5

Metagenomics in Polluted Aquatic Environments

Alexander M. Cardoso[1,5], Felipe H. Coutinho[1] et al.[*]
[1]*Instituto de Bioquímica Médica, Universidade Federal do Rio de Janeiro*
[5]*Instituto Nacional de Metrologia, Qualidade e Tecnologia*
Brazil

1. Introduction

Metagenomics is defined as the culture-independent genomic analysis of biological assemblages providing access to the whole set of genes and genomes from a sample. It encompasses a variety of techniques that are based on (i) total DNA extraction from samples followed by PCR amplification of specific genes, (ii) library construction or amplification and sequencing of the whole genetic material. These methodologies have successfully been applied in studies of composition, dynamics, and functions of microbial communities in a variety of ecosystems including those subjected to anthropogenic modifications (Gilbert & Dupont, 2011).

Culture independent methods allow the analysis of a set of metabolic genes from microbial communities, which can be used to determine how environmental conditions such as pollution can shape community composition and the diversity of genes associated with biogeochemical cycles such as those of carbon, nitrogen, and phosphorus (Singh et al., 2009). This approach is also useful for the discovery of novel environmental microorganisms and genes, with important applications for biotechnology, medicine, and bioremediation (Cardoso et al., 2011).

This applicability has resulted in a recent sharp increase in studies focusing in the metagenomic analysis of polluted sites. Their aim is to characterize microbial communities from a diverse set of environments such as freshwater, marine sediments, open ocean, pelagic ecosystems, soil, and host-associated communities. An example of these initiatives is the Global Ocean Sampling Expedition (GOS), which assessed the genetic diversity of marine microbial communities around the Earth. Since 2003, an enormous amount of data has been generated by GOS helping scientists to reveal the microbial diversity and also allowing them to better understand microbial phylogeny and ecology (Gilbert & Dupont, 2011).

---

[*] Felipe H. Coutinho[1], Cynthia B. Silveira[1], Barbara L. Ignacio[1], Ricardo P. Vieira[1], Gigliola R. Salloto[2], Maysa M. Clementino[2], Rodolpho M. Albano[3], Rodolfo Paranhos[4], Orlando B. Martins[1]
[1]*Instituto de Bioquímica Médica, Universidade Federal do Rio de Janeiro*
[2]*Instituto Nacional de Controle de Qualidade em Saúde*
[3]*Departamento de Bioquímica, Universidade Estadual do Rio de Janeiro*
[4]*Instituto de Biologia, Universidade Federal do Rio de Janeiro*
[5]*Instituto Nacional de Metrologia, Qualidade e Tecnologia*
Brazil

2. Metagenomics and bioinformatics

Microorganisms can be found across all environments. Adequate sample collection is the initial and essential step to achieve a comprehensive coverage of the microbial diversity. For aquatic studies, the collection of large volumes of water followed by filtration is recommended since it increases the chance of retrieving rare groups. After sample collection, an optional enrichment culture step can be performed to maximize the abundance of a targeted group of microorganisms by providing its ideal growth conditions. Further screening for specific phenotypic features may be performed during the cultivation step in order to target microbial species of unusual metabolism.

Genetic material can be obtained from samples by several methods that include physical and/or chemical cell lysis followed by extraction and purification of nucleic acids (DNA or RNA). However, the amount of non-biological matter associated with the biological material may interfere with the extraction, quantification and amplification processes. Thus, different samples will require different extraction methods as a good representation of the biological diversity relies on the efficiency of the nucleic acid extraction step. The extracted genetic material must be free of amplification inhibitors and special attention must be given when dealing with samples retrieved from polluted sites (Cardoso et al., 2010).

Metagenomics may shed light in understanding the complex degradation routes of xenobiotics, which is currently poorly understood. These compounds can be toxic for living organisms and represent a threat to ecosystems. The use of molecular techniques based on genomic analysis can be applied in the monitoring of enzymes associated with the metabolism of xenobiotics, including herbicides and other pollutants (Malik et al., 2008).

Microbial communities can be screened by gene-specific PCR to detect the presence of genes of interest within a community. This method, however, has a bias of favoring previously known genes. An alternative method, which allows for the identification of novel genes and metabolic pathways, is the construction of expression libraries from metagenomic DNA. Through this method, positive clones expressing randomly cloned environmental genes are screened for their capacity to metabolize a specific substrate by plating in media containing a particular compound. In this way, the clones that metabolize this compound can be selected for DNA sequencing.

By gene-specific PCR it is possible to quickly identify genes associated with biodegradation in an environmental sample and sometimes affiliate them with a specific taxonomic group. This approach is limited by the fact that the metabolic pathways to which xenobiotics are submitted can encompass several steps, requiring more than a single enzyme to be catabolized. Therefore, cloning the full set of genes would be required to obtain the desired phenotype, which is not always possible since those enzymes can be encoded by different genes spread throughout the genome.

An alternative procedure for novel gene discovery is by Stable Isotope Probing (SIP). SIP is based on the incorporation of stable isotope-labeled substrates into molecular biomarkers. Once labeled substrates are provided to a microbial community, those microorganisms capable of metabolizing this substrate (e.g. a pollutant molecule) are likely to incorporate labeled atoms in their DNA, RNA and protein molecules. Nucleic acids with labeled atoms can be extracted to retrieve genomic material exclusively from a community capable of metabolizing a desired substrate (Malik et al., 2008). Furthermore, the sequencing of complete genomes of microorganisms can also reveal a set of metabolic genes with important applications for the biodegradation process, representing another important tool for bioremediation strategies (Eyers et al., 2004).

However, metagenomic studies usually result in the production of a great amount of sequence data so an advance in the capacity of bioinformatics tools is expected to deal with it. The processing of metagenomic data for subsequent genetic/environmental analysis requires a series of bioinformatics tools since the abundance of sequences obtained, especially with next generation sequencing equipments, can no longer be processed manually. Considering the high number of bioinformatics tools that have been and are still being developed, this chapter will focus on those most used for microbial community analysis in water pollution studies.

Sequences obtained from environmental metagenomic projects are usually compared to local or global nucleic acid sequence databases to identify the relationship between them and previously obtained data. Popular databases include GenBank (http://www.ncbi.nlm.nih.gov), the RDP (http://rdp.cme.msu.edu), CAMERA (http://camera.calit2.net), SILVA (http://www.arb-silva.de) and others, some of them dedicated to particular taxa. General databases are more fitted for a broad analysis whereas databases dedicated to a specific taxonomic group may contain more rare and reliable sequences. DNA sequence databases can also be divided between curated databases and non-curated ones. While some databases accept all sorts of nucleic acid and protein sequences (non-curated), others are more restrictive and perform a pre-selection of deposited sequences leaving a lower but more reliable amount of sequences available.

