# Microbiological Targets for Ocean Observing Laboratories (MicroTOOLs) 2010 Workshop II

Gordon and Betty Moore Foundation
Palo Alto, California
October 28-29, 2010

## Summary Report

MicroTOOLs seeks to generate comprehensive high throughput, and targeted *in situ* marine microbial community and biogeochemistry approaches for studying and monitoring microorganisms. The goal is to develop a scientific community resource, based on the expertise of the scientific community, and available for application in diverse studies. The MicroTOOLs I workshop (https://sites.google.com/site/microtoolsii/) was held to discuss the issues involved with implementing molecular approaches for high-throughput and *in situ* instrumentation, and was followed by the MicroTOOLs II hands-on workshop.

The goal of the second MicroTOOLs (II) workshop was to design a first generation set of high-throughput microarrays that can be used for temporal and spatial analysis of the biogeochemical functional composition of coastal and open ocean marine microbial communities. Investigators brought their expertise on individual target genes and organisms, and with this knowledge analyzed nucleic acid sequence datasets to select appropriate targets to be incorporated into a high density microarray. Participants used a workflow developed at UCSC in collaboration with CAMERA personnel to select all available homologue sequences of the gene(s) and organism(s) of interest from metagenomic and metatranscriptomic databases in CAMERA and other resources, to obtain comprehensive datasets on genes of biogeochemical importance or for identification of key microorganisms. The resulting gene datasets, which have been derived from BLAST searches of GenBank and the metagenomic/metatranscriptomic data at CAMERA should be all known marine microbial homologues for the target sequences.

The MicroTOOLS II objective is to design microarrays that can be used beginning in early 2011 to probe surface marine microbial communities from typical marine environments. The MicroTOOLs philosophy is a scientific community effort that draws on the collective expertise of the marine microbiology community to design a tool that is broadly applicable for characterizing marine microbial populations and is accessible to the general microbiology community. The focus of this approach is not microbial diversity directly, but microbial community function, targeting key biogeochemical transformations as well as key microorganisms.

The short-term output of this workshop and the subsequent collaborative work will be the design of high-density oligonucleotide (using Roche NimbleGen platform) functional microarrays for coastal and open ocean microbial communities, targeting prokaryotes and eukaryotes (for both microbial diversity and activity) available for use by marine microbiologists. The first of these arrays will be available in early 2011, although it is assumed that improvements and modifications will continue to be implemented following the testing of version 1 of the arrays. Implementation of the assay will be developed as a collaboration among the MicroTOOLs II participants, addressing sampling, sample processing, hybridization, and data analysis issues.

## Specific goals of MicroTOOLs II

I. Collect all extant environmental sequences for the genes of interest from available genomic/metagenomics/metatranscriptomic databases and choose representative sequences for targeting on coastal and/or open ocean microarrays

II. Discuss technical issues: space allocation, array validation, sample processing, and usage of the arrays by the scientific community.

III. Design timeline for implementation of arrays during 2011.

## Detailed Summary

I. Currently, there is enough sequence data for many relevant genes and organisms to design a comprehensive array for assessing major biogeochemical transformations. The sequences have

been obtained from sequencing genomes, Sanger or 454 sequencing of metagenomes and metatranscriptomes, and PCR amplification of environmental samples with degenerate primers. In order to select the appropriate target sequences that will represent diversity of a particular gene in the environment, the sequences should be collected from available databases, compared (aligned), and representatives that differ at least 5% at nucleotide level should be selected. A workflow to select appropriate target sequences was designed at UCSC based on a prototype marine microbial microarray, which used Roche NimbleGen platform. The main idea of the workflow is to use amino acid sequence of the gene of interest to perform a loose BLAST search (TBLASTN for searching nucleotide database with a protein sequence) against all metagenomic (includes metatranscriptomic) Sanger and 454 reads databases in CAMERA to collect all available sequences, then filter the reads that do not originate from the organism of interest or represent the gene of interest by doing reciprocal BLASTN/BLASTX of the collected reads against sequenced genomes and all proteins. After BLASTN/BLASTX, the sequences are clustered, and representatives are selected based on 5% or higher dissimilarity (Figure 1).

