystems. Subsystems from amino acid synthesis, carbohydrate utilization, cofactor synthesis, fatty acids, nucleotide synthesis, and photosynthesis are shown in more detail. Colors are as described for (A). E = eukaryotic subsystem, B = bacterial subsystem, A = archaeal subsystem.

Table 3: Presence of Glycine, Serine, and Threonine subsystems in the AMD, SEED, and Sargasso databases. The table is a subset of the data from the supplemental data [see Additional File 3].

No ¹	Name ²	K ³	Classification ⁴	# AMD ⁵	# SEED ⁶	# SS ⁷	AMD per million ⁸	SEED per million ⁹	SS per million ¹⁰	Prev. ¹¹
206	Glycine synthesis	A	Amino Acids Alanine, serine, and glycine	7	24	34	923	19	35	Sargasso
207	Glycine synthesis	B		2	492	483	264	390	503	Sargasso
426	Serine biosynthesis	A		13	166	244	1713	132	254	Sargasso
427	Serine biosynthesis	B		7	1545	1257	923	1224	1309	Sargasso
475	Threonine synthesis	B	K, T, M, and C	8	1070	952	1054	848	991	Sargasso

¹Subsystem number

²Subsystem name as designated by the curator

³K: Kingdom (A: Archaea; B: Bacteria; E: Eukaryota)

⁴Classification of the subsystem. K, T, M, and C: Lysine, threonine, methionine and cysteine.

⁵Number of proteins present in the AMD sample

⁶Number of proteins present in the SEED sample

⁷Number of proteins present in the Sargasso Sea sample (SS).

⁸⁻¹⁰Number of proteins present in the AMD, SEED, and Sargasso samples normalized per million proteins in each sample.

¹¹Statistically significant prevalence. Prev.: prevalence.