When using DNA sequencing for taxonomic identification ribosomal genes are frequently used but there is no gene or genomic region that is a golden standard for such procedure. In fact, some sequences present high similarity levels with more than one gene or a single species. Usually a DNA sequence from a single gene can confer a trustful identification to a family or genus level, depending on the gene and size of the DNA fragment. However, there is still much debate regarding which portion of genome is reliable for identification of organisms to a species level and whether such a perfect region does exist.

Several sequence alignment tools are available for the comparison of protein and nucleic acid sequences. A reliable alignment is required for a common practice within metagenomic approaches: the construction of phylogenetic trees. These trees, which may be constructed using software such as MEGA and ARB (http://www.megasoftware.net; http://www.arb-silva.de), are widely used to show sequence diversity within samples and to determine how these sequences are related to each other in evolutionary terms. It is important to highlight that sometimes the trees should not be interpreted as a precise evolutionary model of the studied sequences but as mere representations of which sequences are present in a selected sample and how they are related to reference sequences and to each other.

Computational tools can be also used in the quantification of differences between distinct datasets. For example, the software LIBSHUFF (http://whitman.myweb.uga.edu/libshuff.html) generates homologue and heterologous coverage curves to compare gene libraries in order to determine if two sequence libraries are statistically distinct from each other. UniFrac can be used for comparing microbial communities through principal component analysis, allowing several datasets to be compared at once, which helps in the detection of distribution patterns in communities of microorganisms and to correlate them with environmental variants (Lozupone & Knight, 2005).

Correspondence analysis is a similar approach, which helps to quantify how much of the microbial diversity is explained by environmental variables, with major applications for pollution studies as it determines the extent to which microbial communities are affected by pollutants (Vieira et al., 2008). Concerning microbial community diversity, DOTUR can be

used as a tool to assign sequences as operational taxonomic units (OTUs) (Schloss & Handelsman, 2005). In addition, this software can be used to generate rarefaction and collector's curves and diversity indexes. Altogether these features help to quantify how much of the microbial diversity is being covered within a collected sample, to estimate how much of biological diversity is being retrieved through a metagenomic approach from a sample, and to determine how much effort is still required to reach a full coverage of sequence diversity.
Bioinformatics tools are evolving towards packages which aggregate different softwares to facilitate analysis. MOTHUR has emerged as a powerful tool for comparison of microbial communities (Schloss et al., 2009), allowing several steps of bioinformatics from metagenomic analysis to be performed in a single platform. MOTHUR implements tools like LIBSHUFF, DOTUR and UniFrac improving the use of bioinformatics tools to non-specialists. A simple flowchart summarizing the several steps of metagenomic analysis and bioinformatics is shown in Figure 1.

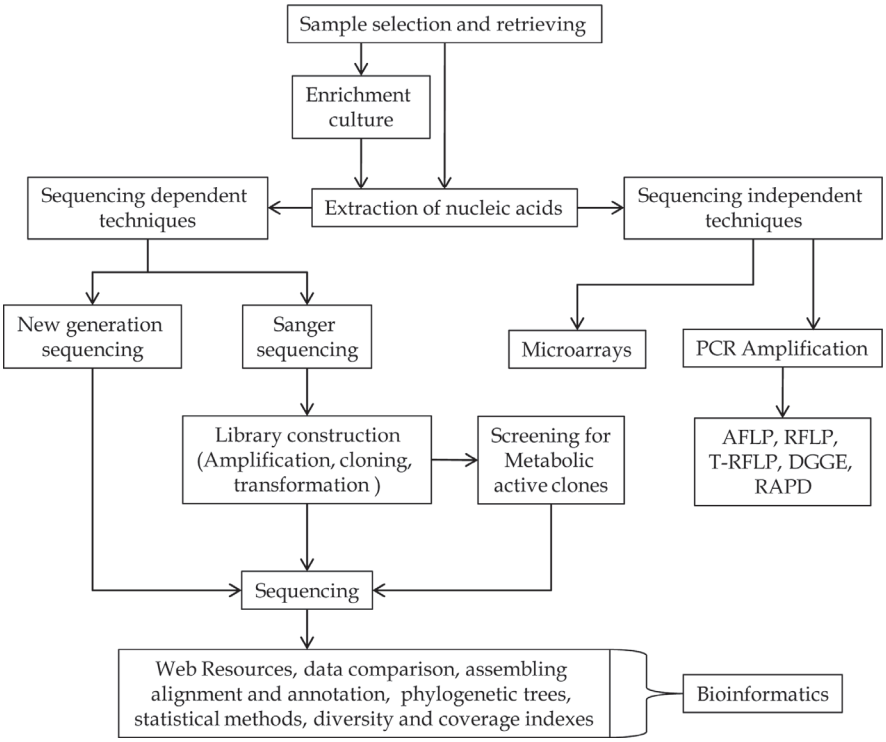

Fig. 1. Flowchart summarizing the several steps of metagenomic analysis and bioinformatics.

3. Metagenomics applied to water pollution studies

3.1 Guanabara Bay

Guanabara Bay is a coastal bay located in Rio de Janeiro state, SE Brazil, centered on latitude S 22°50' and longitude W 43°10', with a perimeter of 131 km and an area of 384 km², of

which 328 km² are of water surface and 56 km² are comprised of islands. The mean water volume is 1.87 x 10^9 m^3. The bay measures approximately 28 km from west to east and approximately 30 km from south to north. The narrow entrance to the bay is 1.6 km wide. Guanabara Bay's drainage basin comprises a set of 32 separate sub-watersheds and is drained by 45 rivers and stream channels spread throughout an area of 4080 km². The bay's water surface area has suffered a reduction of about 10% due to urban expansion.

The bay's climate is tropical humid, with wet warm summers and dry cool winters, a mean annual air temperature of 23.7 °C and a mean rainfall of 1173 mm. The bay is characterized by high salinities and temperatures; the mean salinity is 29.5 ± 4.8 % with a total range from 9.9 to 36.8%, characterizing the bay as an estuarine environment. The average temperature is 24.2 ± 2.6 °C with a total range from 17.0 °C to 31.0 °C. Salinity decreases horizontally from the ocean entrance towards the inner reaches of the bay due to freshwater discharge while temperature increases from the ocean entrance towards the inner reaches (Paranhos & Mayr, 1993).

