Workshops Report and Recommendations
DNA Barcoding of Marine Biodiversity (MarBOL)
Final Version (July 6, 2009)

Executive Summary

MarBOL (DNA Barcoding of Marine Biodiversity, see http://www.marinebarcoding.org/) seeks to effect a huge acceleration in the rate of marine barcoding and to ensure that barcodes are available for members of all key marine groups, with good coverage for groups of highest scientific or societal importance, by 2010. MarBOL, which is led by Dirk Steinke (University of Guelph), will help CoML field projects barcode efficiently and in ways that complement one another (increasing taxonomic and spatial coverage, reducing duplication). The MarBOL goal is to complete barcodes for at least 50,000 species by mid-2010.

A number of CBOL campaigns and CoML field projects are actively engaged in marine barcoding, and several new marine barcoding projects have been initiated with funding from the Alfred P. Sloan Foundation through the MarBOL project. As reported on the BOLD (Barcode of Life Data System, see http://www.barcodinglife.org) website, DNA barcodes have already been prepared for more than 8,000 marine species.

As part of the MarBOL effort, Ann Bucklin (University of Connecticut) organized three one-day workshops which were held at the Alfred Wegener Institute for Polar and Marine Science (April 17, 2009); Woods Hole Oceanographic Institution (May 1, 2009); and Ocean Research Institute, University of Tokyo (May 22, 2009). The workshops immediately followed one-day Marine Barcoding Symposia that featured keynote speakers, who provided overviews on topics of general scientific interest and practical importance for the MarBOL effort. The Symposia also included contributed talks by researchers from the region, and poster sessions.

The goal of the workshops was to identify bottlenecks and facilitate coordination among active marine barcoding centers – especially those associated with CoML projects – and thereby to accelerate progress. We also sought to lay the groundwork for the next steps in barcoding, including use of microarrays and/or metagenetics. The workshop series was designed to reach the global CoML community through meetings at venues in the USA, Europe, and Asia.

This MarBOL Workshops Report summarizes the discussions and recommendations for the major issues related to the MarBOL goals and objectives that were discussed at one or more of the MarBOL workshops. These include

1. Species identification using barcodes
2. Analytical approaches for barcodes
3. Environmental (454) sequencing and metagenetics
4. Prospects for barcode microarrays
5. Phylogenetic and phylogeographic analysis using barcodes
6. Special Applications: Marine Bioinvasions and Harmful algal blooms
7. Barcoding protocols, primer design, and optimization across marine taxa
8. Data quality, data submission, and metadata issues
9. Coordination of barcoding efforts among barcoding centers and countries

Table of Contents

This MarBOL Workshops Report summarizes the discussions and recommendations for the major issues related to the MarBOL goals and objectives that were discussed at one or more of the MarBOL workshops.

Appendices

1. Species identification and discrimination using barcodes
AWI Lead: Paul Hebert
WHOI Lead: Rob DeSalle
ORI Lead: Ryuji Machida

At the AWI, Paul Hebert presented an overview of tree-based identification in BOLD (Barcode of Life Data System, see http://www.barcodinglife.org), which uses a Neighbor Joining (NJ) tree with Kimura-2-Parameter (K-2-P) distances in either DNA or protein. BOLD will highlight and strip out DNA sequences that are wrong. Checking is done by an automatic mismatch program; screening is done for common contaminants (mouse, pig, cow, human). BOLD includes considerable ancillary information or metadata (images, trace files, authorities, etc), as well as links to species pages. Metadata is extremely important; BOLD uses Darwin-Core-1 which is unfriendly and time consuming.

Hebert suggests using a "barcode index number" as an indicator of biodiversity, rather than species names. Names are not canonical and change all the time for most organisms; DNA stays the same. A result of this approach is that many sequences do not have names, and many barcodes will remain unidentified for the foreseeable future. After 250 years, 95% of life is still unregistered, and it may take 2,000 to 3,000 years to finish the job using traditional morphological taxonomy. Barcode clusters or bins will provide a useful picture of diversity that may reflect the actual species diversity rather than the described species diversity. In general, cryptic species count is higher than synonymy from barcodes; mostly taxonomists have not overestimated the number of species. Analysis of additional genes should be used to examine mitochondrial cladogenesis that differs from the morphological analysis. Mitochondrial DNA has a "life of its own" including heteroplasmy, introgression, and aberrant (relict) haplotypes.

Most fishes and zooplankton have overall mean divergences of ~ 7.5%. Barcode clusters or mitochondrial divergences can be treated as a working hypothesis of species divergences. Only a few species have no differentiation (e.g., cichlid fishes). With these divergences, don't need a long barcode. If 200 base-pairs can get almost all species; 50 base-pairs can get 50% of species.

At Woods Hole, Rob DeSalle pointed out the difference between species identification and discovery in terms of goals for barcoding. He reviewed the relationship of DNA barcoding to taxonomy, and reminded everyone that a DNA barcode is not a taxonomic description. Species identification using DNA barcodes can be done using various methods, including: tree-based [e.g., Neighbor Joining (NJ) and other tree building algorithms]; distance-based (e.g., K-2-P distance, barcode gap); and character based (e.g., CAOS; Sarkar et al., 2008 and see https://darchive.mblwhoilibrary.org/handle/1912/2635), and MVSA (REF?). A primary goal for all methods is to "flag" potential new entities for further analysis. All barcoding efforts will yield more descriptive work than can possibly be done by taxonomists.

DeSalle stated that there is a need to catalog, categorize or formalize the barcode data that are emerging from studies. These initiatives will produce many sequences for described species, and also large numbers of samples and sequences that are "flagged" as potential new species. The flagged entities are critical to CoML activities, as well as the MarBOL initiative, and other projects like the Moorea BIOCODE. DeSalle summarized his recommendations as two options:

> Option 1. Expedite taxonomic description or create a category of classification for "flagged" specimens, which should be identified as more than just a novel sequence. This new taxonomic category can be designated as "discovered in a barcode study". Data requirements should be stringent and only rigorous barcode analysis should be included.

Option 2: Create a new reality space or taxonomic bin for these flagged entities that resides outside the formal science of taxonomy. The specimens should be considered provisional new species only for the purposes of ecological and census studies. The goal would be to develop a taxonomic system that operates independently of taxonomy. The provisional species would "graduate" to a position in the morphological taxonomy system after formal description. A model for such a system is a guide for minimal information for a taxonomic description (see Pleijel et al., 2008, Molec. Phylog. Evol. 48:369-371).

