Encuentro No. 93, 117-125, 2012

Overview of Metagenomics for Marine Biodiversity Research[1]

*Barton E. Slatko**

We are in the midst of the fastest growing revolution in molecular biology, perhaps in all of life science, and it appears to be speeding up. We still know very little about the vast diversity of micro-organisms, their metabolic pathways and microbial activity in natural environments. Modern genomic tools are providing deep access to natural microbial diversity and ecology. Interdisciplinary approaches will be required to fully understand microbial ecology by: (1) analysis of genomes, transcriptomes, proteomes and metabolomes and (2) analysis at various levels of individuals, populations, communities and ecosystems. Data gathered is not only theoretical. It holds the promise of practical applications in the control of infectious diseases, in the production of biotechnology goods and services and in environmental remediation. It is an incredibly exciting time in science for the newer generation of scientists, "loaded" with opportunities. It is an excellent time to develop and apply tools to solve problems of local and global importance.

Metagenomics defined

Metagenomics can be defined as the analysis of sequence and/or function of microbial genomes within a selected environmental sample. As such, the analysis represents the organismal diversity in the selected community. Over 340 studies and over 2000 samples have been performed (GOLD database, as of 08.2012), As described by Wooley et al (2010), "With the advent of metagenomics, we are now able to study the genomic potential of a bacterial community and how it is affected by and affects its habitat. Many metagenomic studies have looked to some extent at correlations between sequence data, environment and environmental attributes in an attempt to gain biological insight...."

---

1 This paper was given by the author at the VI Nicaraguan Biotechnology Conference (April 12, 2012).

* *Molecular Parasitology Division, DNA Sequencing Group, New England BioLabs 240 County Road, Ipswich MA 01938. Email: slatko@neb.com*

Because up to 95% of microbes in oceanic waters have not currently been able to be grown in the lab, sampling methods need to avoid a laboratory culturing step. Many tend to think of this as a novel concept, but the idea really goes back to Norm Pace, Ed DeLong and co-workers (Pace et al, 1985; Schmidt et a, 1991) in the late 1970s.

Measuring metagenomic diversity

Microbial diversity in the coastal and marine environments can be measured by various methods. These include species diversity and phylogenetic diversity. Marine microbial diversity is commonly quantified based on evolutionary distances based on a common genetic marker that evolves slowly through evolutionary time. Within a species, genotype diversity and gene diversity are commonly measured. All relate to the probability that two samples in a population will be different.

Decisions, decisions, decisions.....

When considering a metagenomic project, a number of decisions should be considered:

1. The concept: what to sample (microbiome?)
2. Funding: Does the project need to be hypothesis driven or can it be descriptive?
3. Where to sample
4. Sampling technique/contamination minimizing
5. Archiving samples
6. DNA purification/cloning or DNA sampling (labor, $$, equipment, lab space)
7. DNA sequencing (selected genetic loci or whole genome sequence (WGS)
8. Data storage and retrieval
9. Data analysis
10. Resampling/consistency
11. Comparative analysis
12. Use of data, publishing, advocacy

Genomic methods in marine metagenomic ("microbial ecology") research

A number of molecular methods have been, and are, being used to sample metagenomes (Ju, 2006). These include:

- Representational difference analysis (RDA)
- Polymerase chain reaction (PCR)
- DNA cloning systems (plasmid, lambda-phage, cosmid, bacterial artificial chromosome
- or BAC, yeast artificial chromosome or YAC)
- Denaturing gradient gel electrophoresis (DGGE)

- DNA re-association
- Fluorescent *in situ* hybridization (FISH)
- 2-D gel electrophoresis
- Mass spectrophotometry
- Microarray hybridization
- MLST (Multi-Locus Sequence Typing) analysis
- DNA sequencing

DNA sequencing as the method of choice or phylogenetic environmental monitoring

For phylogenetic purposes, it is obviously not necessary to sequence entire microbial genomes, but rather use selected information. After DNA isolation and shearing the genomic DNA into random fragments (hence the "shotgun" sequencing name), fragments can then be cloned into cloning vectors and propagated for DNA sequencing reactions using Sanger dye-deoxy terminator methods. Unless one is assembling genomes, one could screen for selected genes (e.g. ribosomal) by PCR, for example, either at this step or after DNA isolation, above.

The new paradigm shift which has occurred due to the development of "NextGen" sequencing requires no cloning but still utilizes DNA isolated from an environmental sample for library construction. In general, after massively parallel sequencing, informatic selection can be used to identify and utilize particular sequences for phylogenetic analysis (Figure 1; modified from Riesenfeld et al, 2004). The left side of the figure shows approaches which are slower, more time consuming, but are less expensive. Conversely, the right side of the figure describes technology that is more rapid, but technologically more complex. It should be noted that the concept of "filtering" can have two distinct meanings.