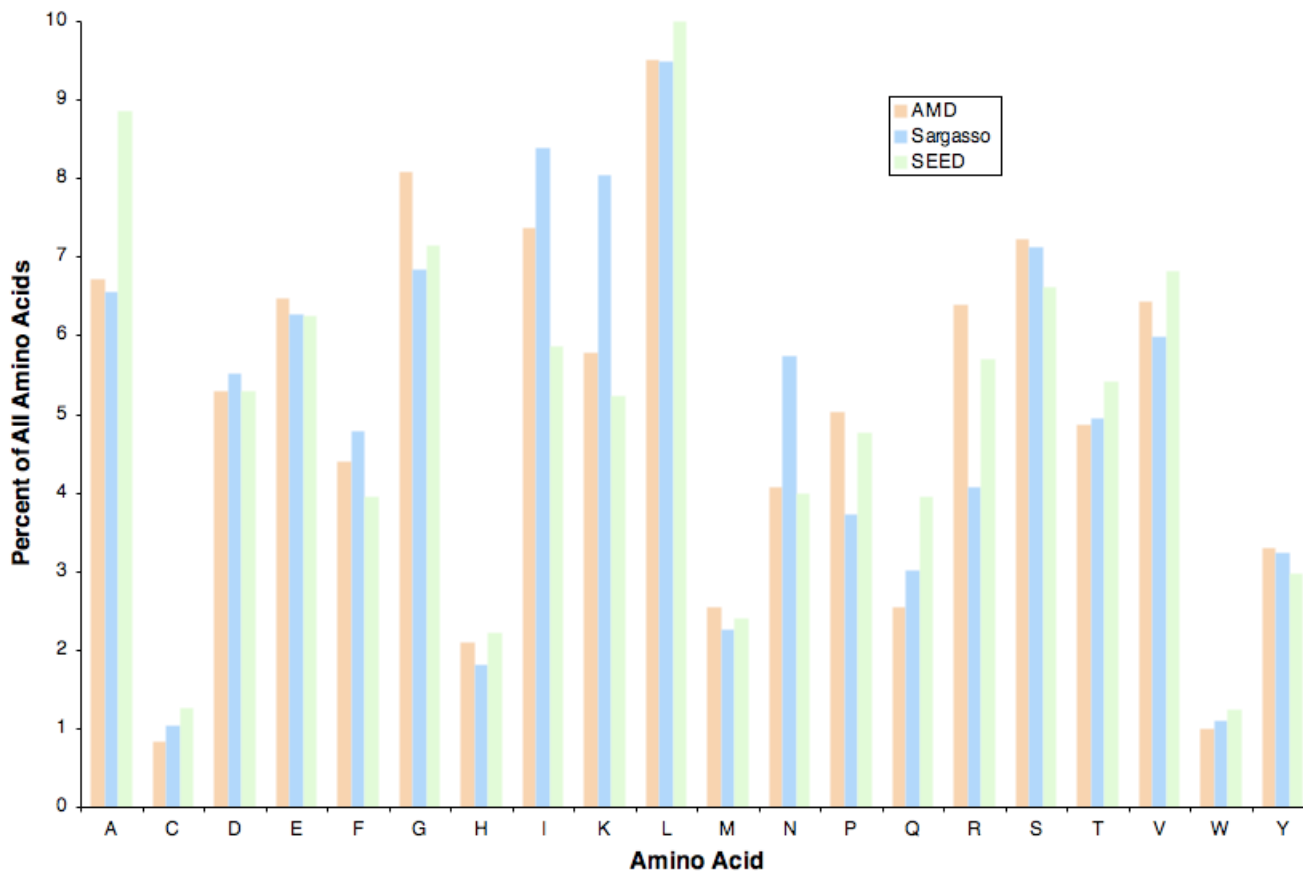


Figure 3

Fraction of amino acids in metagenomes. The fraction of each amino acid in all the predicted proteins in the three data samples was counted and compared.

base. However many marine microbes are thought to use alternative, less well characterized, motility systems, such as the motility mechanism characterized in cyanobacteria [22,23] or twitching motility previously shown in marine microbes [24]. This data leads to the hypothesis that marine microbes are generally not using flagella based motility for movement, and future studies on the genomics of twitching and gliding motility may reveal hints about these mechanisms of movement.

Subsystem differences between the SEED and AMD metagenomes

When the AMD and SEED databases were compared, only phylosubsystems that were in both the AMD and the SEED samples were included. This limited the total number of subsystems that were compared for statistically significant differences. There are far fewer phylosubsystems with significantly different distributions between the AMD and SEED datasets, and phylosubsystems that are significantly more common in the AMD dataset are shown in Table 2. The different occurrences of subsystems reflect the limited complexity of the AMD environment that contains Bacteria and Archaea [8]. The majority of subsystems that are significantly more common in the AMD data set are from archaeal proteins. In the AMD environment, the production of amino acids does not appear to be critical, and only archaeal arginine and histidine degradation and leucine and chorismate synthesis are overrepresented in these samples. Our limited selection of overrepresented subsystems in the AMD sample presumably reflects the current bias in annotated subsystems in the SEED. As the subsystems continue to evolve and expand, and the NIH Project to Annotate 1,000 Genomes [9] matures the impact of these annotations on the AMD sample and other metagenomes will highlight those areas of metabolism and physiology that are critical to survival in different environments.

Subsystem differences between the Sargasso, SEED, farm and whale metagenomes

The SEED and Sargasso subsystems were compared to both the whale fall and farm metagenome samples [6]. For this comparison the individual whale fall samples, and individual farm samples were each combined to create two separate metagenomes. Those metagenomes were compared to the subsystems exactly as described in Methods, using the BLAST algorithm to determine similar sequences. The data shown in the supplemental material [see Additional file 4] was created using 95% confidence, a sample size of 20,000 proteins, and 20,000 replicates. This table shows each of the comparisons with the statistically significant subsystems. The normalized data was used to determine the relative abundance of each KEGG pathway in each sample [25], and these comparisons are

shown in the supplemental material [see Additional File 5].

The KEGG pathways have historically focused on core metabolism, annotating enzymes that have been classified with EC numbers. In contrast, the SEED subsystems include core metabolism and the data is extended to subsystems that cover cellular processes and functions, regulation, and so forth. Although the two classification techniques are not directly comparable, and statistical confidence was not provided with the differences between KEGG pathways in the supplemental data from the previous analysis, some clear parallels can be seen between these analyses [see Additional File 5]. For example, both techniques identified that riboflavin metabolism is more prevalent in the Whale Fall metagenomes than the other samples, however according to the normalized data from Tringe *et al* folate biosynthesis is less abundant in the Sargasso metagenome than either the Whale Fall or Soil Metagenomes whereas this analysis demonstrated that there is significantly more folate biosynthesis in the Sargasso than the other samples. There were 9,311 proteins with similarity to folate biosynthesis subsystem from the SEED database in the Sargasso metagenome, 602 proteins with similarity in the farm soil metagenome and 491 proteins with similarity in the Whale Fall metagenome. In contrast, Tringe *et al.* identified 7,283 proteins, 1,253 proteins, and 889 proteins respectively. These differences are probably due to the difference in annotation of the SEED subsystems and KEGG pathways. These differences also highlight the need for continued careful annotation of genomes, and comparative analysis of different annotation systems and methods.