DeSalle emphasized that barcoders and morphological taxonomists should work together to erase remaining negative feelings by taxonomists, some of whom initially felt encroached upon and threatened. Barcode campaigns such as MarBOL are producing valuable data that must be incorporated into taxonomic thinking. If the taxonomists don't accommodate barcodes then this is a shortcoming for everyone.

The WHOI workshop recommended that MarBOL P.I. Dirk Steinke request that EOL incorporate species pages for the "flagged" specimens of putative new or cryptic species, and especially that EOL designate such taxa by the genus name, with the GenBank Accession Number in place of a species name.

The plan will be to create pages for putative new or cryptic species, which will encourage taxonomists to work toward formal description, and meanwhile allow accurate assessment of species diversity and biogeographic distribution. The pages should include metadata fields from Darwin Core; photo not mandatory. Truly cryptic species are OK, but any additional information (e.g., behavior, ecology) should be included. The group discussed a WIKIPEDIA approaches to species descriptions to speed up such descriptions: SPECIPEDIA

Allen Collins reported that he is planning an experiment using undescribed sponge species to create an EOL Lifedesk page, based on 16S and COI sequence data. He is working toward an automated function that matches up the DNA sequence to undescribed or described species.

At ORI, Ryuji Machida presented an overview of CoML goals, including: who, where, and how many? He reminded the group that molecular approaches can also be useful to address the issue of species abundances. Molecular approaches can provide estimates of biomass at various levels of biodiversity, including species and above. Also, the abundances of many cryptic or closely-related species are difficult to quantify with usual methods, which primarily include imaging. For example, FLOWCAM and FLOWCYTOBOT image particle by particle, and it is possible that these applications of flow cytometry technology may someday be ready for combined application with molecular detection protocols.

At ORI, Linda Amaral-Zettler reminded the group that for the microbial world, 90% of species have only drawings, not real material, and identifications cannot be confirmed.

The ORI workshop recommended that efforts using DNA sequences to characterize microbial diversity should use and retain cultures where possible to provide a means to confirm diversity estimates.

Also, for 50% of protists there is no link between DNA sequences and morphology. This is in contrast to metazoans. For zooplankton, Shuhei Nishida estimated that 80% of species have "type" reference material? A relate issue for species identification is organismal size: above 3 um size fractions marine biodiversity is much better known that for smaller organisms, which are almost completely unknown. The proportion of known to unknown picoplankton is 95%.

2. Analytical approaches for barcodes

AWI Lead: Gert Wörheide
WHOI Lead: Jesse Ausubel
ORI Lead: Ryuji Machida and Linda Amaral-Zettler

All MarBOL workshops affirmed the importance of accurate and valid analysis of biodiversity at the species level. There was some discussion about species concepts – or lack thereof – for viruses and microbes. Despite this, the workshops agreed that a goal of barcoding should be recognition and discrimination of species.

At the AWI, Paul Hebert recommended analysis of barcode data using Neighbor Joining (NJ) tree-building algorithms. Also at the AWI, Gert Wörheide recommended and explained character-based approaches; coalescent methods; Maximum Likelihood (e.g., RAXML; Stamatakis, 2006; http://icwww.epfl.ch/~stamatak/index-Dateien/countManual7.0.4.php); lineage-through-time plots in Statistical Parsimony Networks or other network approaches, and empirical determination of divergences (see Pons et al., 2006; Vogler et al., 2008). The choice of tree-building algorithms matters most for sequences that have very low levels of divergence. In order to distinguish species using barcode data, there are better approaches than clustering on a NJ tree. Any or all of these should yield greater confidence than a simple percent difference (e.g., 2%) used as a benchmark. On a related note, the use of Kimura-2-Parameter (K-2-P) distances was widely viewed as not particularly more appropriate or valid than proportional distances (p-dis), except that reviewers universally reject the use of proportions for publication.

At WHOI, Jesse Ausubel presented “Barcodes through a Macroscope” and demonstrated new statistical approaches to finding patterns in very large barcode datasets, which allow new exploration and learning. The diagnostic characters are nucleotide substitutions.The analysis seeks to exploit the growing barcode database to characterize patterns of molecular evolution, and also to meet the challenge of scale for biodiversity informatics. For example, EOL has 1.8 million species records on line. We need new ways to visualize and analyze large data sets of species and specimens. Barcodes are valuable characters, but their visualization in trees is not possible for very large datasets (but see Goloboff et al., 2009), assumptions may be violated by horizontal gene transfer, and analysis is computationally demanding. Thus, there is a need for the generation of mathematically-objective procedures for classification or relation of life forms applied generically to genomic databases. In this computationally-undemanding approach, every barcode becomes a digital vector; sequence alignments are matrices, indicator vector for any species (average barcode). Ausubel showed results of trial simulations for North American birds, Canadian freshwater fish, and butterflies.

Amaral-Zettler described and recommended the Visualization and Analysis of Microbial Population Structures (VAMPS) software to recognize and evaluate nuclear18S rDNA sequence clusters. She stated that it may be possible to apply and use VAMPS for COI barcode “clusters” for biodiversity assessments and evaluation.

All workshops recommended further investigation and consideration of additional and alternative approaches to the analysis of barcode data, including character-based approaches, coalescent methods, Maximum Likelihood, lineage-through-time plots, Statistical Parsimony Networks, and empirical determination of divergences. A particular focus should be the accurate and statistically valid identification of “barcode clusters” that can be used to produce quantitative estimates of biodiversity.

3. Environmental (454) sequencing and metagenetics
AWI Leads: Linda Medlin and Ryuji Machida
WHOI Leads: Chris Meyer and Ryuji Machida
ORI Leads: Linda Amaral-Zettler and Koji Hamasaki

All MarBOL workshops discussed environmental sequencing, with an emphasis on metazoans at WHOI, microbes and viruses at ORI, and a mixture at the AWI. At WHOI, the discussion began with general questions, such as: How do environmental barcodes compare to and complement gold-standard voucher-based barcodes? What role will barcodes play in a 454 world? Will we still love COI? Jesse Ausubel put the discussion in a strategic context, with a focus on the goal of showcasing marine biodiversity by 2010, and the need to make technology decisions to support program goals – not the other way around.