The physical filtering of environmental samples has two main goals:

(1) Obtaining as much as one can of what one wants

(2) Leaving out as much as one can from what one does not want.

Computational filtering, on the other hand, is used after sequencing. Informatics is used to search for the genes one wants for phylogenetic purposes and to eliminate obvious false positive sequence motifs. This technique can also be used to detect sample contamination.

Sample size and number of samples

The first step in a metagenomic study is to obtain the environmental sample. Samples should be representative the population from which they are taken and thus the method of sampling can be a critical component. Culturing should be avoided as standard culturing techniques account for 5% or less of the bacterial diversity in most environmental samples.

"How much sampling is enough?" is a question that must be answered early in the decision process. To estimate the fraction of species sequenced, a curve plotting the number of species as a function of the number of individuals sampled can be

used. The curve usually begins with a steep slope, which at some point begins to flatten as fewer species are being discovered per sample (Wooley et al, 2010).

Sequencing Ribosomal Genes

The approach of sequencing of ribosomal RNA genes (5S, 16S rRNA) has enabled generation of a set of culture-independent approaches to:

- reconstruct phylogenies
- compare microbial distributions among samples
- quantify the relative abundance of each taxonomic group

However, rDNA has been criticized when used as the only marker, and evidence of horizontal gene transfer involving rDNA may confound its reliability even more. Secondly, 16S rDNA may exist in multiple different sequence copies in a single bacterium; this would cause a variance in both the estimated individual bacterial count (the mean number of bacterial ribosomal operons per genome may vary between 1 and 15). Alternative markers, such as rpoB, amoA, pmoA, nirS, nirK, nosZ, and pufM genes have also been suggested and can be used as MLST sets using multi-alignment tools.

For individual gene/locus profiling, one strategy is to perform a PCR on total genomic DNA from the metagenomic sample. This provides a representative sample of that locus in the population, provided that the PCR primers are "universal" and the starting DNA material is representative. Because the PCR products represent a "pool", they should be cloned into sequencing vectors, transformed into bacteria to create "libraries" of colonies. Assuming no cloning or cell viability bias, the colonies should be representative of the starting DNA material. Sanger dideoxy DNA sequencing of the cloned inserts provides the basis for analysis.

A review of DNA Sequencing Technology

Standard Sanger dideoxy DNA is almost universally based on fluorescent detection in conjunction with capillary electrophoresis of fluorescently labeled dideoxy nucleotides ("Big Dye" energy transfer dyes). A dideoxy molecule and the "Big Dye" structure are shown in Figure 3.

A review of Sanger technology can be found in Slatko et al (2011).

For many metagenomic projects this technology has been replaced by "NextGen" technology (Shendure et al, 2011). Thus "roomfuls" of Sanger automated DNA sequencers have been replaced with one or a few of the newer technology machines.

A general description of "NextGen" technology is provided in Figure 4 and a description of the current major platforms is provided in Table 1. In the table, those technologies with "red stars" are most often used for metagenomic projects. Intranet links to the various technologies are provided in the reference list.

Informatic analysis after the sequencing enables the data to be analyzed for the particular parameters of interest (ribosomal gene sequences, for example). From this analysis, phylogenetic profiling can be performed and conclusions made as to

the distribution and make up of the metagenomic population, including frequency estimates of the various species. Figure 5 provides some examples of phylogenetic profiling (Yun et al, 2010; Luton et al, 2002). One can also perform a detailed sequence analysis of particular genes among the various species which were identified (Figure 6). For the detection of all known orders of methanogen, Luton et al. (2002) have described a methanogen-specific PCR approach. In that method *mcrA* and small subunit rRNA gene phylogenies were remarkably similar, validating their approach.

Conclusions

The oceans and coastal environments are extremely diverse as is the life in the seas, including microbial life. This diversity is being explored by traditional sequencing methods as well as through large scale DNA sequencing approaches. Through many different studies in the past few decades it has been found that that the oceans harbor unparalleled microbial diversity. Novel genes and proteins families have been discovered through the application of metagenomics in various marine environments, including extreme environments. Novel molecules of relevance to health and industrial biotechnology are being discovered from these uncultured and highly diverse microbial ecosystems. Young scientists from the tropics and particularly from Nicaragua and Central America may find a lot of joy in the application of metagenomics in their exploration of the ocean biome.