Guanabara Bay is located in a region that was originally covered by the Atlantic rainforest, an ecosystem of which only a very small fraction was conserved. Vegetation was cut down along the centuries as the region developed economically to become the second largest metropolitan area of Brazil. Most of the vegetation preserved along the basin is restricted to conserved portions of mangrove wetlands, in small fringes of the bay and covering mountain slopes. Studies have demonstrated a possible role for mangrove ecosystems in the attenuation of the anthropogenic impact imposed over the bay, raising awareness for the need to preserve and restructure those environments (Gomes et al., 2008; Santos et al., 2011).

Guanabara Bay is surrounded by the metropolitan region of Rio de Janeiro city, encompassing several towns and smaller communities, totaling around 10 million inhabitants. Sewage runoff is discharged, mostly untreated, directly into the bay and in the rivers and streams that flow into it, which have been shown to undergo a severe eutrophication process (Mayr et al., 1989).

Rio de Janeiro went through an abrupt industrialization and intense population increase in the 1950s when the bay started to suffer with a severe anthropogenic impact. It is estimated that nowadays 540 tons of high biochemical oxygen demand waste and 5.5 tons of garbage are deposited daily into the bay (FEEMA, 2008). A large number of industries are present on its borders and thousands of ships circulate in its waters every year. Also, the bay's drainage basin possesses several urban, cattle raising, aquaculture, and agricultural areas in its surroundings, besides harboring one of the main industrial poles in the country. The bay is also subjected to several types of pollutants such as heavy metals, pesticides, antibiotics, oil, polycyclic aromatic hydrocarbons and organochlorines, raising the bay as one of the most impacted coastal environments in Brazil. Bay pollution has implications in public health issues, since its waters are utilized for recreational purposes and commercial fishing.

In the beginning of the 1990s a cleansing program was created for Guanabara Bay. Until 2010, almost a billion US dollars have been spent in an attempt to decrease the levels of pollutants discharged into the bay, which nonetheless persists as a highly impacted estuarine environment. Recent studies have shown that the bay has a high overload of organic and inorganic nutrients. Higher mean nutrient concentrations are found in the inner bay margins, due to sewage runoff and less efficient water renewal. At these same sites significant differences are observed in nutrient concentration between surface and bottom waters while nutrient levels are strongly reduced towards the bay's entrance (Vieira et al., 2007; Vieira et al., 2008).

Organic matter can either be deposited into the bottom sediments or undergo bacterial decomposition, consequently increasing nutrient availability in the water column and promoting eutrophication. The innermost reaches of the bay present elevated levels of biochemical oxygen demand as a result of the consumption of dissolved oxygen for degradation of organic matter derived from sewage runoff. High levels of chlorophyll a are also highest in the inner portions of the bay, presumably due to elevated nutrient overload allowing for higher phytoplankton production.

3.1.1 Metagenomics studies in Guanabara Bay

Estuaries can be biologically very diverse due to the mixing of seawater and freshwater so more knowledge about the composition and the ecological roles played by estuarine microbial communities is necessary. In 2007 and 2008, Vieira and colleagues studied the free-living planktonic Archaea and Bacteria diversity through the construction of 16S rRNA gene libraries, in a transect along Guanabara Bay, extending from a heavily polluted inner channel to the pristine costal seawater of the southern Atlantic Ocean. Nutrient levels along the transect revealed a clear trophic and pollution gradient along the bay, with higher nutrient levels in inner bay sites that decreased towards the bay`s entrance and beyond. This gradient reflected in bacterial abundance and production, which peaked in inner sites. Bacterial diversity in three of the sampled sites revealed clones associated with marine, estuarine and brackish waters and also with species present in sewage, indicating the effect of pollution in structuring the microbial communities of the bay. Additionally, in the highly polluted anoxic inner channel, several clones were affiliated with opportunistic pathogenic genera. This finding illustrates the risks to public health associated with impacted environments.

Exploring the diversity of bacterioplankton communities in a latitudinal gradient along Latin America through pyrosequencing of the V6 hipervariable region of the 16S rRNA gene, Thompson and colleagues, revealed in 2011 that among 7 different coastal seawater sites, that included Guanabara Bay, *Proteobacteria* followed by *Cyanobacteria*, *Bacteroidetes* and *Actinobacteria*, are the most abundant phyla. *Proteobacteria* accounted for more than 75% of the 17631 sequence tags obtained from the bay, with similar proportions observed in the other sites, indicating that this phylum is dominant among coastal seawater environments. Potentially pathogenic microorganisms were also detected within the bay, which is probably a result of nutrient overload that promotes the growth of heterotrophic bacteria, which include a diverse set of opportunistic pathogens (Thompson et al., 2011). This explains how a previously less harmful microbiota from a pristine aquatic environment is replaced by a set of new microorganisms that present higher risks to human health upon pollution impacts. This work is an excellent example of the applications of pyrosequencing to uncover rare members and perform a quantitative analysis of microbial communities which can be used to determine distribution patterns of microorganisms along the globe. On the other hand, the low extension of V6 tags obtained in this approach (~60 nt), hampers a full taxonomic characterization of sequences to a species level.

Archaeal species diversity in water samples from 4 distinct bay sites, with different levels of anthropogenic impact, was evaluated. Denaturing gradient gel electrophoresis (DGGE), revealed that free-living and particle attached archaeal communities are significantly different (Vieira et al., 2007). Suspended particulate material (SPM) is abundant in several sites of Guanabara Bay and SPM has been considered a hotspot of microbial activity because particles function as sites of intense heterotrophic metabolism (Azam et al., 2001). Interestingly, the highest levels of archaeal diversity found by Vieira and colleagues were detected in sites

located in the interface between sewage polluted freshwater and coastal seawater, indicating that water mixing patterns promote increases in archaeaplankton diversity. Clones obtained through 16S rRNA gene library construction were mostly affiliated with other uncultured environmental Archaea. Libraries were dominated by typical estuarine *Euryarchaeota* followed by *Crenarchaeota* but in the anoxic heavily polluted inner channel library, most clones were associated with *Thermoplasmatales* and anaerobic methanogenic archaeal groups. The libraries also presented an abundance of species known for their hydrocarbon degradation activities, revealing a diverse set of microorganisms with potential applications for the bioremediation of ecosystems subjected to oil contamination.

