http://www.mbl.edu/it/itservices/network/wireless/mbl-register/

choose MBL-PRIVATE

user: dm00834

password: dm00834

http://registration.mbl.edu/
Welcome to STAMPS 2015

Strategies and Techniques for Analyzing Microbial Population Structure

Directors:
David Mark Welch @ Marine Biological Laboratory
Mitch Sogin @ Marine Biological Laboratory

Course Date: August 5 - August 15, 2015

Twitter hash: #STAMPS15

Deep DNA sequencing using massively-parallel, next-generation technology has enabled nearly comprehensive environmental surveys that can describe the different kinds of microbes in a community and their relative abundance. These descriptions of richness and evenness make possible estimates of microbial diversity, but the size of the required data sets pose enormous computational challenges. The rapidly expanding flow of information from next generation DNA sequencing platforms has fueled healthy debate about best practices for data analysis while at the same time building a user demand for tools that can address important ecological questions. The STAMPS course will promote dialogue and the exchange of ideas between experts in analysis of metagenomic data and offer interdisciplinary bioinformatic and statistical training to practitioners of molecular microbial ecology and genomics.

Topics to be covered include but are not limited to next-generation strategies for analysis of microbial communities; acquisition and organization of next generation sequence data; principles of quality control of sequence data; the theory of cluster and rarefaction analyses; taxonomic assignments for high-throughput data; statistical models for estimating microbial diversity; microbial community comparison

Announcements

- Editing the wiki: Log in by clicking on the link in the upper right corner. Follow the instructions, including the email address confirmation. You can elect to change your password; it is not automatically synced to your server password.
- Lectures and Labs will be in Loeb G70
NOTE: Times of meals and morning/afternoon breaks are approximate!

This schedule is subject to change!

First Week

<table>
<thead>
<tr>
<th>Time</th>
<th>Thursday 6 August</th>
<th>Friday 7 August</th>
<th>Saturday 8 August</th>
<th>Sunday 9 August</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30-12:00</td>
<td>Mitch Sogin Introduction</td>
<td>Sue Huse</td>
<td>Tracy Teal</td>
<td>Jennifer Martiny</td>
</tr>
<tr>
<td></td>
<td>David Mark Welch Overview</td>
<td></td>
<td></td>
<td>Beta-diversity, biogeography, and microbial traits</td>
</tr>
<tr>
<td></td>
<td>Sue Huse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Illumina Quality Filtering</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00-14:00</td>
<td>Lunch break (tour of Keck Facility)</td>
<td>Lunch break (tour of Keck Facility)</td>
<td>Lunch break</td>
<td>Lunch break</td>
</tr>
<tr>
<td>14:00-17:00</td>
<td>Sue Huse</td>
<td>Sue Huse</td>
<td>Tracy Teal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chimeras in Amplicon Sequencing</td>
<td>Intro to R</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assigning Taxonomy to Marker Genes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clustering OTUs by Sequence Similarity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:00-19:00</td>
<td>Dinner break (variable)</td>
<td></td>
<td></td>
<td>Course Banquet 17:30-21:30</td>
</tr>
<tr>
<td>19:00-22:00</td>
<td>David Mark Welch and TAs 2014 Unix tutorial</td>
<td>John Bunge and Amy Willis Alpha and Beta Diversity theory and practice</td>
<td>Susan Holmes PhyloSeq</td>
<td></td>
</tr>
</tbody>
</table>
# Faculty

## 2015 Course Faculty

- **John Bunge**, Cornell University
- **A. Murat Eren**, Marine Biological Laboratory
- **Susan Holmes**, Stanford University
- **Sue Huse**, Brown University
- **Curtis Huttenhower**, Harvard University
- **Rob Knight**, University of California, San Diego
- **Morgan Langille**, Dalhousie University
- **David Mark Welch**, Marine Biological Laboratory
- **Jennifer Martiny**, University of California, Irvine
- **Hilary Morrison**, Marine Biological Laboratory
- **Mihai Pop**, University of Maryland
- **Christopher Quince**, University of Glasgow
- **Mitch Sogin**, Marine Biological Laboratory
- **Tracy Teal**, Michigan State University

## 2015 Course TAs

- **Galeb Abu Ali**, Harvard University
- **Ben Callahan**, Stanford University
- **Adam Robbins-Planka**, University of Colorado at Boulder
- **Will Van Treuren**, University of Colorado at Boulder
- **Yoshiki Vázquez Baeza**, University of California, San Diego
- **Amy Willis**, Cornell University
# Participants

To add a brief description of your interests, choose edit from the tab above and add text to the line below your email, leaving the "I" at the beginning of the line. **Do not** add extra lines beginning with "I"

Feel free to add a link to your own wiki page or home institution web page.