NextGen references: (some with videos of the technology)

- Ion Torrent: *www.iontorrent.com/*
- 454: *454.com/*
- SOLiD: *www3.appliedbiosystems.com/SOLiDSystemSequencing/*.
- Illumina: http://www.illumina.com/technology/sequencing_technology.ilmn
- PacBio: *www.pacificbiosciences.com/*

Figure 1: Sanger sequencing and NexGen sequencing

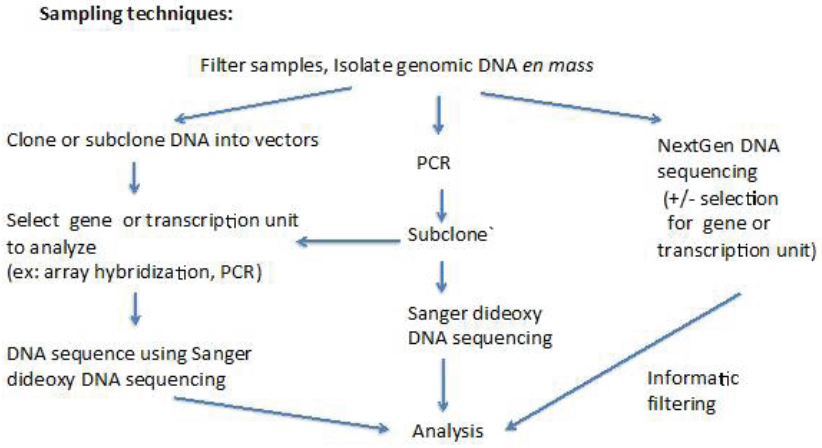


Figure 2: How much sampling is enough?

