NESCent Catalysis Meeting Report

High-throughput biodiversity research using eukaryotic metagenetics
January 24-27, 2011

Meeting Coordinators: Holly Bik (Postdoctoral Researcher, Univ. of New Hampshire)
Kelley Thomas (Professor, Univ. of New Hampshire)

Meeting Leaders (in addition to coordinators):
• Simon Creer (Senior Research Fellow, Univ. of Bangor, Wales)
• Dorota Porazinska (Assistant Professor, Univ. of Florida)
• Robin Giblin-Davis (Professor, Univ. of Florida)
• Way Sung (Graduate Student, Univ. of New Hampshire)

Participants:
• Jan Pawlowski (Professor, Univ. of Geneva)
• Franck Lejzerowicz (Graduate Student, Univ. of Geneva)
• Tony Walters (Graduate Student, Univ. of Colorado, Boulder)
• Florent Angly (Postdoctoral Researcher, Univ. of Queensland)
• Mitchell Sogin (Professor, Marine Biological Lab)
• Greg Caporaso (Postdoctoral Researcher, Univ. of Colorado, Boulder)
• Chris Quince (Research Fellow, Univ. of Glasgow)
• Yijun Sun (Research Scientist, Univ. of Florida)
• Shawn Polson (Assistant Professor, Univ. of Delaware; NECC/CAMERA database)
• Mike Pfrender (Associate Professor, Notre Dame)
• Eric Allen (Assistant Professor, Scripps/CAMERA)
• Erick Matsen (Postdoc, Fred Hutchinson Cancer Research Center)
• Alexandros Stamatakis (Group Leader, Scientific Computing group, Heidelberg Institute for Theoretical Studies)
• Simon Berger (Graduate Student, Heidelberg Institute for Theoretical Studies)
• Samantha Lewis (Graduate Student, Univ. of California, Riverside)
• Anthony Chariton (Research Scientist, CSIRO, Australia)
• Erik Pilgrim (Research Biologist, EPA Cincinnati)
• Axayácatl Rocha Olivares (Research Scientist, CICESE, Mexico)

Meeting Goals

Eukaryotic meiofauna (organisms 45μm-1mm) are abundant and ubiquitous across every ecosystem on earth. Yet, there exists a well-recognized gap in the understanding of their global biodiversity. The emerging field of eukaryotic metagenetics encompasses en mass meiofaunal biodiversity assessment using high-
throughput (HTP) sequencing of traditional molecular loci (e.g. ribosomal rRNA). Despite the promising outlook of HTP approaches, we currently have a poor understanding ribosomal RNA gene evolution and continue to lack the cyberinfrastructure needed for effective interpretation of such large datasets. The NESCent meeting aimed to bring together current collaborators and other investigators who are keen to share expertise and contribute towards computational/theoretical advances within the field. Ultimately, we planned to use the meeting discussions and interactions as a basis for a longer-term Research Coordination Network funding proposal through NSF in 2011.

During the meeting sessions, our goal was to address three central themes:

1. Discuss how to generate effective data with maximum global usefulness (by coordinating the use of common loci), and how to do so in the face of evolving sequencing technology and an expanding number of potential target loci (e.g. whole mitochondrial genomes).

2. Outline the current tools available and future development of resources for primary data analysis, including: clustering raw sequence reads into Molecular Operational Taxonomic Units (MOTUs), phylogenetic pipelines (interpreting short reads in an evolutionary context), and ecological tools (e.g. rarefaction, algorithms for community comparisons).

3. Design data sharing mechanisms and plan cyber resources that meet the needs of the eukaryotic community, by leveraging tools/resources developed currently geared towards microbial projects.

Summary of activities and discussion

Day 1 (January 24th): Pre-meeting discussion and planning session attended by meeting coordinators (Bik/Thomas) and meeting leaders (Creer, Porazinska, Gliblin-Davis, and Sung). All participants in this group are long-term collaborators working on eukaryotic metagenetic projects across diverse environments (deep-sea, marine sediments, and soil communities). Thus, the pre-meeting session was necessary to discuss recent progress on individual projects, and outline the ongoing challenges encountered in data analysis. Group input helped to solidify the meeting focus around a central set of key questions. We also divided up responsibilities for chairing discussion sessions and presenting an overview of our methods/projects (essential to set the stage for participants from diverse backgrounds, e.g. computational biologists)

Days 2 and 3 (January 25th -26th): Meeting sessions attended by all participants. Pre-meeting participants delivered overview presentations during the morning session of Day 2; these presentations outlined the current state of the field (based on our own research and interpretation of the literature) in order to clearly define the meeting goals.
The remainder of the two days was structured as a series of open discussion sessions, organized under five themes: 1) Collecting environmental samples, 2) Generating high-throughput sequence data, 3) Clustering sequence reads into Operational Taxonomic units (OTUs), 4) Placing and visualizing sequence data in an evolutionary framework, and 5) Interpreting sequence data in a biological/ecological context. Based on their published research and personal expertise, a few meeting participants were assigned to lead the discussion on each topic area (and present data/relevant background information to the group if necessary).

**Day 4 (January 27th):** On the final day, participants were split up into breakout groups. Based on the previous two days of meeting discussions, we decided that three breakout groups were needed to address the most pressing questions: 1) What simulation/benchmarking studies do we need for high-throughput data, and how do we design them? 2) What database resources do we need and how do we design them? and 3) What information do we need to make biological/ecological conclusions from high-throughput data? Breakouts met during the morning session and presented their group reports during the afternoon; we asked that reports address two questions: a) What knowledge/empirical gaps currently exist, and b) What outputs can be produced from the NESCent meeting?