Clementino and co-workers analyzed the archaeal diversity among several sites that included the bays water, agricultural soil, wastewater treatment plant water, halomarine sediment and a landfill leacheate, located in the bay`s surroundings. 16S rRNA gene library clones revealed that wastewater samples were exclusively of the *Euryarchaeota* phylum, most of them affiliated with methanogenic Archaea. These high levels of pollutants in inner bay anoxic environments are promoting the growth of anaerobic Archaea with methane producing metabolism. Bay water samples presented clones of the *Euryarchaeota* and *Crenarchaeota* phyla affiliated with marine water samples (Clementino et al., 2007). It is likely that more archaeal phyla are present in this environment than could be detect by the low number of sequences obtained.

In 2010, Turque and colleagues published a work comparing the diversity of archaeal communities associated with the marine sponges *Hymeniacidon heliophila, Paraleucilla magna* and *Petromica citrina*, from Guanabara Bay with those from an offshore less impacted environment, the Cagarras Archipelago. The diversity of the *amoA* gene was also analyzed. This gene belongs to the operon *amo* that encodes the catalytic subunit of ammonia oxygenase, which catalyzes an ammonia oxidation reaction being therefore extremely relevant in the nitrogen biochemical cycling with potential applications in bioremediation processes. Results showed that although the two sampling sites are located very close to each other, Guanabara Bay waters presented higher levels of NH_3, total phosphorus, chlorophyll a, prokaryotic abundance and production when compared to the Cagarras Archipelago. No operational taxonomic units (OTUs) were shared between water samples of the two sites, indicating that differences in the levels of pollutants drastically change the archaeal communities of those environments. Some OTUs were also related to the archaeon *Methanoplanus petrolarius*, a species associated with petroleum contaminated anoxic environments, indicating a possible contamination of Guanabara Bay by petroleum spills. Another relevant finding was that sponges of the same species separated by a short distance presented distinct archaeal communities suggesting that environmental conditions can shape these sponge associated microbial communities. The *amoA* sequence diversity within sponge associated samples indicated that this gene can play a central role in the adaptation of sponges to eutrophicated environments, which usually present high levels of ammonia. This raises the hypothesis that sponges harboring a microbiota more capable of performing detoxification of their tissues would be more fitted to survive in an environment with high ammonia levels such as Guanabara Bay (Turque et al., 2010).

In 2011, Gonzalez and colleagues published the analysis of *amoA* and *nifH* gene diversity from bacterial communities Guanabara Bay water samples. The *nifH* gene encodes one of the nitrogenase protein complex subunits present in diazotrophic microorganisms, capable of reducing atmospheric nitrogen to ammonia. Most *amoA* sequences were related to uncultured environmental organisms, which probably correspond to new species that

participate in ammonia oxidation. A similar pattern was observed for the *nifH* gene. The diversity of nitrogen fixing and nitrifying activity genes of these bacterial communities brings to light a complex set of microbial metabolic routes associated with nitrogen biochemical cycling in Guanabara, which can be extrapolated to other coastal estuarine ecosystems. Future studies are still necessary for a better characterization of the pollution impact of nutrients in this and other polluted ecosystems to expand the knowledge in the dynamics of urban estuarine biochemical cycles to understand how these impacted environments can function to remediate themselves (Gonzalez et al., 2011).

Mangroves are ecologically important ecosystems located in the interface between terrestrial, freshwater and seawater environments. Guanabara Bay harbored an extensive mangrove system which suffered significant reduction along the years due to anthropogenic impact. Very little is known about the microbial biodiversity in these habitats.

A 16S rRNA gene DGGE study of the bacterial communities in sediments from three distinct sites in Guanabara Bay's mangrove system, all subject to distinct levels of heavy hydrocarbon pollution, revealed that the bacterial microbial communities from the three sites were significantly different. This provides evidence that each mangrove site harbors distinct bacterial community patterns, which could be the result of the different levels of hydrocarbon contamination that each sampled site is subjected to (Gomes et al., 2008). When some of the DGGE bands were sequenced a diverse set of species that are capable of metabolizing hydrocarbons was revealed, illustrating the potential for bioremediation of environments subjected to oil spills.

Sediments of three mangrove forests from Guanabara Bay presenting distinct levels of PAH contamination (one of them near a petrochemical refinery) were analyzed regarding the diversity of genes encoding the multicomponent enzyme system naphthalene dioxygenase (NDO) that initiates the degradation metabolism of low molecular weight polycyclic aromatic hydrocarbons (PAH),). They revealed that PAH pollution is capable of shaping NDO gene diversity within these microbial communities (Gomes et al., 2007).

Hydrocarbon degrading bacterial communities from Guanabara Bay mangrove sediments were characterized utilizing mesocosm systems by Brito and colleagues in 2006. In this study, bacteria obtained from bay's mangrove sediments were inoculated into *in situ* 700 cm^2 surface mesocosms systems installed in the Guapimirim mangrove. Mesoscosms received 350 ml of petroleum and after 130 days of incubation bacteria retrieved from the mesocosms were isolated and inoculated in culture media containing one of five different hydrocarbons (octadecane, pristane, naphthalene, pyrene or fluoranthene) as a sole carbon source. Measurements of hydrocarbon degrading activities revealed that several isolates were capable of degrading hydrocarbons and some of them presented degrading activity of up to four kinds of aliphatic or aromatic hydrocarbons (Brito et al., 2006).

Sequencing of 16S rRNA genes from isolates showed that bacteria retrieved belonged to the following groups: *Gammaproteobacteria*, with a large number of isolates associated with genera known for their hydrocarbonoclastic activities; *Alphaproteobacteria*, of which most of the bacterial isolates were affiliated with hydrothermal vent strains, raising the possibility that these environments may harbor organisms with potential oil degrading activities; and *Actinobacteria*, which were affiliated with genera that due to their metabolic flexibility are capable of degrading different kinds of materials, like rubber, oils and hydrocarbons. More studies are still required to describe how microbial communities respond to hydrocarbon pollution in mangroves and other environments, especially regarding archaeal groups which present an unexplored set of metabolic activities.