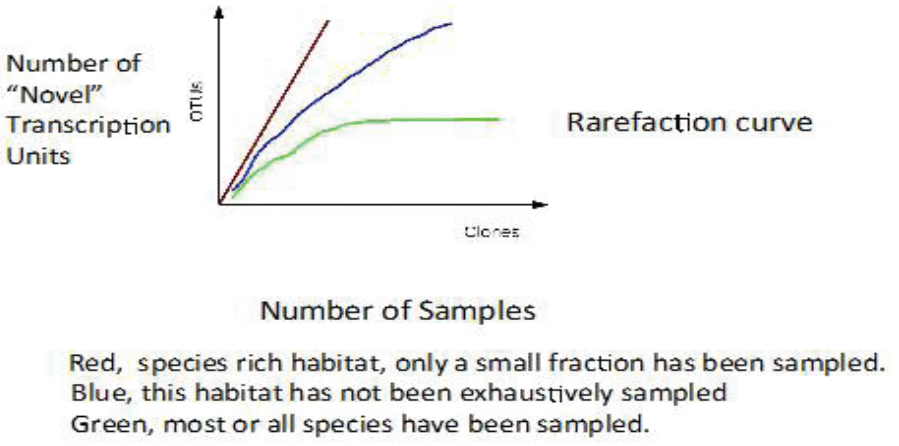


Figure 3: A dideoxy molecule and the "Big Dye" structure

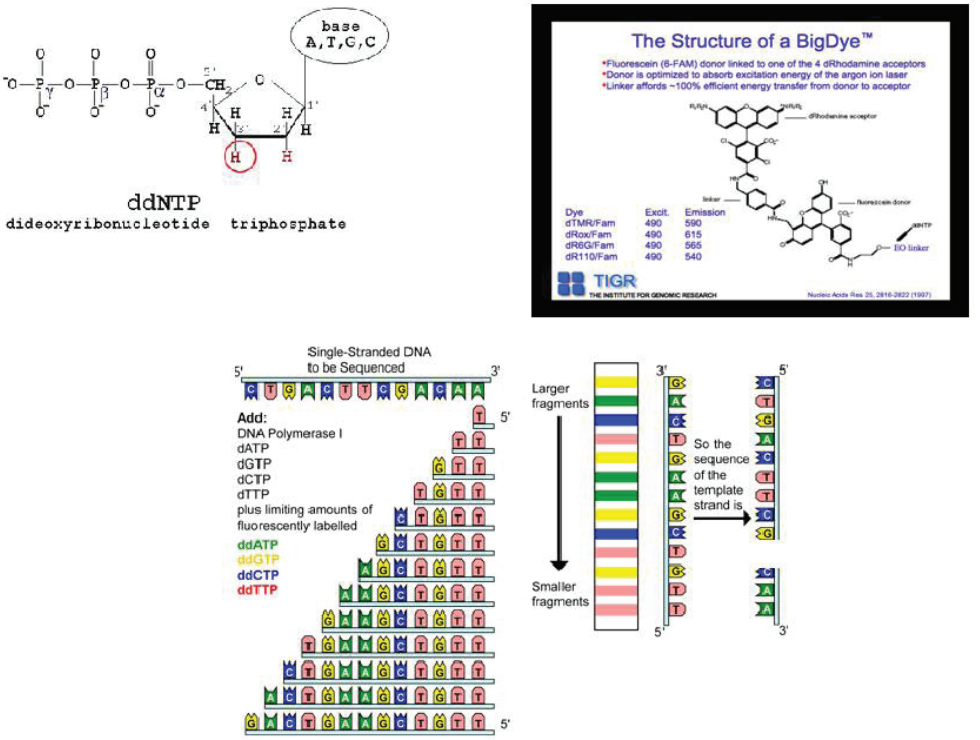
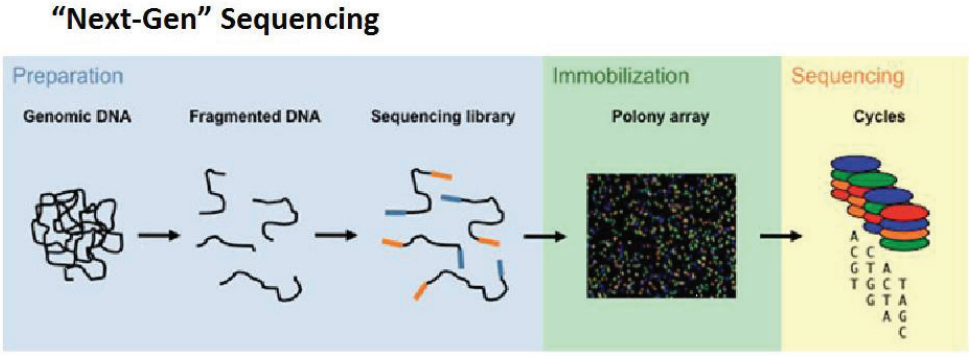


Figure 4: Next-Gen sequencing workflow.



General sequencing workflow: DNA preparation, immobilization, and sequencing.

-- random fragmentation of the genomic DNA
-- addition of adapter sequences to the ends of the fragments
-- immobilized DNA on a solid support to form detectable sequencing features
--massively parallel cyclic sequencing reactions to interrogate the nucleotide sequence

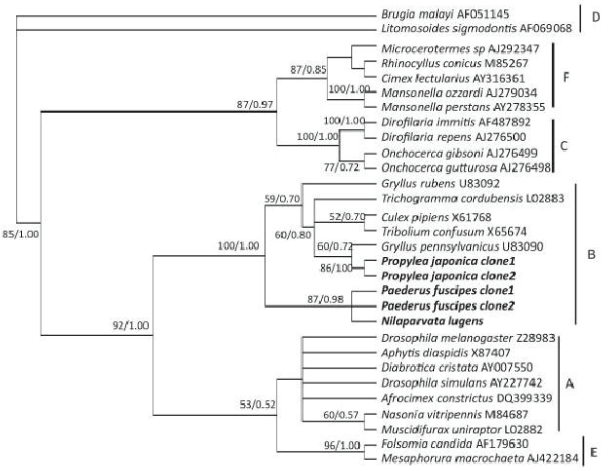
modified from : Samuel Myllykangas , Jason Buenrostro , and Hanlee P. Ji, 2012. Overview of Sequencing Technology Platforms, in, Bioinformatics for high throughput sequencing. N. Rodriguez-Ezpeleta, M. Hackenberg and A. M. Aransay (eds). Springer.