At ORI, the discussion of environmental sequencing began with general considerations of applications for which environmental sequencing is needed, including a number of questions that gold-standard barcoding can't answer. Examples included large-scale spatial patterns of vertical distribution and spatial patchiness of zooplankton.

Among all workshops there were many different views on what this activity should be called and whether it is barcoding, per se. There was no clear consensus on the term to be used, and suggestions included: community genomics (Ryuji Machida), environmental sequencing (Nancy Knowlton), environmental genetics (Ann Bucklin), and environmental barcoding (Paul Hebert). Regardless of the name, all workshops agreed that the general approach of sequence-based analysis of diversity has an important – in fact, necessary – role for MarBOL.

The group distinguished between barcoding, which requires an identified specimen and is ideal for analysis of museum and culture collections, and environmental sequencing (the term to be used herein). The latter must meet metadata standards, with the exception that voucher specimens may not be available and cannot be required. However, the material that is the basis of environmental sequencing analysis must be preserved and archived for further examination and analysis.

There are many questions and technical issues that need further investigation in order to ensure that environmental sequencing yields accurate estimates of biodiversity. These include:

- ✓ How valid are environmental collections? Will it be possible to accurately assess biodiversity in highly patchy environments (e.g., the benthos)?

- ✓ How accurate are environmental sequences? Will you get lots of artefactual sequences?

- ✓ Will it be possible to estimate abundance, relative abundance, or biomass, or will you only determine presence / absence?

- ✓ Will you get all species? Will there be extraction bias? Will there be amplification bias (i.e., differential success among species and higher taxa in PCR amplification)? Is COI the best choice or would 18S rRNA be better? Will the results be selective for only a few species, and thus not reflective of the species diversity as a whole? What are the effects of dramatic differences in species abundances and organismal body sizes?

- ✓ What bioinformatics infrastructure (i.e., analytical procedures and capacity) will be needed to identify the specimen to species using only DNA sequences, without a voucher?

Workshop participants discussed results of projects already underway: Ryuji Machida noted that taxonomic diversity in his zooplankton samples was quite broad. Chris Meyer said that at Moorea, they only find fish and human sequences. Keith Crandall noted that COI primers do not amplify about 50% of crustaceans.

The ORI workshop recommended validation experiments for environmental sequencing that compare splits of the same sample: one analyzed by voucher-based barcoding and the other analyzed by environmental sequencing protocols.

The ORI workshop consensus was that environmental sequencing studies of metazoans will need a reference database to tie barcodes to species names. The barcoding database should be verified, and linked to metadata databases that include the names of the people who identified the specimens and determined the DNA sequences, as well as indices of "reliability level". This information is absolutely necessary for MarBOL since there is no quality control process in place for GenBank entries. One suggestion was to invite taxonomists to participate in WIKI correction processes. But Syed Ajmal Khan worried that it will be difficult to ensure we have enough taxonomists, and that India has good taxonomic expertise for the initial period of building the barcode library.

At ORI, Linda Amaral-Zettler and Koji Hamasaki discussed progress by ICOMM and reported on technological developments in DNA-based microbial diversity studies. Deep sampling is very beneficial for microbial diversity estimation. ICOMM has progressed from 1,000s of sequences per year, to 454 sequencing of 10,000s sequence tags per sample. Amaral-Zettler reported some results of replication experiments. Technical replicates (tag sequencing same DNA four times) gives good reproducibility (ranging between 17,000 – 21,000 sequences), but biological replicates (replicate samples analyzed by clone libraries for full-length sequences) vary by environment, quite good for water column and very variable for sediment samples. ICOMM is using heat maps to evaluate reproducibility (Bray-Curtis indices translated into color).

The ORI workshop recommended increased consideration for the need for replication in environmental sequencing studies to characterize microbial diversity. These should include technical, biological, and environmental replicates.

Taxonomic identification for microbes is based on the GAST – not BLAST – reference database and uses full-length sequences. Katie Barrott said that BLAST misidentifies virus sequences by finding short areas of high homology. Instead, viruses are identified using a database with only whole-genomes for taxonomic identification.

At ORI, Linda Amaral-Zettler and Katie Barrott stressed the importance of ancillary contextual environmental data for environmental sequencing. The quality of metadata is critical and also impacts the value of the genetic information. A problem is that such data collection is usually very expensive. It was noted that OBIS does not keep track of environmental data, so ICOMM created MICROBIS to capture these data.

Linda Amaral Zettler described the Genomics Standards Consortium (GSC; see http://darwin.nox.ac.uk/gsc/) which recommends standards for metadata submission when genomes are sequenced. For "orphaned" genomes that are submitted to GenBank but never published, GenBank should require minimal information to be submitted along with the data.

There are new guidelines for “Minimal Information about Environmental Sequence Data (MIENS), see Fields (2008) and link on GSC website. The GSC is open for new members, and has a new journal for publishing genomes that may otherwise not be published. It was also noted that NCBI and DDBJ have a format for single-gene submission information, and unstructured fields for additional information.

4. DNA barcode microarrays
AWI Lead: Marc Kochzius
WHOI Lead: Christoffer Schander

At the AWI, Marc Kochzius weighed the advantages and disadvantages of microarrays for biodiversity analysis. Among the disadvantages are that they can detect known sequences. Linda Medlin countered by recommending the use of hierarchical probes for detection of unknowns. Still, microarrays do not give you new sequence information. At WHOI, Christoffer Schander added to the list that phylochips must be (geographic) site-specific, are expensive, and do not detect new species. Clearly microarrays are useful for some but not all questions.

Discussion of technical issues for using microarrays included:

- ✓ How many species can be detected on each microarray? Answer: depends on the specificity of the probes. Linda says “thousands” for 18S probes, but Marc says 250 or so (for COI and mt16S). Difference is probably due to different genes, patterns of variation, position of label effect (how close label is to the probe), signal strength, differences among probe hydribization conditions, too short sequences for hierarchical probe design.

- ✓ Can you distinguish mitochondrial and nuclear genes? Can you use histone-wrapping of nuclear genes to separate them and use pyrosequencing to sequence the mitochondrial genome?

- ✓ Can microarrays be used to quantify species? Answer: not easily, since the probes have different signal strengths; organisms with variable biomass will also create problems for quantification; probes may not be specific and can cross-react. Focus should be on presence/absence, especially for multicellular organisms. Linda Medlin countered that she is hopeful for quantification of microbes; use calibration curves for different species and different growing conditions; regression analysis with signal strengths has given good data on cell numbers.