The key insights from our meeting discussions and breakout groups were as follows:

**Environmental Sampling**
- There should be a push towards effective collection of metadata (following MIENS standards)
- A nesting approach (pseudoreplicates) is a good approach, as it attempts to cover spatial community variability; reviewers ask for replication, but sequencing is not free—looking at the effectiveness of subsampling has not been done because of cost (and informatic) limitations, even though the questions may be interesting.

**Data Generation**
- Amongst the possible new platforms for HTP sequencing, microbial groups are already moving towards Illumina because of higher number of reads (10x) for an equivalent cost to 454 sequencing.
- However, we are still unclear about the nature of Illumina sequencing errors (are they non-random? Reproducible?)
- Eukaryotic studies are consistently hindered by sparse reference databases, which prevent the use of other genetic loci and shorter sequence reads
- PCR methods should aim to minimize chimera formation: fewer cycles, longer extension time, shorter amplicons, avoiding amplicons with a conserved region in the middle, increasing primer melting temperature
OTU clustering

- The use of homologous loci will be imperative for conducting future metadata analysis and phylogeographic studies.
- Pairwise sequence alignment is recommended for OTU clustering over Multiple sequence alignment, since the latter severely overestimates biodiversity which propagate to downstream analyses.
- Noise removal pipelines (PyroNoise, AmpliconNoise) are a potentially critical step for removing sequencing/PCR errors before clustering is undertaken.
- Computational pipelines can handle raw reads, although the sheer volume of data means that OTU clustering is a more practical, and thus likely to persist.

Placing data in an evolutionary framework

- Even for short sequences (above 200bp), Evolutionary Placement Algorithms allow for robust phylogenetic placement—no more than one node away from the true evolutionary placement (inferences based on empirical tests of mock datasets). However, these analyses are critically dependant on a robust guide tree.
- Phylogenetic placement of short reads can help you identify taxon sampling problems in the reference dataset that would not be obvious by BLAST searches.
- There are upcoming pipelines for defining important ‘edges’ in tree topologies—e.g. identifying key lineages that define community differences
- **There is a clear mandate from NSF to link high-throughput sequence data and taxonomy, but in the absence of a robust reference database (and lack of funding to develop one), this database issue presents a substantial barrier for robust evolutionary interpretations of metagenetic data.**

Interpreting data in a biological/ecological context

- If we are able to prove that high-throughput sequencing offers a deeper picture of diversity (even without abundances), then it represents a much more valuable tool for environmental assessment/monitoring versus traditional taxonomic surveys that can only offer small snapshots of communities. This is an important selling point for promoting metagenetic techniques across a wider community.
- Quantification isn’t what drives our ability to determine differences across sites. Based on real data from California—presence/absence data provides plenty of signal to differentiate the communities (based on real environmental data from California)
- A greater understanding of intragenomic rRNA variation across species will be key for interpreting environmental datasets. **Full-length reference sequences will be crucial for quantifying this variation in poorly described taxa.**
Plan for follow-up activities

We believe this meeting was extraordinarily fruitful, and expect that many participants will maintain contact and develop new collaborations in the future. From our perspective (the meeting coordinators/leaders), we intend to focus on synthetic research that merges the expertise of computer scientists and biologists. It became clear during this meeting that such a relationship can lead to substantial progress within the field—as biologists we will aim to provide real environmental datasets for computer scientists to test their models/algorithms, as well as provide feedback on the multitude of tools available for data analysis. In addition, there may be scope to submit several working group proposals that narrowly focus on specific questions; we will follow-up to assess participants interest and willingness to prepare working group proposals.

Anticipated outcomes and products

- We anticipate that most immediate meeting output will be a comprehensive review paper (aimed at TREE or another similar journal) outlining the current state of the field, available tools and existing challenges. The target audience will be biologists and ecologists with a non-computational background.

- The fast-moving nature of the field and obvious parallel efforts of research groups has highlighted the pressing need for coordinated efforts in eukaryotic-focused metagenetic research. Thus, we will move forward to submit a Research Coordination Network proposal through NSF at the next deadline in early July.

- The lack of robust, reference guide trees for eukaryotes was an overarching theme within the meeting discussions and breakout groups. There must be a push towards obtaining full-length rRNA reference sequences from taxa that are currently underrepresented or missing from public sequence repositories. We aim to highlight this issue with funding bodies and put more effort into identifying (and filling) glaring gaps in the Tree of Life.

- As a result of discussions with Eric Allen (CAMERA database) and Greg Caporaso (QIIME pipeline), we will be developing new tools and database structures that specifically meet the needs of the eukaryotic community. We expect two primary outcomes along these lines: 1) Incorporating eukaryotic sequences into OTU databases being developed by the Knight lab at the University of Colorado, Boulder and 2) Working with camera as a ‘test case’ to develop a useful tool for eukaryotic users within the CAMERA database. This will most likely be a graphical web tool utilizing Evolutionary Placement Algorithms to phylogenetically place short sequence reads (uploaded as a FASTA file) onto a reference guide tree.

- We hope that a less tangible (but equally important) meeting outcome will be a more open dialogue between computer scientists and biologists. Informative,
easy-to-use bioinformatic tools will become increasingly important as larger datasets are generated. Intuitive pipelines (e.g. graphical program interfaces, web tools) are needed to reach a wider community of non-computationally trained ecologists and biologists. Merging knowledge from disparate fields will help to address these needs and produce effective, community-driven resources.