OCN 750: Microbial Community Sequence Data Analysis
Sample Syllabus
Craig Nelson + 15 students

This is an experimental graduate level course designed to address a need among students at UH Mānoa for training in the analysis of amplicon-based community biological sequence data. The widespread and growing interest in “metagenomics”, sequencing raw pooled genomic DNA extracted from both diverse environmental habitats and host tissues, is the scientific motivation for this course. The structure and direction of the course is fluid, and will be guided by student needs and interest.

Initial surveys of students established some clear consistencies in what the majority of students are interested in and have prior experience in.

- Almost all of the students are interested in working with amplicons of the 16S gene, typically used to profile the taxonomic composition of microbial communities.
- Almost all of the students have or are acquiring data comprising paired-end nucleotide reads from amplicon libraries of 16S sequenced on the Illumina MiSeq platform.
- Most of the students have experience with basic bioinformatics, including the idea of using a tool such as BLAST to perform a local alignment of a sequence to a database to infer both the organismal source and the likely gene source.
- Most of the students have some familiarity with the Unix/Linux operating system where bioinformatics platforms are typically run (aka “Command Line Interface”).
- The students are split between experience using personal computers (running bioinformatic routines on either the UH cluster, or a handful of students have used AWS or Google Cloud).

Based on these generalities the course will focus on the practical bioinformatic analysis of 16S sequences using widely available “pipelines”, with a focus on the post-bioinformatic analysis and visualization of data. The work will be structured do that students can move at their own pace, with key concepts in bioinformatics and statistics presented in a lecture-based format when a majority of students are “ready” for each topic. Students will work together to learn software packages and share code, and there will be only minimal effort to homogenize the nuts and bolts of pipelines or statistics. That said, the core software packages used in the course will be mothur for amplicon bioinformatics, R and JMP for statistics and visualization. None are required or even recommended, and students are encouraged to use any software that interests them as long as they are willing to work with the instructor to troubleshoot.

A general proposed conceptual outline of the course timeline is presented below.
W1: Basic amplicon generation, sequencing, and processing
   • Library construction
   • The MiSeq platform
   • Quality scoring
   • Demultiplexing
   • Paired end merging
   • Quality control
   • Interpreting failed amplifications

W2: Working with sequence files
   • Fasta format
   • BLAST
   • Dereplication
   • Denoising

W3: Multiple sequence alignments
   • Global vs. local alignment and algorithms
   • Incorporating secondary structure into alignments
   • Databases (for 16S this would be SILVA, RDP, Greengenes)
   • Alignment algorithms
   • Classification algorithms

W4: Operational Taxonomic Units
   • Philosophies of OTU construction
   • QIIME vs. mothur
   • Definitions of OTUs
   • ASVs, ESVs, sOTUs

W5: Phylogenies
   • Building trees from MSAs
   • Models of evolutionary change
   • Phylogeny computational algorithms
   • Tree visualization tools

W6: Pipeline Catchup

W7: Community diversity
   • Calculations of alpha diversity (richness, evenness, rarefaction)
   • Beta diversity (what it is and what it isn’t)
   • Community distance metrics (Bray-Curtis, Unifrac)

W8: Pipeline Catchup

W9: Community Visualization
   • Visualization of community composition (heat maps, bar charts, krona)
   • Visualization of community variation (clustering and ordination)

W10: Spring Break

W11: Community Statistics (Testing distance matrices)
   • PERMANOVA and ANOSIM
• DISTLM and Mantel Testing (incorporating outside parameters)
• Community dispersion and interpretation

W12: Population Statistics (Testing OTUs)
• Linear models (ANOVA, regression, etc.)
• Meeting distributional assumptions
• Discriminant models (SIMPER, Random Forest, Indicator Species)
• Volcano plots and interpretations
• Networking (Correlation and Co-occurrence)

W13: Servers and Logistics for Efficiency
• AWS and Google Cloud
• MiniLIMS
• Galaxy
• UHHPC
• C-MAIKI tools

W14: Holiday – Good Friday

W15: Metagenomics
• Methods
• Difficulties
• Pipelines
• Databases
• Statistical analyses