- ✓ Can microarray analysis be automated? Are results reliable and reproducible?

- ✓ Which gene is best for chips: mitochondrial COI, 16S, CytB; nuclear 18S rRNA? COI seems to give lots of irregularities perhaps due to intraspecific variability.

The advantages of microarrays were discussed in the context of example projects, most of which are designed to repeatedly survey known, small assemblages.

At WHOI, Christoffer Schander reviewed ongoing projects using microarrays for monitoring and detection of known species, including one he is involved in to track 30 most important species near oil platforms. Other examples include identifying harmful algal bloom species (Linda Medlin and Uwe John); identifying species in ballast water; identifying fish species, such as ichthyoplankton in the North Sea or 1,000 fish typically found in marketplace (see

http://www.fish-and-chips.uni-bremen.de/PostNuke/html/); invasive species in San Francisco Bay; and fish parasites around an aquaculture facility. Chris Meyer suggested using microarrays to monitor perimeters of Marine Protected Areas (MPAs) to characterize larval dispersal. Jesse Ausubel reminded the group of a 1997 CoML report suggesting molecular detection of organisms based on DNA in water (see Ausubel, 1999), and suggested an aquarium test of dissolved (free) DNA to detect pathogens or presence of large species.

Keith Crandall described a project he is working on to "ask the bug" about which gene is best for probe design for microarray applications. He is using a variety of primers and genes, including both species- and group-specific primers. Dirk Steinke noted that BOLD has a database with COI primers. Jesse Ausubel wondered whether "mini barcodes" might be a useful source of information for primer and probe design.

5. Phylogenetic and phylogeographic analysis using barcodes
AWI Leads: Katja Peijnenburg & Silke Laakman
WHOI Lead: Keith Crandall
ORI Lead: Erica Goetze

Katja Peijnenburg (at the AWI) and Erica Goetze (at ORI) surveyed COI uses for phylogeography. COI can help explain possible reasons for genetic patterns, including barriers to gene flow and allopatric speciation, remembering that dispersal ability can be countered by differential selection, survivorship, and divergences from neutrality.

At ORI, Erica Goetze summarized some advantages (e.g., generation of strong hypotheses) and disadvantages (e.g., not enough information for phylogeny) of single-gene approaches. The most usual case is that COI is analyzed as part of larger phylogeographic or phylogenetic studies. Of interest here is the diversity of COI copies within a species. Pelagic species typically have 1-2% intraspecific variation, but there are examples of apparent dramatic rate acceleration, e.g., the copepods, *Tigriopus* (Edmands et al.) and *Microsettella norvegica* (Eberl et al.). Also, cosmopolitan marine species may have larger divergences (e.g., of 30 chaetognath species, 29 species have 10% difference between ocean basins). For these species, barcode projects should use larger sample sizes to resolve intraspecific geographic variation. Low mtDNA diversity and small divergences in other small, numerous organisms may thus require explanation: perhaps selective sweeps or bottle-necks? What is the evidence for these?

Use of barcodes for phylogeographic analysis will rely upon geo-referencing, which GenBank still does not require. The ORI workshop stressed that all DNA sequences including COI barcodes must indicate the geographic source of the material.

At WHOI, Keith Crandall cautioned that COI can't determine a phylogenetic "backbone", even when the COI database is more complete. There was some discussion about whether this is a priority for MarBOL, with general feeling that broader taxonomic coverage is needed, but perhaps there is no need for higher classification. Clearly, deeper branches will analysis of another gene, which can be specific to each group and need not be standardized. Best approach for MarBOL and CBOL is collaboration with the Assembling the Tree of Life (AToL) program.

The WHOI workshop recommended "gap and overlap analysis" to spot taxa that are important for both phylogeny (AToL) and barcoding (MarBOL). This analysis should be used to guide, encourage, and accelerate sample collection and exchange. AToL

coverage focuses on the family level and may include thousands of species also among the ~50,000 marine species targeted by MarBOL.

At WHOI, Keith Crandall led a discussion about the phylogenetic signal in COI. This needs formal analysis and a focused effort to improve the phylogenetic software used (since COI patterns of evolution do not fit the Maximum Likelihood models). A general conclusion is that nucleotide variation of COI resolves species, but not genera (although resolution of genera may improve with better analytical approaches). Amino acid variation of COI resolves deep branches among metazoans. Between these shallow and deep levels, COI is not phylogenetically informative.

Concerns were expressed about BOLD's use of phylogenetic distance (PD) as a diagnostic tool. Although the BOLD diagnostic tools are intended as a means of visualizing barcode clusters and binning species, the tree-based visualization of barcodes across diverse taxa can be misleading and erroneous.

The WHOI workshop recommended further investigation and consideration of additional and alternative approaches to the display and analysis of barcode data, including accurate and statistically valid identification of "barcode clusters". (See also Section 3, above).

6. Special Applications: Marine bioinvasions, harmful algal blooms, species' abundance
AWI Lead: Uwe John
WHOI Lead: Jon Geller
ORI Lead: Kirsty Smith and Florence Pradillon

At the AWI, Uwe John discussed molecular approaches to early-warning systems for harmful algal blooms. He noted that there are 100,000 protists and only ~3,000 are toxic; different strains have different relative abundances. The detection of HAB-forming species is made more difficult by their very small size and little morphological variation. The issue is that surveys for toxin levels must differentiate the strains. There seem to be stable group divergences over time. Successful application is the moored detection system for HABS by Chris Scholin (Monterey Bay Aquarium Research Institute, USA), which can discriminate strains in situ.

At WHOI, Jon Geller noted that marine invasions in North America are still increasing. Uses of barcodes include 1) identification of larvae and juveniles; 2) discrimination of sibling and cryptic species complexes that confuse our understanding of patterns of invasion; and 3) general issues of species identification and taxonomy. Research topics in this regard include mechanisms and pathways of invasion and rapid detection.

Approaches include barcoding studies to identify invertebrate larvae (later matched to barcodes for adults); determine how many species of invaders, even cryptic or undescribed; reveal global-scale geographic patterns of invasion; and detect range expansions, range shifts, and changes in species' biogeographic boundaries.

At the ORI, Kirsty Smith called out the use of barcodes in New Zealand's biosecurity surveillance programs. In addition to the uses listed above, barcoding will allow source population tracking based on haplotypic diversity of mitochondrial genes. Requirements for barcode analysis of invasive species include the use of validated specimens and a global approach.