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>E-mail</th>
<th>Interests</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nicole Adams</td>
<td>University of Southern California</td>
<td>nicoleea AT usc DOT edu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deepta Agashe</td>
<td>National Center for Biological Sciences</td>
<td>dagashe AT nbcs DOT res DOT in</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emrah Altindis</td>
<td>Joslin Diabetes Center</td>
<td>altindise AT gmail DOT com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Julia Baum</td>
<td>University of Victoria</td>
<td>baum AT uvic DOT ca</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luc Bijnens</td>
<td>Janssen Pharmaceutical Companies of Johnson and Johnson</td>
<td>LBijnens AT its DOT jnj DOT com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eelke Brandsma</td>
<td>University Medical Centre Groningen</td>
<td>e DOT t DOT brandsma AT umcg DOT nl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anya Brown</td>
<td>University of Georgia</td>
<td>anyabrown AT uga DOT edu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lucelia Cabral</td>
<td>University of Campinas - UNICAMP</td>
<td>luc DOT g DOT cabral AT gmail DOT com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Egle Cekanaviciute</td>
<td>University of California San Francisco</td>
<td>egle DOT cekanaviciute AT ucsf DOT edu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wen-Chi Chou</td>
<td>Broad Institute</td>
<td>wcc957 AT gmail DOT com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bruno Chumpitazi</td>
<td>Baylor College of Medicine</td>
<td>chumpita AT bcm DOT edu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anne Dekas</td>
<td>Lawrence Livermore National Laboratory</td>
<td>aedekas AT gmail DOT com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joshua Denny</td>
<td>University of Louisville</td>
<td>joshua DOT denny AT louisville DOT edu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natalie Fischer</td>
<td>Institut Pasteur Paris</td>
<td>Natalie DOT Fischer211 AT gmail DOT com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jessica Grembi</td>
<td>Stanford University</td>
<td>jgrembi AT stanford DOT edu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sergio Guajardo-Leiva</td>
<td>Pontificia Universidad Catolica de Chile</td>
<td>seguajardo AT uc DOT cl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leah Guthrie</td>
<td>Albert Einstein College of Medicine</td>
<td>ltcguthrie AT gmail DOT com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Julia Haas</td>
<td>Umea University</td>
<td>julia DOT haas AT umu DOT se</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theresa Hattenrath-Lehmann</td>
<td>Stony Brook University</td>
<td>theresa DOT hattenrath AT stonybrook DOT edu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vicki Hertzberg</td>
<td>Emory University</td>
<td>vherzb AT emory DOT edu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kuan-Hsi</td>
<td>Northwestern University</td>
<td>shih AT uw DOT uw</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Strategies for Analysis of Microbial Population Structures

• Tag/Marker Gene Metagenomics
  (sequencing amplified markers, “amplicons” e.g. SSU rRNA)
  – Who’s in the community?

• Shotgun Metagenomics (sequencing total community DNA)
  – What is the functional potential of the community?
  – (Partial) Assembly of community genomes

• Metatranscriptomics (sequencing tag or shotgun cDNA)
  – What fraction of the community is active?
  – What is the community doing?

• Metaproteomics (sequencing community peptides)

• Meta-metabolomics
  (detecting and quantifying metabolites used by the community)

Total Community Sampling vs Targeted or Enriched Sampling
Tag/Marker Gene Sequencing Metagenomics

0. Design your experiment to meet your goals!
   - Sampling depth (more reads/sample) vs. sampling breath (more samples)
   - Targeted approach vs trying to sample everything
   - Importance of rare taxa?
   - Importance of biological replicates
   - Collect appropriate contextual metadata
   - Standardize experimental approach **before you begin**
1. Design primers to a region of interest
   - Usually SSU rRNA (“16S”)
   - conserved primers flanking variable regions(s)
   - length dependent on sequencing platform

2. Extract DNA (RNA)
   (2a Create cDNA from RNA)

3. Amplify DNA

All of these steps introduce biases!