These findings described above suggest that Guanabara Bay microbial communities are strongly affected by the high levels of anthropogenic impact to which this environment is subjected. The input of organic and inorganic nutrients promotes shifts in the diversity and abundance of microbial species that, in turn, are involved in biogeochemical cycles.
Different sites from Guanabara Bay present distinct levels of bacterial production and abundance (Vieira et al., 2008), which can be associated with the impact levels that each of those sites is subjected. To further understand how pollution affects the metabolism of these microbial communities a better understanding of the complex biochemical pathways that undergo in the Bay is still necessary therefore new generation sequencing techniques may play an important role in the process of expanding knowledge in this area in the near future.
The degradation of Guanabara Bay and other aquatic ecosystems seems to promote the emergence of opportunistic pathogenic species, suggesting that pollution of aquatic environments represents a direct threat to public health and stresses the importance of preserving ecosystems for maintaining the quality of life of humans and other species in the planet. Opportunistic pathogens are common in polluted environments but a further description of pathogenic bacteria in the bay is still required, with special attention to antibiotic resistant bacteria, which endanger the millions of people living in the bay's surroundings. The microbial communities of the Bay's aquatic and surrounding mangrove ecosystems are repeatedly affected by hydrocarbon pollution due to nearby petrochemical facilities. Coincidently, the resident communities at these same sites seem to harbor the metabolic machinery necessary to remediate those impacts, increasing the relevance of studying their diversity more deeply. Hydrocarbon pollution is a global ecological issue and metagenomics can provide insights into the effects of hydrocarbon contamination on microbial communities worldwide and consequently help in the development of bioremediation strategies to bypass this common kind of environmental degradation.

3.1.2 Worldwide metagenomic studies of polluted habitats

Metagenomics has been applied to several studies of aquatic pollution worldwide to understand how the presence of pollutants affects microbial community composition and ecology. An example is the characterization of archaeal and bacterial sediment communities from two distinct portions of Western Europe's largest freshwater reservoir, Lake Geneva in Switzerland (Haller et al., 2011).
The Bay of Vidy is the most contaminated area of the lake: it is subjected to discharges from a wastewater treatment plant and shows the higher levels of nutrients and heavy metals while the Ouchy area is a nearby less polluted site. The sites were compared based on 16S rRNA gene library constructions from environmental DNA. Rarefaction analysis showed a higher number of bacterial OTUs in the Ouchy site, suggesting a higher diversity. Most of the retrieved bacterial sequences were associated with *Proteobacteria* and *Bacteroidetes* at both sites. However, phylogenetically distinct OTUs from these two bacterial groups were found at each site. A Multiple Factor Analysis (MFA) indicated that differences in bacterial community composition were statistically correlated to differences in the levels of organic matter, nutrients and heavy metals. All archaeal sequences belonged to the Euryarchaeota division. Several clones from the Vidy site were associated with archaea that present methanogenic metabolism, which can be associated with organic matter degradation in this polluted site. This work showed that microbial communities from nearby sites can suffer compositional changes due to treated sewage contamination, indicating not only that those communities are extremely sensible to alterations in environmental conditions but also that even after treatment sewage drastically affects species diversity within aquatic ecosystems.

Huang and colleagues (2011) published a metagenomic analysis of heavily polluted small streams in China. As those environments seem to be more sensible to pollution than large masses of water, the description of how they are affected by anthropogenic impacts is extremely relevant. The authors performed a DNA profiling technique based on Terminal Restriction Fragment Length Polimorfism (T-RFLP) in parallel to construction of bacterial 16S gene libraries to analyze three distinct streams subjected to industrial, agricultural and urban pollution. T-RFLP analysis showed significant differences in community composition between the three sites, even though physicochemical parameters were very similar among them. The results suggest that a large number of other factors that go beyond nutrient levels may be responsible for shaping the composition of the bacterial communities (Huang et al., 2011).

Gene libraries showed that *Betaproteobacteria* were widespread through all the three streams, although divisions such as *Alpha* and *Gamma-Proteobacteria*, *Bacteroidetes* and *Cyanobacteria* had distinct distribution patterns between the three samples, suggesting that the proportion between these taxa varies between streams. Furthermore, clones affiliated with several genera that have been previously associated with polluted environments such as *Flavobaterium*, *Rhodobacter* and *Hydrogenophaga* were retrieved. In addition to an evaluation of 16S rRNA gene libraries, this study presents a sequencing independent metagenomic analysis based on T-RFLP using environmental DNA. This technique can be applied for a crude characterization of microbial communities, with potential applications for biomonitoring strategies of aquatic environments. However, sequencing based techniques provide a deeper understating of bacterial community composition even though they are more expensive and time consuming.

Contamination of aquatic environments by sewage originated pollution poses a threat to human health. Fecal bacteria may infect humans that consume contaminated water for recreational, feeding or drinking purposes. Although quantification of *Enterobacteria* has been extensively used for water quality analysis, many species of this bacterial group can colonize different host species. Therefore, this analysis provides poor information about the origin of pollution (i.e. whether its source is human feces or fecal bacteria of other animals).

Wéry and colleagues (2010) studied the human specific fecal bacteria originated from wastewater treatment plant effluents focused on the following groups: *Bacterioidales, Clostridiales, Bifidobacteria*, and the *Bacillus-Streptococcus-Lactobacillus* (BSL) cluster. Construction of V6 region of the 16S rRNA gene libraries from genomic material of effluents from five French wastewater treatment plants was performed. A set of specific primers to target the selected bacterial groups were utilized (Wéry et al., 2010). Besides library construction, an analysis based on Capillary Electrophoresis Single Stranded Conformation Polymorphism (CE-SSCP) was performed to characterize the profiles of the four bacterial groups between samples, which suggested an smaller diversity of *Bifidobacterium* within samples when compared to the three other bacterial groups. This can possibly be accounted for by the small number of species from this group that is part of the human gut microbiota.

Sequences obtained were affiliated with bacteria originated from feces, although some groups of *Bacteroides* commonly found in the human gut could not be retrieved from environmental samples, probably because some species are not fitted to survive in the wastewater environment and are more adapted to a host associated lifestyle.

Comparisons of bacterial species diversity between the wastewater treatment plants showed that the species profile of *Bacteroides, Clostridiales, Bifidobacterium* and BSL cluster is very similar, suggesting that specific bacteria from those groups are ubiquitous in wastewater treatment systems effluents. To identify the putative source of bacterial 16S rRNA gene sequences obtained through metagenomics the authors utilized a bioinformatics tool to

associate sequences retrieved from samples to previously obtained fecal samples and observed that all *Bifidobacterium* and *Clostridiaceae* were associated with database sequences of fecal source while the other two groups were affiliated with sequences of fecal and non-fecal source. Overall, the results indicate *B. adolescentis, B. caccae, L. pectinoschiza and H. filiformis* as potential indicators of human fecal contamination in aquatic environments.