FISH and DNA barcoding

At ORI, Florence Pradillon described her application of Fluorescent In Situ Hybridization (FISH) for quantification of species in space and time. FISH is a useful application of DNA barcoding, since you must know the sequence data for the target species as well as the other closely-related species in the area, to prevent cross-reactions. Ribosomal genes are the best: 16S or 18S rDNA is the target of choice because of the availability of sequence data. It is possible to develop probes by checking sequence databases for hybridization to transcribed ribosomal rRNA.

There was a detailed discussion about ideal genes, copy numbers, levels of variability, and design of optimal probes. Among the questions addressed were:

- ✓ How different should probes be to maximize specificity? Note that competitive / multiplexed reactions increase specificity.

- ✓ Will it be possible to hybridize directly to DNA, especially a high-copy number genome such as mtDNA? DNA is more stable and easier to use that rRNA, but it may be necessary to use 300 or 400 base-pair probes. Consider chromosomal painting, entailing DNA:DNA hybridization using nick-translation to incorporate label into DNA. In that case wouldn't need to choose genes, but need to start with ethanol-preserved organism, use cultured organisms for probe creation to get enough DNA, or perhaps whole-genome amplification.

- ✓ How many different species can be detected using fluorchrome colors? There are many colors, but problems include autofluorescence and confusion of colors for different species. Consider combinatorial imaging using software to discriminate different colors of wavelengths, with mixtures of different colors to allow detection of many more different probes. Now technology is available for 14 to 40 different colors.

7. Barcoding protocols, primer design, and optimization across marine taxa
AWI Leads: Gert Wörheide and Florian Leese
ORI Leads: Linda Amaral-Zettler, Katie Barrott, and Erica Goetze

Barcoding metazoans

Although the Folmer COI primers work reliably and consistently for some groups, notably fish, PCR success rates for many invertebrate groups are typically much lower, in the neighborhood of 50% (for crustaceans) or less (10% - 20% for sponges). For the remaining species, group-specific primer design is necessary.

At the AWI, Gert Wörheide discussed results for problematical groups, including polychaetes, sponges, and echinoderms. He noted that the sponge barcoding project (see http://www.spongebarcoding.org/) has detailed protocol and primer information available. Additional strategies for overcoming difficulties in obtaining barcode sequence data were discussed, including:

- ✓ Primer cocktails especially in high-throughput facilities and for some groups (e.g., fish, echindoderms);
- ✓ Sequencing a larger fragment and then re-amplify shorter regions with specific primers;
- ✓ Increase target template concentration; reducing the non-target DNA;
- ✓ Whole-genome amplification, which may be good for protists;

- ✓ Dilution to remove contaminants and inhibitors;
- ✓ High-low stringency (touch-down) reactions;
- ✓ Cloning (but watch for PCR error);
- ✓ Try different Taq polymerases (e.g., platinum Taq);
- ✓ Remove inhibitory secondary metabolites using the DNAeasy Plant kit or by dilution;
- ✓ Watch for pseudogenes (with stop codon), multiple gene copies).

At the AWI, Florian Leese talked about the "formalin problem". There seemed to be general consensus that commercial kits for analysis of damaged DNA do not work well for formalinized tissues. Ann Bucklin recommended that the group read a report on DNA sequencing from formalin preserved samples (Tang, 2006; see http://www.nap.edu/catalog.php?record_id=11712).

All MarBOL workshops recommended that MarBOL work to improve approaches to sharing detailed information on barcoding protocols, group-specific primer sequences, and other information to accelerate marine barcoding progress for difficult groups. Suggestions included a website clearing house or list of websites with this information.

Barcoding viruses

At ORI, Katie Barotte said that only 22% of viral sequences are known, providing great opportunity for discovery in "viral sequence space". Viral biodiversity is calculated as the overlap of sequences from 454 sequencing, but must do metagenomics since there are no conserved sequence regions in any virus. Whole-genome sequencing for DNA viruses is also useful, since they also want to know the function of the virus, and ~25% of genes can be identified to function.

Barcoding microbes

At ORI, Linda Amaral Zettler described sequencing of the V6-Tag region (60-100 base pairs) at the 3' end of 18S rRNA. Among 1M sequences, there are likely lots of plankton, and MarBOL should take advantage of these to match to sequences for eukaryotic species.

The ORI workshop recommended that MarBOL facilitate coordination with ICOMM to compare results from the same samples using the V6 "tag" region of 18S rRNA and COI barcodes.

Also at ORI, Koji Hamasaki described his ICOMM 454 "Keck" project. Microbial diversity is usually calculated at 97% similarity to define separate "species" (see Pedrós-Alió, 2006). Rare taxa are important; depth of prokaryote diversity is unprecedented (see Sogin et al., 2006; Huber et al., 2007). Exhaustive sequencing is less than ideal, but still help estimate diversity and – with additional sampling – can characterize temporal and spatial variation.

Barcoding protists and fungi

At ORI, Linda Amaral-Zettler summarized progress in protistan barcoding, including cilates (using COI), dinoflagellates (COI, COB), diatoms (COI, RBCL, 18S, and ITS). She stated that there should be a requirement for DNA preservation of new species, using cryo-preservation if possible (although this will not preserve all forms). There likely will not be a single barcode gene for protists, especially since some groups lack mitochondria. Also, there is considerable intraspecific variation (>3%) within species and large differences between life stages.

Barcoding fungi

Linda Amaral-Zettler summarized barcoding of yeast, which has used the D1/D2 region of large subunit rRNA, as well as other genes. COI has some advantages (see Hanner and Gregory, 2007), including easy alignment across lineages and good resolving power for some groups. There would be many advantages in expanding COI use across fungi and protists, as well as animals. Disadvantages of COI for fungi include length variation due to intron variability (cDNA may range in length from 1504 to 1905), and difficulty of amplification for some groups.

8. Data quality, data submission, and metadata issues

At WHOI, Chris Meyer emphasized the need for consistent metadata schema for barcoding. Consistent metadata standards are required for efficient processing, tracking and interoperability of information management systems. The database working group (DWG) from the Consortium for the Barcode of Life (CBOL) has set forth proposed standards and required documentation data elements to receive the "BARCODE" flag from NCBI and its collaborators (See http://barcoding.si.edu/PDF/DWG_data_standards-Final.pdf).