Discussion

Community genome sequencing – metagenomics – can provide fine detailed analysis of the metabolism occurring in different ecosystems. However, metagenomics analysis is limited to a purely descriptive science without a rigorous statistical comparison of the prevalence of different genes in different environments. Our analysis demonstrates an application of statistics to identify those areas of metabolism that are significantly over represented in different environments.

The method described here is predicated on the expectation that genes that are more useful in an environment are more commonly found in that environment. Or put another way, there is an enrichment or selection for sets of genes in different environments. A statistical analysis, using a resampling with replacement technique, was developed to generate both the difference in occurrence of each subsystem in each sample, and to generate confidence intervals for the likelihood that these differences are observed by chance. By using these statistical techniques

to compare the genetic composition of different environments, the areas of metabolism and biochemistry that are important in a particular environment, in comparison to other environments, can be identified. Like other studies, this analysis demonstrated that microbes in the surface waters of the ocean are much more likely to contain genes involved in photosynthesis than the control data set. The non-redundant database used as a control is not expected to contain large numbers of photosynthetic organisms because it is skewed towards microbial pathogens.

Our analysis also demonstrated more than 150 other subsystems that are over represented in the Sargasso Sea sample when compared to the control set. The skew in the database alone cannot explain this difference, and these subsystems must be important for survival in the ocean. Some examples, such as the synthesis of serine, threonine, and glycine, directly testable hypotheses can be generated from these analyses. For other examples, the explanation of the differences between samples may be more elusive. Several pieces of evidence will assist in determining the roles of different subsystems in different environments. For example, the inclusion of more environmental data with each sample will allude to some of the differences in metabolic potential between samples; the careful dissection of the presence of different subsystems in different organisms will identify which organism in which environment is performing the different biochemical reactions; and the extension of other techniques such as metabolic modelling into the environmental arena may provide insights into the critical biochemical mechanisms in each environment.

Conclusion

Comparative metagenomics is a powerful mechanism for highlighting the ecological differences between environmental samples. Our analysis has generated several hypotheses that can be readily tested on microbes from the Sargasso Sea:

1. Serine, threonine, and glycine betaine are primarily being used as osmoprotectants. Increased intracellular concentrations of serine may protect and against the osmolarity of the ocean. In contrast, sucrose and trehalose are not being used as readily for osmoprotection.
2. Microbes in the Sargasso Sea are more limited for phosphate than nitrogen.
3. Microbes in the ocean are not generally using flagella based motility but are probably using one of the less-characterized mechanisms of locomotion.
4. Archaea in the AMD sample are degrading arginine and histidine.

The subsystems approach to investigating environmental genomes demonstrates the intricate interplay between the abundance of genes in the environment and the biology of that environment. In addition to answering that age-old knock-knock joke [26] by cataloging the organisms that are present in an environment and looking for novel proteins and structures, metagenomics also provides critical insights into our understanding of the physiology, biology, and ecology of an environment. Using subsystems to compare the ecology of sites that have been sampled by metagenomics can be applied to any other metagenome samples to provide similar insights into the ecology of those environments.

Methods

Sequence databases

The complete SEED database v4 was used as the source of all data [27]. Construction and annotation of the subsystem database is described elsewhere [9,28]. The environmental sequences were removed from the SEED database for the analyses presented here. Furthermore, any sequences with principal homology to either *Shewanella* sp. or *Burkholderia* sp. were removed from the Sargasso Sea metagenome because of contamination concerns [29]. This dataset contained 960,561 predicted proteins. The AMD data set contained 7,588 predicted proteins. For these analyses the term "protein" is used when referring to predicted proteins.

Assignment of proteins to subsystems and phylosubsystems

Each protein from the AMD, Sargasso, or the SEED database was compared to proteins in the SEED database previously assigned to particular subsystems by the SEED annotators [9]. A protein was considered a member of a subsystem if the protein had significant similarity (designated as an E value less than 1×10^{-20}) to another protein previously assigned to a subsystem. Each protein was also classified as Bacteria, Eukarya, and Archaea, based on the Domain assignment of the most similar protein. There were a total of 276 annotated subsystems in the SEED. Bacteria had proteins in 257 of the subsystems. Archaea and Eukarya had proteins in 132 and 134 of the subsystems, respectively. This means that each there were a total of 523 potential data points. The term phylosubsystem is used to reflect that fact that the assignments are based both on the subsystem and Domain.