Table 1

"NextGen" Sequencing platforms

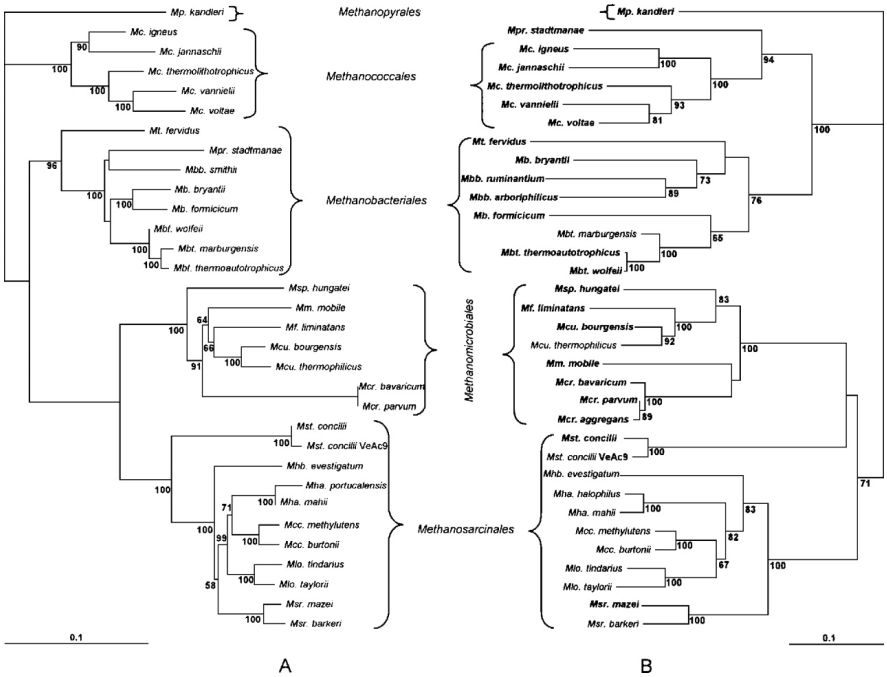
Platform	Support	Feature generation	Sequencing reaction	Detection method
GS FLX 454 Life Sci.	Picotiter plate	Emulsion PCR	Synthesis	Pyrosequencing
Genome Analyzer Illumina	Flow cell	Bridge PCR	Synthesis	Fluorophore labeled reversible terminator nucleotides
PGM Ion Torrrent Applied Biosystems	Flow cell	H+ ion release	Synthesis	Semiconductor
SOLiD Applied Biosystems	Flow cell	Emulsion PCR	Ligation	Fluorophore labeled oligonucleotide probes
CGA platform Complete Genomics	DNA nanoball arrays	Rolling circle amplification	Ligation	Fluorophore labeled oligonucleotide probes
PacBio RS Pacific Biosci.	Zero mode waveguide	Single molecule	Synthesis	Phospholinkedfluor o-phore nucleotides

Figure 5: An example of phylogenetic profiling of Wolbachia (Yun et al, 2010). The distribution of Wolbachia in arthropods was detected by diagnostic PCR amplification of the wsp (Wolbachia outer surface protein gene) and 16S rDNA genes.



Unrooted phylogeny of *16S* rDNA of *Wolbachia* reconstructed using the maximum likelihood (ML) method. The names of taxa are those of the hosts. Levels of confidence for each node are shown as bootstrap values. Trees inferred from Bayesian analyses were similar, and the posterior probabilities are shown following the bootstrap values from ML analyses. Sequences from this study are indicated in bold. *Wolbachia* supergroups (A-F) are indicated.

Figure 6: Phylogenetic tree showing the relationship between 16S rDNA sequences (A) and partial mcrA DNA sequences (B) of methanogens (Luton et al. 2002). No major differences were apparent between DNA and amino acid sequences, nor between the different algorithms used.



References

GOLD database: http://www.genomesonline.org/cgi-bin/GOLD/index.cgi

Luton, P., Wayne, J., Sharp, R. and Riley, P. 2002. The *mcrA* gene as an alternative to 16S rRNA in the phylogenetic analysis of methanogen populations in landfill. *Microbiology* 148, 3521–3530

Pace NR, Stahl DA, Lane DJ, Olsen GJ. 1985. Analyzing natural microbial populations by rRNA sequences. *ASM News* 51:4–12

Riesenfeld, C., Schlos, P. and Handelsman, J. 2004. Metagenomics: Genomic Analysis of

Microbial Communities. Annu. Rev. Genet. 2004. 38:525–52

Schmidt TM, DeLong EF, Pace NR. 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* 173:4371–78.

Shendure JA, Porreca GJ, Church GM, Gardner AF, Hendrickson CL, Kieleczawa J, Slatko BE. 2011. Overview of DNA sequencing strategies. Curr Protoc Mol Biol. 2011 Oct;Chapter 7:Unit7.1.

Slatko BE, Kieleczawa J, Ju J, Gardner AF, Hendrickson CL, Ausubel FM. 2011. "First generation" automated DNA sequencing technology. Curr Protoc Mol Biol.1Oct;Chapter 7:Unit7.2.

Wooley, J., Godzik, A. and Friedberg, I. 2010. PLoS Computational Biology 6:e1000667.

Doi:10.1371/journal pcbi.1000667).

Xu, J. . 2006. Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. Molecular Ecology 15: 1713–1731.

Yun, Y., Peng, Y., Liu, F., Lei, C. 2011. Wolbachia screening in spiders and assessment of horizontal transmission between predator and prey. Neotropical Ento.40: 152