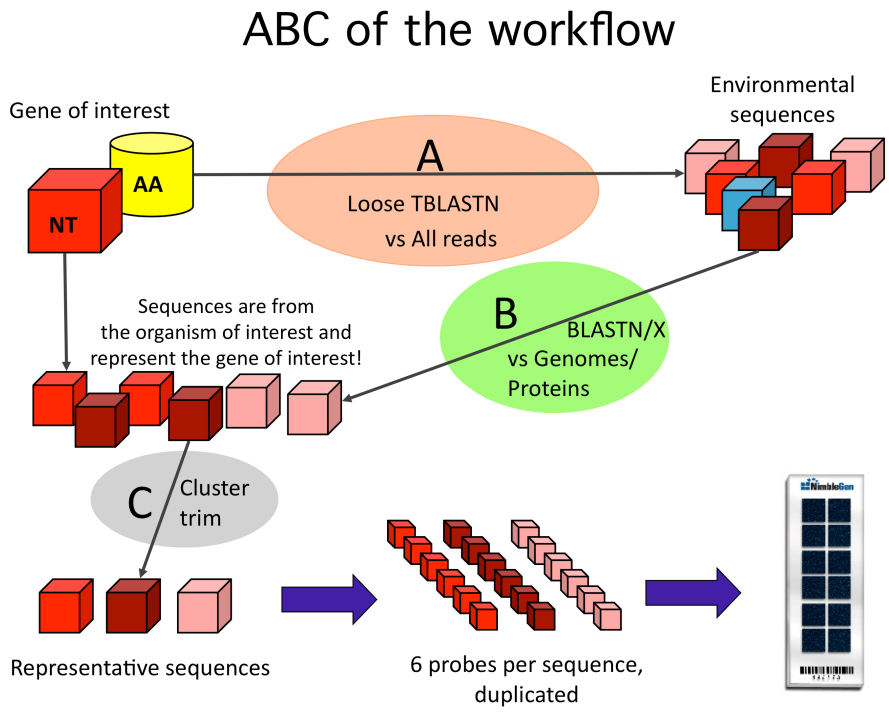


**Figure 1.** A scheme for collecting target sequences from environmental databases.

While acquiring and validating homologous sequences for a gene of interest are more or less automatic processes, choosing representative sequences and a region for targeting requires the knowledge and expertise of the investigators. Roche NimbleGen chip technology allows synthesis of 60-mer long oligonucleotides providing high sensitivity of the array. The probes are designed to a specific region and overlap each other, which increases the specificity of the hybridization so that the signals from up to 95% similar sequences can be distinguished. For this set of arrays, we decided to start with four-six probes per sequence group and a targeting region of at least 150 bp long (Figure 2). When selecting representative sequences and a region to be targeted with the probes within the gene, the following should be taken into consideration:

1. The data for targeted sequences will be meaningful and interpretable.
2. Geographical region where arrays will be applied.
3. Abundance and degree of conservation within the gene (depending on the gene of interest and why this gene is included in the array, either the most or the least conserved region in the gene sequence will be targeted).

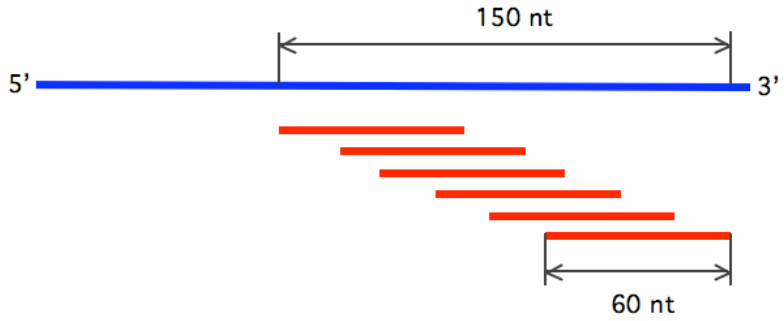


**Figure 2.** Probe design for a gene of interest.

II. Space allocation, array validation, sample processing, and usage of the arrays by the scientific community.

Once target sequences are selected, the next step is to allocate space for different organisms and processes to be targeted on open ocean and coastal chips. The Roche NimbleGen (http://www.nimblegen.com/) technology allows the user to build custom designed chips that

feature 72K, 135K, or 385K probes. Considering that the array will have at least 10% of all probes designated to controls, at least 4-6 probes are designed for each target sequence in the first set (number of probes can be reduced later), and the probes are duplicated, the maximum number of target sequences can be about 8K, 15K, or 43K, respectively. The relevant targets (organisms, genes) determined by investigators' interest and during the previous MicroTOOLs workshop (https://sites.google.com/site/microtoolsii/workshopI) represent central biogeochemical processes such as C, N, P, S, Fe, vitamin B, Zn, Si metabolisms, stress responses (light, temperature, oxidative, toxicity, O2), osmoregulation, quorum sensing, biosynthesis, microbial interactions, phytoplankton bloom indicators, general metabolism, growth rate (cell cycle), circadian (who's expression is under circadian rhythm and who's expression is not), and ecotype differentiation. The gene targets that were selected for the first version of the microarray are listed in Table 1.

After target sequences are submitted, probe design will be done at NimbleGen, and specificity of probe sequences will be validated by BLASTN against all available nucleotide databases including database on ribosomal RNAs. This will not guarantee absence of cross-hybridization in environmental samples, but it will help to eliminate the probes that would most likely cross-hybridize.

The specificity and sensitivity of the arrays will be validated experimentally by hybridizing samples from several cultures (*Prochlorococcus*, *Synechococcus*, SAR11, and a representative of eukaryotic phytoplankton) and environmental samples from coastal and open ocean regions. Below are the questions that will be addressed in the first testing of the arrays:

1. Degree of cross-hybridization of the probes between species
2. Sample processing development and standardization, including nucleic acid extraction and amplification, and ds cDNA synthesis for RNA samples, in order for data to be comparable between different experiments.
3. Eliminating bad probes/targets (that do not light up in any of the samples) and reducing number of probes per sequence to include more targets on the arrays.
4. Developing a data analysis pipeline.