Comparisons of metagenomic databases

A flowchart of the methods used to identify subsystems with statistically significant differences between databases is shown in Fig. 4 and described in the supplemental material [see Additional File 2]. In addition, source code to software to calculate these differences is provided as supplemental material [see Additional File 1]. In order to decide whether a subsystem was over-represented

1. Predict Protein Sequences from genomes or metagenomes (SEED and GenDB developers)



Genome ID	Organism	Variet Code	PrsA	PurF	PurD	PurN	PurT	PurL	PurQ	PurS	PurM	PurE	PurK	PurC	PurB
167540.1	Prochlorococcus marinus subsp. pastoris MED4 [B]	28	1175	452	881	1407	1276	453	1538	1539	585	281	1665	834	707
74547.1	Prochlorococcus marinus str. MIT 9313 [B]	28	1072	4	1410	709	2258	3	539	540	1665	2020	668	1808	1478
167539.1	Prochlorococcus marinus subsp. marinus str. CCMP1375 [B]	28	1116	4	1418	943	1182	3	852	851	1352	191	1006	1417	1616
59910.1	Prochlorococcus marinus subsp. pastoris str. CCMP1986 [B]	28	1076	4	1376	888	1000	3	727	726	1385	167	680	1335	1461

2. Annotations and Subsystem Characterizations (SEED Annotators)

3. Compare all predicted proteins with all other predicted proteins (precalculated in SEED)

ASSIGN to	Similar sequence	E-val	region in similar sequence (color explained)	region in fig16487.1.ppt.1425 (color explained)	ASSIGN from	In Sub	Function (GeneID color explained)	Organism
SELECT	fig16588.1.ppt.258	0.0 81.2%	1-528 (528/528)	5-532 (528/532)	~	1	D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95)	Synechococcus sp. WH 8102
	fig16539.1.ppt.1433	0.0 77.2%	1-527 (527/528)	5-531 (527/532)	~	2	D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95)	Prochlorococcus marinus subsp. marinus str. CCMP1375
	fig16519.1.ppt.1350	0.0 72.54%	1-528 (528/528)	5-532 (528/532)	~	2	D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95)	Prochlorococcus marinus subsp. pastoris str. CCMP1986
	fig16583.1.ppt.669	0.0 72.54%	1-528 (528/528)	5-532 (528/532)	~	2	D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95)	Prochlorococcus marinus subsp. pastoris MED4

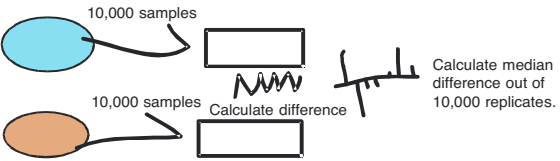
Subsystems in which this cds is present

Subsystem	Role
Serine Biosynthesis	Phosphoserine aminotransferase (EC 2.6.1.52)

4. Find predicted proteins in subsystems or predicted proteins with homologs ($E < 10^{-20}$) in subsystems

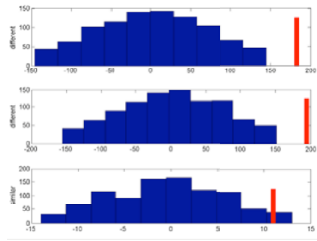
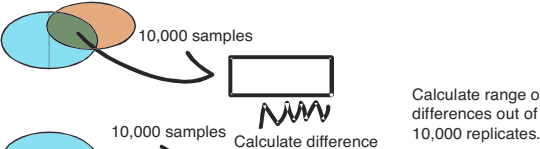
5. Classify protein, homolog, domain of homolog, and subsystems for each predicted protein in Sargasso, AMD, and SEED-nr databases

426	Serine Biosynthesis	Archaea	13	166	244	1715.231418	131.5262993	254.1182248	Sargasso
427	Serine Biosynthesis	Bacteria	7	1545	1257	922.5092251	1224.145376	1308.610281	Sargasso
428	Serine Biosynthesis	Eukaryota	0	52	11	0	41.2010942	11.4E164128	SEED-nr



6. Choose 10,000 samples with replacement from each database. Calculate the difference for each subsystem. Repeat 10,000 times and report the median differences