These standards generally require any entity that is generating DNA barcodes to track all data elements through the barcoding pipeline. Currently, there is no turnkey, universal management system to aggregate and track these metadata, and this responsibility falls to each research project or program to manage the information on its own. For smaller scale operations, this can be done relatively easily, but when multiple investigators are involved in large expeditions, or the physical elements are handled in different locations (collection sites vs. sequencing centers), data management becomes imperative, but significantly more difficult. Moreover, data elements required by the DWG of CBOL are a subset of potentially useful metadata associated with specimens that many research communities would like to have available (see OBIS data schema at: http://www.iobis.org/tech/provider/schemadef1.html).

This OBIS schema is an extension of the Darwin Core (v.2), and the standards are meant to facilitate interoperability between databases. However, the current OBIS schema is relevant only for specimen data.

The WHOI workshop recommended that tracking tissues, extractions, primers, trace files and sequences be incorporated into DNA barcoding workflows to enable efficient recovery for posting to the appropriate public repositories (e.g. Genbank or BOLD). The ability to track such derivatives (tissue and extracts) will become increasingly more important with genetic resource/CBD obligations and requirements associated with BioIP and material transfer agreements.

The ORI workshop recommended that all barcode data should be verified and linked to metadata databases that include the names of the people who identified the specimens and determined the DNA sequences, as well as indices of "reliability level". This information is absolutely necessary for MarBOL since there is no quality control process in place for GenBank entries.

Allen Collins and Chris Meyer reported that various research programs are engaged in developing databases and tracking systems for handling these workflows. Researchers wishing to investigate this further are directed to either the CMarZ Barcoding Association (see http://www.cmarz.org/jg/dir/CMarZ) or the Moorea Biocode Project (see http://biocode.berkeley.edu) and the contacts listed at each site. Among commercial products

for meeting metadata and specimen-tracking goals is software under development by Geneious for the Moorea BIOCODE project.

9. Coordination of barcoding efforts among countries

ORI Lead: Syed Ajmal Khan

At ORI, Syed Ajmal Khan noted the irony that countries that are poor in diversity are frequently rich in technology, yet countries that are rich in biodiversity are poor in technology. We must do away with this mismatch! Cases in point include the Red Sea (30% endemism), Madagascar (70% endemism), and India and the Indian Ocean, which many endemic species and very high species diversity in coastal waters. In the case of India, national biodiversity of laws prevent sending out samples or DNA; exceptions to laws are almost impossible to secure. Thus, personnel should be trained to be able to barcode material in India.

An ORI recommendation was that MarBOL should place priority on marine biodiversity hotspots and areas with high endemism.

Countries that should be given particular consideration for training programs for barcoding marine organisms (especially invertebrates and microbes) include: Brazil, China, India, Indonesia, Phillipines, and Malaysia. Training should involve assistance with trouble-shooting, primer design, technical issues associated with marine DNA barcoding. In addition, websites should provide more details, and should serve to consolidate the information on protocols.

10. Literature cited

Ausubel, J.H. (1999) Toward a Census of Marine Life. Oceanography 12(3):4-5, 1999.

Field, D., et al. (2008) The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 26(5):541-7.

Goloboff, P.A., et al. (2009) Phylogenetic analysis of 73,060 taxa corroborates major eukaryotic groups. Cladistics 25:1-20.

Hanner, R.H. and T.R. Gregory (2007) Genomic diversity research and the role of biorepositories. Cell Preservation Technology 5: 93-103.

Huber, J.A., D.B. Mark Welch, H.G. Morrison, S.M. Huse, P.R. Neal, D.A. Butterfield, and M.L. Sogin (2007) Microbial population structures in the deep marine biosphere. Science 318 (5847): 97-100

Pedrós-Alió, C. (2006) Marine microbial diversity: can it be determined? Trends Microbiol. 14, 259 (2006).

Pons, J. et al. (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55(4):595-609; doi:10.1080/10635150600852011

Sarkar, I.N., P.J. Planet and R. DeSalle (2008) CAOS software for use in character-based DNA barcoding. Mol Ecol Resources 8(6):1256-1259.

Sogin, M.L. et al. (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proc. Nat. Acad. Sci. 103 (32): 12115-12120

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22(21):2688-90.

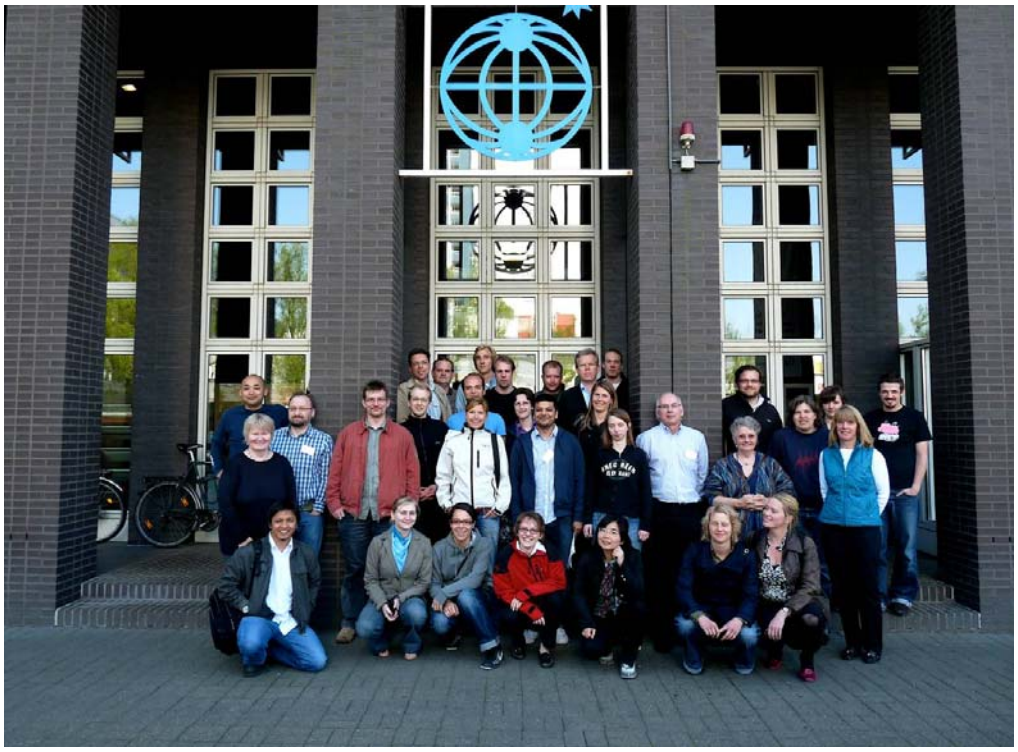
Tang, E.P.Y. (2006) Path to effective recovering of DNA from formalin-fixed biological samples in natural history collections: Workshop summary. National Academies Press, Washington DC. 64 pages.