Although the exclusivity of those bacteria as human gut associated organisms is debatable, this work provides insights into the development of molecular techniques based on metagenomics and bioinformatics analysis, for fecal source tracking on aquatic environments, which can shed light into the quantification of impacts caused by wastewater effluents to which environments are subjected.

Bacterial resistance to antibiotics has recently become a public health issue. Aquatic environments are ecosystems where bacteria can easily exchange antibiotic resistance genes through horizontal gene transfer, promoting the spread of antibiotic resistant bacteria (Martinez, 2008). Analyses conducted in the Seine River in France and in the Zenne and Scheldt rivers in Belgium, subjected to urban and industrial sewage discharge, were performed focusing on antibiotic resistant bacteria, revealing an alarming truth for public health. Those environments presented a diverse set of heterotrophic bacteria that possess antibiotic resistance, including genera of important human pathogens (Garcia-Armisen et al., 2010). The 16S rRNA gene libraries showed that among bacteria that presented multiple resistance to several antibiotics obtained from sewage-contaminated rivers, most of them were affiliated with the *Bacterioidetes* or *Proteobacteria* phyla, reflected by the isolation of multi-resistant bacteria.

Based on the abundance of sequences associated with pathogenic bacteria, the authors raised the awareness to the health risks associated with polluted rivers, which could function as hotspots of dissemination of antibiotic resistance and as environments that promote shifts in bacterial behavior towards a pathogenic state. Additionally, the analysis of a 16S gene library originated from water samples of the heavily impacted Zenne river, revealed the severe impact to which microbial communities undergo upon sewage contamination, suffering drastic compositional changes that may result in alterations in ecosystem functioning.

Heavy metal contamination is a common form of aquatic pollution worldwide. Those compounds are toxic to living organisms and also present mutagenic activity. Microbial communities are capable of interacting with these compounds, consequently affecting their availability and impact in the environment. Rastogi and colleagues (2011) studied the Couer d'Alene river, located in the United States as a model environment of severe metal contamination. Metagenomic analysis integrating gene library construction and a microarray based technique, the PhyloChip were used to describe microbial communities from this site (Rastogi et al., 2011).

Couer d'Alene River has suffered an impact of 125 years of acidic ore mining, which resulted in extremely high levels of metal in its waters, that include As, Cr, Cu, Ni, Pb and Zn. Ribosomal small subunit gene libraries revealed only a modest bacterial diversity in this environment, as suggested by rarefaction curves, but the PhyloChip approach was capable of covering a much larger portion of the bacterial diversity which reached 40 phyla while sequences retrieved by library construction retrieved only 6 phyla. Most OTUs originated from river sediment samples were associated with the Phylum *Proteobacteria*, followed by *Firmicutes*, *Actinobacteria, Bacteroidetes, Acidobacteria* and *Chloroflexi.* In addition, several other phyla with a smaller number of OTUs could only be detected by the PhyloChip. Several OTUs obtained in this study were affiliated with sequences obtained from samples subjected to heavy metal pollution but also with soil and aquatic samples that were not subjected to such impacts.

Additionally, several OTUs where associated with microorganisms whose metabolism is involved in processes of heavy metal bioavailability and mobilization. The authors also explored the diversity of the *amoA* and *mcrA* genes, associated with microbial ammonia oxidation and methanogenic metabolism, respectively, through library construction. Libraries produced a small number of OTUs evidencing that few species harboring such genes are present in this environment. While the *amoA* gene library was dominated by species of the ammonia oxidizing genus *Nitrosospira*, the *mcrA* gene library was dominated by methanogenic species of the genus *Methanosarcina*, evidencing an important role of archaea in the process of methane production within the river. Although little information could be retrieved from the *amoA* and *mcrA* gene libraries, they provide the first insights into the characterization of the roles of ammonia oxidizers and methanogenic microorganisms in this environment.

Interestingly, the *amoA* and *mcrA* genes are associated with very distinct environments. The former commonly occurs in strictly aerobic microorganisms while the latter is usually associated with anaerobic metabolism. So, the apparent paradox of both genes being retrieved in the same environments is still to be elucidated. This work is a good example of how distinct metagenomic techniques can complement each other. Since microarrays rely on previously known sequences it cannot be applied for detecting new species. Furthermore, this technique provides very little information concerning the composition of DNA sequences, revealing only if they are present or absent. On the other hand, the detection of novel taxa can be reached by 16S rRNA gene library construction. One limitation is that rare groups tend to be neglected by this approach, even with a very large sequencing effort.

Hydrocarbon pollution is a global issue, oil spills cause contamination of aquatic environments affecting living organisms in most areas of the world. Efforts to avoid, reduce, and remediate those continuous and/or acute contaminations are extremely necessary. Some advancement in this subject, however, has been recently achieved. For example, Marcos and colleagues (2009) identified bacterial populations in sub-Antarctic marine sediments, subjected to severe hydrocarbon pollution, capable of degrading polycyclic aromatic hydrocarbons (PAH). These findings revealed potential bioremediation strategies based on bacterial oxygenase genes obtained from those environments. In this work, samples were retrieved from the Ushuaia Bay, Argentina, a region subjected to constant small accidental hydrocarbon spillages, due to loading and unloading of petroleum-derived substances in a nearby pier. Using metagenomic DNA, construction of gene libraries were performed using primers designed for the gram-negative bacterial dioxygenase gene (Marcos et al., 2009).

The dioxygenase is the catalytic subunit of a multicomponent enzyme complex that catalyzes the insertion of molecular oxygen into benzene rings, an important step in PAH degradation pathways. Phylogenetic analysis of the dioxygenase protein sequences deduced from these gene sequences revealed a diverse set of enzymes in Ushuaia Bay's bacterial communities. Results revealed important biological indicators of PAH pollution and highlight the potential applications of metagenomics in the search for enzymes to be used for the development of PAH bioremediation strategies.

An recent evaluation of the bacterial diversity within mangrove sediments revealed several groups sensible to oil contaminations, which represent potential oil spillage indicator organisms in ecosystems (Santos et al., 2011).