7. Choose 20,000 samples at random from both databases to generate two mixed populations and calculate the difference. Repeat 10,000 times and obtain the range of differences



8. Compare the median of #6 with the range from #7 for a given confidence interval to determine whether it is significant

Figure 4
Flow chart of methods used to identify statistical differences between phylosubsystems.

between metagenomes, a comparison between the median number of proteins assigned to specific subsystems was performed for a given confidence level. The following steps were carried out: 1) A number N of proteins were drawn at random, with replacement, from each of two metagenomes. Each protein was classified into a subsystem. Then, the difference between the two sets of samples was calculated for each subsystem, resulting on a list of differences between metagenomes by subsystems. These differences were expected to be more pronounced where the differences between metagenomes were more remarkable, but with just one set of differences this was not guaranteed. Thus, this procedure was repeated M times. Then, for each subsystem the median of all the M differences was calculated. 2) To build a confidence interval, a number N of "mixed" samples were drawn at random, with replacement, from a mixture of both metagenomes. For each sample, two random numbers were drawn, the first one to decide which metagenome would be selected, and the second one to choose a particular element of the metagenome, which was the accounted into its corresponding subsystem. This process was repeated for another set of N samples and with these two mixed sets a difference was calculated. This whole process was repeated M times again, and for each subsystem the resulting differences were ordered and used to build a confidence interval. The limits of the confidence interval for each subsystem were estimated by using the corresponding quantile elements on the ordered set of differences of each subsystem as limits to the confidence interval. For example, if $M = 1,000$ and the confidence level was 90%, the limits for the confidence interval were at the 5% and 96% percentiles on the ordered values of the differences for that subsystem, that is, the 50th and 951st elements would be the upper and lower limits on the confidence interval. 3) For any given subsystem, if the median from step 1 was inside the confidence interval determined in step 2 it was considered that the subsystems were not significantly different at the given confidence level. If the median was outside the confidence interval, it was assumed that the difference was significant at the given confidence level.

Software to calculate the statistics

A software package that calculates the statistics from appropriately formatted files is provided as supplemental material [see Additional File 1]. The software is released under the GPL license and is available from the BMC website.

Authors' contributions

BRB developed the statistics and compared the samples. FR contributed to the discussion on roles of subsystems in the samples. RE generated the connections

between proteins and subsystems. All authors contributed to authoring and editing the manuscript.

Additional material

Additional File 1

Step-by-step description of the methods used to calculate the differences between subsystems

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-7-162-S1.tgz>]

Additional File 2

A tar gzip compressed archive of source code and instructions for calculating difference of medians as described here. Version 0.01 of the software, released under the terms of the Gnu General Public License (GPL).

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-7-162-S2.PDF>]

Additional File 3

Presence of subsystems in different environmental samples in tab-separated text format.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-7-162-S3.txt>]

Additional File 4

Presence of subsystems in the AMD, Sargasso, and SEED samples in tab-separated text format.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-7-162-S4.txt>]

Additional File 5

Abundance of selected Kegg Pathways and Subsystems that are approximately similar as detected either based on the Normalized differences of Tringe et al (reference 6) or using the method described herein. Note that the relative gene content of the subsystems and Kegg pathways is never identical.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-7-162-S5.doc>]

Acknowledgements

The authors thank Matt Cohoon, Sveta Gerdes, Andrei Osterman, Ross Overbeek, Rick Stevens, Olga Vassieva, Veronika Vonstein, Olga Zagnitko, and the other developers and annotators working on the SEED for their invaluable contributions. The authors also thank Mya Breitbart, Ed DeLong, Stanley Maloy, Braudel Maqueira, Ross Overbeek, and Anca Segall, for helpful discussion and critical reading of the manuscript. The Gordon and Betty Moore Foundation Marine Microbiology Initiative grant to FLR and the NSF Biocomplexity Program (NSF0221763) to John Paul (University of Southern Florida) and Anca Segall (San Diego State University) funded this work.

References

1. Riesenfeld CS, Schloss PD, Handelsman J: **Metagenomics: genomic analysis of microbial communities.** *Annu Rev Genet* 2004, **38**:525-552.

2. Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P, Rohwer F: **Diversity and population structure of a near-shore marine-sediment viral community.** *Proc R Soc Lond B Biol Sci* 2004, **271(1539)**:565-574.
3. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F: **Metagenomic analyses of an uncultured viral community from human feces.** *J Bacteriol* 2003, **185(20)**:6220-6223.
4. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F: **Genomic analysis of uncultured marine viral communities.** *Proc Natl Acad Sci U S A* 2002, **99(22)**:14250-14255.
5. Cann AJ, Fandrich SE, Heaphy S: **Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes.** *Virus Genes* 2005, **30(2)**:151-156.
6. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM: **Comparative metagenomics of microbial communities.** *Science* 2005, **308(5721)**:554-557.
7. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304(5667)**:66-74.
8. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428(6978)**:37-43.
9. Overbeek R, Begley T, Butler R, Choudhuri J, Chuang H, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank E, Gerdes S, Glass E, Goessmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy A, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch G, Rodionov D, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.** *Nucleic Acids Res* 2005.
10. Keil RG, Kirchman DL: **Utilization of dissolved protein and amino acids in the northern Sargasso Sea.** *Aquatic Microbial Ecology* 1999, **18(3)**:293-300.
11. Suttle CA, Chan AM, Fuhrman JA: **Dissolved free amino-acids in the Sargasso Sea - uptake and respiration rates, turnover times, and concentrations.** *Marine Ecology-Progress Series* 1991, **70(2)**:189-199.
12. Oren A: **Diversity of halophilic microorganisms: Environments, phylogeny, physiology, and applications.** *Journal Of Industrial Microbiology & Biotechnology* 2002, **28(1)**:56-63.
13. Roesser M, Muller V: **Osmoadaptation in bacteria and archaea: common principles and differences.** *Environ Microbiol* 2001, **3(12)**:743-754.
14. Sleator RD, Hill C: **Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence.** *Fems Microbiology Reviews* 2002, **26(1)**:49-71.
15. Yancey PH, Clark ME, Hand SC, Bowlus RD, Somero GN: **Living with water stress: evolution of osmolyte systems.** *Science* 1982, **217(4566)**:1214-1222.
16. Galinski EA: **Osmoadaptation in bacteria.** *Adv Microb Physiol* 1995, **37**:272-328.
17. Mackay MA, Norton RS, Borowitzka LJ: **Organic osmoregulatory solutes in cyanobacteria.** *Journal Of General Microbiology* 1984, **130(SEP)**:2177-2191.
18. Deiwick J, Nikolaus T, Erdogan S, Hensel M: **Environmental regulation of Salmonella pathogenicity island 2 gene expression.** *Mol Microbiol* 1999, **31(6)**:1759-1773.
19. Wu JF, Sunda W, Boyle EA, Karl DM: **Phosphate depletion in the western North Atlantic Ocean.** *Science* 2000, **289(5480)**:759-762.
20. Grossart HP, Riemann L, Azam F: **Bacterial motility in the sea and its ecological implications.** *Aquatic Microbial Ecology* 2001, **25(3)**:247-258.
21. Mitchell JG, Pearson L, Bonazinga A, Dillon S, Khouri H, Paxinos R: **Long lag times and high velocities in the motility of natural assemblages of marine-bacteria.** *Applied And Environmental Microbiology* 1995, **61(3)**:877-882.
22. McCarren J, Heuser J, Roth R, Yamada N, Martone M, Brahamsha B: **Inactivation of swmA results in the loss of an outer cell layer in a swimming synechococcus strain.** *J Bacteriol* 2005, **187(1)**:224-230.
23. Waterbury JB, Willey JM, Franks DG, Valois FW, Watson SW: **A cyanobacterium capable of swimming motility.** *Science* 1985, **230(4721)**:74-76.
24. Henrichsen J: **The occurrence of twitching motility among gram-negative bacteria.** *Acta Pathol Microbiol Scand [B]* 1975, **83(3)**:171-178.
25. **Supplemental Data on the String Website** [http://string.embl.de/metagenome_comp_suppl/keggmap.detection.fre_quencies.txt]
26. Oremland RS, Capone DG, Stolz JF, Fuhrman J: **Whither or wither geomicrobiology in the era of 'community metagenomics'.** *Nat Rev Microbiol* 2005, **3(7)**:572-578.
27. **The SEED** [<http://theseed.uchicago.edu/FIG/index.cgi>]
28. Meyer F, Goessmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, Puhler A: **GenDB--an open source genome annotation system for prokaryote genomes.** *Nucleic Acids Res* 2003, **31(8)**:2187-2195.
29. DeLong EF: **Microbial community genomics in the ocean.** *Nat Rev Microbiol* 2005, **3(6)**:459-469.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