Vogler, C., J. Benzie, H. Lessios, P. Barber, and G. Worheide (2008) A threat to coral reefs multiplied? Four species of crown-of-thorns starfish. Biology Letters 4: 696-699.

Appendix I. Invited speakers, organizers, and participants for MarBOL Workshops

Last	First	Institution	Country	Email	Attended?
BREMERHAVEN GERMANY					
Invited Speakers					
Hebert	Paul	Univ. Guelph	Canada	phebert@uoguelph.ca	Yes
John	Uwe	AWI	Germany	uwe.john@awi.e	Yes
Kochzius	Marc	Univ. Bremen	Germany	kochzius@uni-bremen.de	Yes
Laakmann	Silke	University of Bremen	Germany	laakmann@uni-bremen.de	Yes
Medlin	Linda	AWI	Germany	lkmedlin@awi-bremerhaven.de	Yes
Peijnenburg	Katje	Univ Amsterdam	Netherlands	K.T.C.A.Peijnenburg@uva.nl	Yes
Probert	Ian	CNRS	France	probert@sb-roscoff.fr	Yes
Wörheide	Gert	Munich	Germany	woerheide@lmu.de	Yes
Organizers					
Blanco Bercial	Leo	University of Connecticut	USA	leocadio@uconn.edu	Yes
Bucklin	Ann	University of Connecticut	USA	ann.bucklin@uconn.edu	Yes
Cornils	Astrid	Alfred Wegener Inst	Germany	Astrid.Cornils@awi.de	Yes
Machida	Ryuji	University of Tokyo	Japan	ryuji@ori.u-tokyo.ac.jp	Yes
Schiel	Sigrid	Alfred Wegener Inst	Germany	sschiel@awi-bremerhaven.de	Yes
Steinke	Dirk	University of Guelph	Canada	dsteinke@uoguelph.ca	Yes
Participants					
Ivanenko	Viacheslav	Moscow State University	Russia	ivanenko.slava@gmail.com	Yes
Khelifi-Touhami	Meriem	Universit Badji Mokhtar	Algeria	khelifi_meriem@yahoo.fr	Yes
Kiko	Rainer	Institut für Polarökologie	Germany	rkiko@ipoe.uni-kiel.de	Yes
Krabbe	Kathrin	Ruhr University Bochum	Germany	kathrin.krabbe@rub.de	Yes
Kuropka	Jana	Biologische Anstalt Helgoland	Germany	jana.kuropka@awi.de	Yes
Leese	Florian	Ruhr University Bochum	Germany	florian.leese@rub.de	Yes
Pollupuu	Maria	Estonian Marine Institute	Estonia	maria.pollupuu@ut.ee	Yes
Trunk	Simone Beatrice	Ruhr University Bochum	Germany	simone.trunk@ruhr-uni-bochum.de	Yes
Wilmsen	Eileen	Ruhr-University-Bochum	Germany	Eileen.wilmsen@goolgemail.com	Yes
WOODS HOLE, MASS					
Invited Speakers					
Collins	Allan	Smithsonian	USA	CollinsA@SI.edu	Yes
Crandall	Keith	BYU	USA	Keith_Crandall@byu.edu	Yes
Creer	Simon	University of Wales, Bangor	UK	s.creer@bangor.ac.uk	No
DeSalle	Rob	AMNH	USA	desalle@amnh.org	Yes
Geller	Jon	Moss Landing Lab, CSU	USA	geller@mlml.calstate.edu	Yes
Knowlton	Nancy	Smithsonian	USA	Knowlton@si.edu	Yes
Meyer	Chris	Smithsonian	USA	meyerc@si.edu	Yes
Schander	Christoffer	University of Bergen	Norway	christoffer.schander@bio.uib.no	Yes
Organizers					
Blanco Bercial	Leo	University of Connecticut	USA	leocadio@uconn.edu	Yes
Bucklin	Ann	University of Connecticut	USA	ann.bucklin@uconn.edu	Yes
Copley	Nancy	Woods Hole Oceanogr Inst	USA	ncopley@whoi.edu	Yes
Cornils	Astrid	Alfred Wegener Inst	Germany	Astrid.Cornils@awi.de	Yes
Machida	Ryuji	University of Tokyo	Japan	ryuji@ori.u-tokyo.ac.jp	Yes
Steinke	Dirk	University of Guelph	Canada	dsteinke@uoguelph.ca	Yes
Wiebe	Peter	Woods Hole Oceanogr Inst	USA	pwiebe@whoi.edu	Yes
Participants					
Aarbakke	Ole	University of Tromsoe	Norway	ole.n.aarbakke@uit.no	Yes
Allison	Dicky	Woods Hole Oceanogr Inst	USA	dallison@whoi.edu	Yes
Batta Lona	Paola	University of Connecticut	USA	paola.batta_lona@uconn.edu	Yes
Cepeda	Georgina	INIDEP	Argentina	gcepeda@fiba.org.ar	Yes
Giridhar	Rajesh	National Inst Oceanogr	India	rajeshgiridhar@gmail.com	Yes
Jennings	Rob	University of Massachusetts	USA	rob.jennings@umb.edu	Yes
Rhodes	Adelaide	University of South Florida	USA	acerhodes@ufl.edu	Yes
Questel	Jennifer	University of Alaska	USA	jenn.questel@sfos.uaf.edu	Yes
Radulovici	Adrianna	Dept Fisheries Oceans	Canada	adriana.radulovici@gmail.com	No
Xavier	Francis	National Inst Oceanogr	India	franciskx@gmail.com	Yes
TOKYO JAPAN					
Invited Speakers					
Ajmal Khan	Syed	Annamalai University	India	seyedajmal@gmail.com	Yes
Amaral-Zettler	Linda	Marine Biological Laboratory	USA	amaral@mbl.edu	Yes
Barrott	Katie	San Diego State University	USA	katiebarott@gmail.com	Yes
Goetze	Erica	University of Hawaii	USA	egoetze@hawaii.edu	Yes
Hamasaki	Koji	University of Tokyo	Japan	hamasaki@ori.u-tokyo.ac.jp	Yes
Kogure	Kazohiro	University of Tokyo	Japan	kogure@ori.u-tokyo.ac.jp	Yes
Miya	Masaki	Natural Hist Museum Inst	Japan	miya@chiba-muse.or.jp	Yes
Nishida	Mutsumi	Ocean Research Inst.	Japan		Yes
Pradillon	Florence	JAMSTEC	Japan	fpradillon@jamstec.go.jp	Yes
Song	Sun	Inst. Oceanology, China	China	sunsong.cn@gmail.com	No
Utsuki	Nozomu	Tokyo Metropolitan University	Japan	asilidae@gmail.com	Yes
Participants					
Kuriyama	Mikiko	Nat Res Inst Fish Science	Japan	mikiko@affrc.go.jp	Yes
Li	Chaolun	Inst. Oceanology, China	China	lcl@ms.qdio.ac.cn	No
Miyamoto	Hiroomi	University of Tokyo	Japan	miyamoto@ori.u-tokyo.ac.jp	Yes