In this work, bacterial communities in microcosms containing mangrove sediments exposed to 2% and 5% v/w of fuel oil were analyzed prior to oil exposure and after 23 and 66 days of oil contamination. Pyrosequencing of partial 16S rRNA gene sequences revealed that bacterial communities from all samples were dominated by the *Proteobacteria* phylum, mostly of the *Gamma*, *Delta* and *Alpha-proteobacteria* classes. Interestingly, petroleum contaminated

microcosms presented higher indexes of species richness when compared to their control samples. Additionally, the authors observed significant shifts in bacterial genera abundance associated with hydrocarbon degradation (*Alcanivorax*) and hydrocarbon contaminated environments (*Marinobacterium*, *Marinobacter, Clostridium,* and *Fusibacter*). On the other hand, an intense decrease in the abundance of sequences associated with the genera *Helia* after hydrocarbon contamination indicated that this group seems to be very sensitive to oil pollution.
A similar study focused on the diversity of microeukaryotes retrieved from the same mangrove system within microcosms subjected to the same levels of oil contamination. DGGE and 18S rRNA gene library construction revealed organisms within the microeukaryote group that could be used as potential bioindicators in the monitoring of oil pollution (Santos et al., 2010).
Results suggested that Fungi and Metazoa, the originally dominant group in the mesocosm, are the most sensitive to oil contamination, being replaced by *Stramenopiles* as the most abundant group after oil contamination. Therefore, the balance between those two groups can reflect the level of oil impact to which an ecosystem is submitted. Interestingly microeukaryote species richness and diversity indexes were lower in the oil contaminated mesocosm, in opposition to the prokaryotic pattern of higher diversity and richness in contaminated mesocosm.
Determining the reason why these communities behave differently upon oil contamination will shed light on the variables that regulate microbial community composition on polluted environments and possibly establish new ecological relationships by which prokaryotic and microeukaryotic populations affect each other in pristine and polluted ecosystems. Also, the comparison of microbial communities in microcosms subjected to oil contamination in different time periods suggests that shifts in prokaryotic and microeukaryotic communities are time-dependent, undergoing distinct alterations according to the time passed since contamination. Altogether these results provide precious information to be used in the development of bioindicators based on microbial diversity, abundance and community composition. Even though those results are very promising, further study is required to determine if microbial communities respond to oil contamination in a similar pattern as they respond in microcosm environments before reliable biomonitoring strategies can be developed.

4. Conclusions

We now know that pollution severely affects microbial communities indicating how fragile aquatic biota can be, so it is necessary to explore how exactly those impacts affect microorganisms directly (e.g. characterizing which of pollutants cause death of organisms and which ones promote slowing of growth). Rarefaction curves generated from metagenomic analysis indicate that a large amount of the diversity of Bacteria and Archaea is not being covered, calling for further efforts to fully identify these organisms, especially when focusing on rare groups.
New generation sequencing techniques seem to be the option for a complete access to Earth's Biota. Also, due to the capability of those methodologies to generate very large datasets, they are more fitted for quantitative inference, which is only poorly explored through library construction due to the smaller number of sequences that is generated through this method when compared to these new generation sequencing technologies.
Future studies focusing on microbial communities from aquatic environments subjected to anthropogenic pollution, to determine their composition, structure, metabolic capacities,

and ecological relationships are still necessary for a deep understanding of how those complex environments function and how the severe impacts to which they subjected can be reversed or at least attenuated. Function driven metagenomics, focused on specific genes associated with biogeochemical cycles (e.g. *amo* genes) contribute to a more detailed interpretation of the data gathered so far, so scientists can know not only which microorganisms are living in impacted environments but also how they are behaving biochemically. Furthermore, the utilization of metagenomic techniques will help to determine what effects pollution produces in aquatic microorganisms in a global level, specifying which alterations are common to all impacted environments and which ones are associated with a specific ecosystem, pollutant or a microbial group.

A great number of polluted sites where microbial diversity was explored through metagenomics are subjected to more than one kind of pollutant (industrial, wastewater, agricultural, etc.) Analyzing similar sites subjected to different pollutants will contribute to the understanding of which pollutants are more aggressive against environmental microbial communities and what sorts of alterations each one of them is capable of promoting.

For example, several studies suggest that while some sorts of pollutants, such as heavy metal and hydrocarbons, tend to alter or decrease microbial diversity, sewage contamination usually produces increases in microbial diversity. Sequences originated from polluted environments are often associated with pristine sources, indicating that those sites may harbor an important microbiota capable of surviving and growing in heavily polluted sites. Further exploring the diversity of organisms in these ecosystems will provide relevant insights for biomonitoring strategies as these microorganisms may present important metabolic traits that can be applied for bioremediation. Therefore, pristine environments are as much important as polluted ones in the development of biotechnology strategies, and thus require further efforts to elucidate their microbiota.

Certainly, an enormous amount of species, genes, proteins, enzymes, and metabolic pathways are still to be discovered. Guanabara Bay has been shown to be a potential site harboring those unexplored biological units, which presents potential applications for biomonitoring, bioremediation and even sanitary policies.

Metagenomic analysis showed that *Proteobacteria* are widely spread throughout polluted environments, therefore, further studies focused on this phylum will contribute to our understanding of the anthropogenic impacts in aquatic environments. The study of the behavior of microbial communities in Guanabara Bay during and after the cleansing program to which it is being submitted can provide insights into how aquatic life in heavily polluted aquatic environments responds to attempts to revert impacts, to determine if such an attempt is successful and to quantify the extent of the damage that can be reversed. Due to its characteristics, Guanabara Bay could be used as a model ecosystem for this sort of analysis. Therefore, Guanabara Bay has an unexplored biological and biotechnological richness requiring further research, specially focusing on microbial groups that are poorly studied such as environmental microeukaryotes, viruses and archaeas. For these future studies metagenomic approaches will surely be indispensable and will certainly contribute much valuable information.

5. Acknowledgement

This work was partially funded by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

6. References

Azam F & Long RA. (2001). Sea snow microcosms. *Nature.* 414:495-498.

Brito EM, Guyoneaud R, Goñi-Urriza M, Ranchou-Peyruse A, Verbaere A, Crapez MA, Wasserman JC & Duran R. (2006). Characterization of hydrocarbonoclastic bacterial communities from mangrove sediments in Guanabara Bay, Brazil. *Researh in Microbiology*. 157:752-762.

Cardoso AM, Vieira RP, Paranhos R, Clementino MM, Albano RM & Martins OB. (2011). Hunting for extremophiles in rio de janeiro. *Front Microbiol*. 2:100.

Cardoso AM, Clementino MM, Vieira RP, Cavalcanti JJV, Albano RM & Martins OB. (2010). "Archaeal metagenomics: bioprospecting novel genes and exploring new concepts" in Metagenomics: Theory, Methods, and Applications, ed. D. Marco (Wymondham: Caister Academic Press), 159–169.