Once arrays are designed and validated, the samples from any laboratory can be sent to and hybridized at NimbleGen at the cost of sample processing. Data analysis will likely be carried out by each laboratory following a single pipeline. The data from microarray experiments presented following the MAQC standards can be stored in a data base and shared. Due to its complexity, development of a data analysis pipeline and a data accessibility network may require an additional MicroTOOLs meeting in 2011.

**Table 1.** List of genes/organism to be targeted on the first set of open ocean and coastal arrays.

| | PHYLUM | GENES |
|---|---|---|
| Eukaryota | *Alveolata*<br>*Chlorophyta*<br>*Chrysophyceae*<br>*Cryptophyta*<br>*Haptophyceae*<br>*Pelagophyte*<br>stramenopiles | *arg, dca, isiB, gap, isiP, nir, nr, petF, phoA, sit,* others |
| Bacteria | *Prochlorococcus*<br>*Synechococcus*<br>*Candidatus Pelagibacter*<br>other alpha<br>gamma | *bop, ddd, dmdA, ftsZ, glnB, glnA, isiB, kaiC, narB, nifH, nirA, nirS, nrtP, ntcA, phn, phoA/X/D, psa, psb, pstS, rbcL, ure, idiA, zwf,* others |
| Archaea | *Crenarchaeote* | *amoA, rbcL, ure* |
| Viruses | *Podovirus*<br>*Myoviridae*<br>*Phycodnavirus*<br>*Picornavirales* | *gp23 , g20, mcp, RdRp, pol* |

III. The following due dates were set:

1. January 15th

   i. Assemble all target sequences. This set of sequences will be used to design probes at NimbleGen by the end of February. First test of the arrays will be carried out during spring, and decisions will be made about modifications and re-designing, if needed.

2. July 15th

i. Collect additional target sequences that could not be ready by the first due date.

ii. Invite other people with relevant expertise who may be interested to be involved in the project.

iii. Automate the workflow for selecting target sequences as new sequences data become available (in collaboration with CAMERA)

iv. Develop a data analysis pipeline

## Conclusions

The MicroTOOLs workshops were held to initiate a scientific community discussion and collaboration to develop better ways to study marine microorganisms given the large spatial and temporal variability of the ocean environment. Methods for the future include high throughput analyses in the laboratory (such as sequencing and chips) as well as remote instrumentation (currently limited throughput qPCR, but ultimately higher throughput technologies, DNA chips, proteomics and perhaps even sequencing).

The implementation of such approaches need to be comprehensive in order to target the multiple characteristics of complex microbial communities and biogeochemical processes, and need to be based on the extant genetic information on marine microbial diversity. Such tools provide a common comprehensive assay that can be used in multiple locations and by numerous investigators. The MicroTOOLs approach requires the input and expertise of the scientific community, but also will become a community resource.

# Participants

| Participants | Institute |
|---|---|
| Adam Monier | Monterey Bay Aquarium Research Institute |
| Andy Millard | University of Warwick |
| Anne Thompson | Massachusetts Institute of Technology |
| Annika Mosier | Stanford University |
| Anton Post | Woods Hole Oceanographic Institution |
| Bess Ward | Princeton University |
| Bethany Jenkins | University of Rhode Island |
| Boris Wawrik | University of Oklahoma |
| Chris Francis | Stanford University |
| Dreux Chappell | University of Rhode Island |
| Holly Simon | Oregon Health & Science University |
| Irina Ilikchyan | UC Santa Cruz |
| Jana Grote | University of Hawai`i at Manoa |
| Jim Tripp | UC Santa Cruz |
| Jody Wright | University of British Columbia |
| Jon Zehr | UC Santa Cruz |
| Jonathan Magasin | UC Santa Cruz |
| Julie LaRoche | Leibniz Institute of Marine Sciences |
| Julie Robidart | UC Santa Cruz |
| Kendra Turk | UC Santa Cruz |
| Libusha Kelly | Massachusetts Institute of Technology |
| Louie Wurch | Woods Hole Oceanographic Institution |
| Mahdi Belcaid | University of Hawai`i at Manoa |
| Mariya Smit | Oregon Health & Science University |
| Martin Ostrowski | University of Warwick |
| Micaela Parker | University of Washington |
| Philip Heller | UC Santa Cruz |
| Rhona Stuart | UC San Diego |
| Ryan Paerl | UC Santa Cruz |
| Scott Gifford | University of Georgia |
| Sebastian Sudek | Monterey Bay Aquarium Research Institute |
| Shellie Bench | UC Santa Cruz |
| Shulei Sun | UC San Diego, CAMERA |
| Tim Hollibaugh | University of Georgia |