Appendix II. MarBOL workshop group photos



Alfred Wegener Institute for Polar and Marine Science (April 17, 2009)

Woods Hole Oceanographic Institution (May 1, 2009)

Appendix III. Agendas for MarBOL Workshops at AWI, WHOI, and ORI

Workshop

DNA Barcoding of Marine Biodiversity (MarBOL)

Alfred Wegener Institute for Polar and Marine Science

Bremerhaven, Germany - Friday, April 17, 2009

Agenda

8:30 am	Coffee and Pastries
9:00 am	Welcome – Ann Bucklin and Dirk Steinke
9:15 am	***Species identification and discrimination using barcodes*** (Lead: Paul Hebert)
9:45 am	***Community analysis using barcodes*** (Leads: Linda Medlin & Ryuji Machida)
10:30 am	Coffee Break
10:45 am	***Phylogenetic and phylogeographic analysis using barcodes*** (Leads: Katja Peijnenburg & Silke Laakman)
11:15 am	***Barcoding protocols, primer design, and optimization across marine taxa*** (Leads: Gert Wörheide & Florian Leese)
12:00 pm	LUNCH on site
1:30 pm	***DNA barcode microarrays*** (Lead: Marc Kochzius)
2:00 pm	***Statistics and bioinformatics challenges for barcode data*** (Leads: Dirk Steinke & Ann Bucklin)
2:30 pm	***Detection of harmful algal blooms using DNA barcodes*** (Lead: Uwe John)
3:00 pm	Coffee Break
3:15 pm	***Submission of barcode data to BOLD and GenBank BARCODE*** (Leads: Dirk Steinke & Ann Bucklin)
3:45 pm	***Coordination among marine barcoding centers*** (Lead: Ann Bucklin)
4:15 pm	***Realistic marine barcoding goals by October 2010*** (Lead: Dirk Steinke)
4:45 pm	Open discussion
5:15 pm	Action items and next steps
5:30 pm	Adjourn

Workshop

DNA Barcoding of Marine Biodiversity (MarBOL)

Woods Hole Oceanographic Institution

Woods Hole, USA - Friday, May 1, 2009

Agenda

9:00 am	Welcome – Ann Bucklin and Dirk Steinke
9:15 am	***Species identification using barcodes*** (Lead: Rob DeSalle)
9:45 am	***Promise and pitfalls of environmental (454) sequencing and metagenetics*** (Leads: Chris Meyer, Si Creer, Ryuji Machida)
10:15 am	***DNA barcode microarrays*** (Lead: Chris Schander)
10:45 am	Coffee break
11:00 am	***Phylogenetic analysis using barcodes*** (Leads: Keith Crandall & Allen Collins)
11:30 am	***Phylogeographic analysis using barcodes*** (Lead: Tim Shank)
12:00 Noon	***Analysis of marine bioinvasions using barcodes*** (Lead: Jon Geller)
12:30 pm	Lunch on site
2:00 pm	***Barcoding protocols, trouble-shooting, and optimization across marine taxa*** (Leads: Rob Jennings & Leo Blanco Bercial)
2:30 pm	***Submission of barcode data to BOLD and GenBank BARCODE*** (Leads: Dirk Steinke & Ann Bucklin)
3:00 pm	***Analytical approaches for publication*** (Lead: Rob DeSalle)
3:30 pm	Coffee break
4:00 pm	***Coordination among marine barcoding centers*** (Lead: Ann Bucklin)
4:30 pm	***Realistic marine barcoding goals by October 2010*** (Lead: Ann Bucklin)
5:00 pm	Open discussion
5:30 pm	Action items and next steps
6:00 pm	Adjourn

Workshop

DNA Barcoding of Marine Biodiversity (MarBOL)

Ocean Reearch Institute, University of Tokyo

Tokyo, Japan - Friday, May 22, 2009

Agenda

8:30 am	Coffee
9:00 am	Welcome – Ryuji Machida
9:15 am	***Species identification using barcodes - Promise and pitfalls***
	Lead: Ryuji Machida
9:45 am	***Environmental sequencing and DNA barcoding***
	Leads: Linda Amaral-Zettler, Koji Hamasaki
10:15 am	***FISH and DNA barcoding***
	Lead: Florence Pradillon
10:45 am	Coffee break
11:00 am	***Advantages and disadvantages of single-gene approaches***
	Leads: Erica Goetze, Masaki Miya
11:30 am	***Barcoding viruses***
	Lead: Katie Barotte
12:00 Noon	***Barcoding protists and fungi***
	Lead: Linda Amaral-Zettler
12:30 pm	Lunch on site
2:00 pm	***Barcoding metazoans***
	Leads: Leocadio Bercial Blanco, Astrid Cornils
2:30 pm	***Coordination of barcoding efforts among countries***
	Leads: Sun Song, Chaolun Li, Syed Ajmal khan, Nozomu Utsuki
3:00 pm	***Databases and bioinformatics***
	Leads: Ann Bucklin, Ryuji Machida, Linda Amaral-Zettler
3:30 pm	Coffee break
4:00 pm	***DNA barcoding and invasive species***
	Lead: Kirsty Smith
4:30 pm	Open discussion, action items, and next steps
5:00 pm	Adjourn