Clementino MM, Fernandes CC, Vieira RP, Cardoso AM, Polycarpo CR & Martins OB. (2007). Archaeal diversity in naturally occurring and impacted environments from a tropical region. *Journal of Applied Microbiology*. 103:141-151.

Eyers L, George I, Schuler L, Stenuit B, Agathos SN & El Fantroussi S. (2004). Environmental genomics: exploring the unmined richness of microbes to degrade xenobiotics. *Applied Microbiology and Biotechnology*. 66:123-130.

FEEMA. (1998). Qualidade de água da Baía de Guanabara 1990/1997, *Fundação Estadual do Meio Ambiente*. pp. 100.

Garcia-Armisen T, Vercammen K, Passerat J, Triest D, Servais P & Cornelis P. (2011). Antimicrobial resistance of heterotrophic bacteria in sewage-contaminated rivers. *Water Research*. 45:788-796.

Gilbert GA & Dupont CL. (2011). Microbial metagenomics: beyond the genome. *Annual Review of Marine Science*. 3:347-371.

Gomes NC, Borges LR, Paranhos R, Pinto FN, Krögerrecklenfort E, Mendonça-Hagler LC & Smalla K. (2007). Diversity of ndo Genes in Mangrove Sediments Exposed to Different Sources of Polycyclic Aromatic Hydrocarbon Pollution. *Applied and Environmental Microbiology*. 73:7392–7399.

Gomes NC, Borges LR, Paranhos R, Pinto FN, Mendonça-Hagler LC & Smalla K. (2008). Exploring the diversity of bacterial communities in sediments of Urban mangrove forests. *FEMS Microbiology Ecology*. 66:96-109.

Gonzalez AM, Vieira RP, Cardoso AM, Clementino MM, Albano RM, Mendonça-Hagler L, Martins OB & Paranhos R. (2011). Diversity of bacterial communities related to the nitrogen cycle in a coastal tropical bay. *Molecular Biology Reports*. DOI: 10.1007/s11033-011-1111-9.

Haller L, Tonolla M, Zopfi J, Peduzzi R, Wildi W & Poté J. (2011). Composition of bacterial and archaeal communities in freshwater sediments with different contamination levels (Lake Geneva, Switzerland). *Water Research*. 45:1213-1228.

Huang Y, Zou L, Zhang S & Xie S. (2011). Comparison of Bacterioplankton Communities in Three Heavily Polluted Streams in China. *Biomedical and Environmental Sciences*. 24:140-145.

Lozupone C & Knight R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*. 71:8228-8235.

Malik S, Beer M, Megharaj M & Naidu R (2008). The use of molecular techniques to characterize the microbial communities in contaminated soil and water. *Environment International*. 34:265-276.

Marcos MS, Lozada M, Dionisi HM. (2009). Aromatic hydrocarbon degradation genes from chronically polluted Subantarctic marine sediments. *Letters in Applied Microbiology*, 49:602-608.

Martinez, JL. (2008). Antibiotics and antibiotic resistance genes in natural environments. *Science.* 321:365-367.

Mayr LM, Tenembaum DR, Villac, MC, Paranhos R, Nogueira CR, Bonecker SLC & Bonecker (1989). Hydrobiological characterization of Guanabara Bay. In Coastlines of Brazil, eds. O. Magoon & C. Neves. American Society of Civil Engineers. Charleston July, 1989.

Rastogi G, Barua S, Sani RK & Peyton BM. (2011). Investigation of Microbial Populations in the Extremely Metal-Contaminated Coeur d'Alene River Sediments. *Microbial Ecology*. 62:1-13.

Paranhos R & Mayr LM. (1993) Seasonal patterns of temperature and salinity in Guanabara Bay, Brazil. *Fresenius Environmental Bulletin*. 2:647-652.

Santos HF, Cury JC, Carmo FL & Rosado AS, Peixoto RS (2010). 18S rDNA sequences from microeukaryotes reveal oil indicators in mangrove sediment. *PLoS ONE*. 5:12437.

Santos HF, Cury JC, Carmo FL, dos Santos AL, Tiedje J, van Elsas JD, Rosado AS & Peixoto RS (2011). Mangrove bacterial diversity and the impact of oil contamination revealed by pyrosequencing: bacterial proxies for oil pollution. *PLoS ONE*. 6:16943.

Schloss P & Handelsman J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology*. Vol.71, No.3 (May 2004), pp. 1501-1506, ISSN 0099-2240

Schloss P, Westcott S, Ryabin T, Hall J, Hartmann M, Hollister E, Lesniewski R, Oakley B, Parks D, Robinson C, Sahl J, Stres B, Thallinger G, Van Horn D, Weber C. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*. 75:7537-7541.

Singh J, Behal A, Singla N, Joshi A, Birbian N, Singh S, Bali V & Batra N. (2009). Metagenomics: Concept, methodology, ecological inference and recent advances. *Biotechnology Journal*. 4:480-494.

Thompson FL, Bruce T, Gonzalez A, Cardoso A, Clementino M, Costagliola M, Hozbor C, Otero E, Piccini C, Peressutti S, Schmieder R, Edwards R, Smith M, Takiyama LR, Vieira R, Paranhos R & Artigas LF. (2011). Coastal bacterioplankton community diversity along a latitudinal gradient in Latin America by means of V6 tag pyrosequencing. *Archives of Microbiology*. 193:105-114.

Turque AS, Batista D, Silveira CB, Cardoso AM, Vieira RP, Moraes FC, Clementino MM, Albano RM, Paranhos R, Martins OB & Muricy G. Environmental Shaping of Sponge Associated Archaeal Communities. *PLoS ONE*. 5:15774.

Vieira RP, Clementino MM, Cardoso AM, Oliveira DN, Albano RM, Gonzalez AM, Paranhos R & Martins OB. (2007). Archaeal Communities in a Tropical Estuarine Ecosystem: Guanabara Bay, Brazil. *Microbial Ecology*. 54:460-468.

Vieira RP, Gonzalez AM, Cardoso AM, Oliveira DN, Albano RM, Clementino MM, Martins OB & Paranhos R. (2008). Relationships between bacterial diversity and environmental variables in a tropical marine environment, Rio de Janeiro. *Environmental Microbiology*. 10:189-199.

Wéry N, Monteil C, Pourcher A-M & Godon J-J. (2009). Human-specific fecal bacteria in wastewater treatment plant effluents. *Water research*. 44:1873-1883.