4. Sequence
Also biased, but probably less so than the earlier steps!
5. Clean up the data
   - Remove low quality reads
   - Remove other error-prone reads
   - Remove chimeric reads
   - Remove non-target reads

6. Organize the reads into units
   - Assign reads to a taxonomic identifier
   - Assign reads to a phylogenetic clade
   - Group reads into (or assign reads to) Operational Taxonomic Units (OTUs)
     • By sequence similarity
     • By sequence information

Sue Huse, Today!
Tracy Teal, Sat
Susan Holmes, Sat
Rob Knight, Tues
Meren, Wed
Chis Quine, Wed
Tag/Marker Gene Sequencing Metagenomics

7. Analyze your groups of sequences
   - Assumption: each read represents an individual
   - Assumption: steps 1-6 are effectively unbiased, or equally biased across samples
   - Quandary: what to do about low abundant groups (singletons)?
   - Quandary: how to compare samples of unequal size?

Within a sample or combined samples
   - Richness, number of observed and estimated groups
   - Evenness, relative abundance of the groups

Between samples or groups of samples
   - Beta diversity
   - Relating diversity to metadata
An Abridged History of DNA Sequencing

1971 Wu & Taylor sequence overhang of phage λ (12 bp)
1974 Sogin, Woese and Pace sequence a 5S rRNA (116 nt)
1977 Sanger et al. sequence phiX174 (5,375bp)
1982 Sanger et al. sequence phage λ (48,501bp)
1985 Pace, Olsen, Sogin, and others sequence 16S rRNA genes
1996 ABI Prism 310 Genetic Analyzer
1998 Phred automated base calling becomes standard
1998 ABI 3700; high-throughput capillary sequencing becomes widely available
2005 454 GS20
2006 Solexa Genome Analyzer
2011 Illumina HiSeq
Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts.
The Short Read Archive (SRA)

Next-generation sequencing technologies

- Ion Torrent
- Oxford Nanopore
- PacBio
- SOLiD
- Illumina
- 454 GSFLX
Q: To what degree does your research drive next-gen sequencing technology?
The Pattern of Perception

The Collins Curve

initial idea -> hype -> reality check -> true acceptance performance

enthusiasm -> disillusionment

technology 1  technology 2  technology 3
75,000,000 reads

375,000,000 reads per lane
fastq format
fastq format

- machine
- run
- flowcell
- lane
- tile
- x:y coordinates
- pair (1, 2, 0)
- failed filter?
- multiplex ID

Phred quality score from 0 to 93 using ASCII 33 to 126

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>34</td>
<td>35</td>
<td>36</td>
<td>65</td>
<td>66</td>
<td>67</td>
<td>68</td>
<td>124</td>
<td>125</td>
</tr>
<tr>
<td>!</td>
<td>&quot;</td>
<td>#</td>
<td>$</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Phred Quality Scores

\[ Q = -10 \times \log(P) \]

<table>
<thead>
<tr>
<th>Q</th>
<th>P</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.10</td>
<td>1 in 10</td>
</tr>
<tr>
<td>20</td>
<td>0.01</td>
<td>1 in 100</td>
</tr>
<tr>
<td>30</td>
<td>0.001</td>
<td>1 in 1,000</td>
</tr>
<tr>
<td>40</td>
<td>0.0001</td>
<td>1 in 10,000</td>
</tr>
</tbody>
</table>

In the era of high-throughput capillary sequencing, the “gold standard” to which next-gen methods are measured, the goal was to produce reads with an average phred score of 20.
Tag Sequencing Metagenomics

7. Analyze your groups of sequences
   - Assumption: each read represents an individual
   - Assumption: steps 1-6 are effectively unbiased, or equally biased across samples
   - Quandary: what to do about low abundant groups (singletons)?
   - **Quandary: how to compare samples of unequal size?**

Within a sample or combined samples
   - Richness, number of observed and estimated groups
   - Evenness, relative abundance of the groups

Between samples or groups of samples
   - Beta diversity
   - Relating diversity to metadata
Tags Sampled vs OTUs Observed

Rarefaction Curve

- Y-axis: OTUs Observed
- X-axis: Tags Sampled

The curve illustrates the relationship between the number of tags sampled and the number of observed OTUs. As the number of tags sampled increases, the number of observed OTUs also increases, but at a decreasing rate, indicating that rarefaction helps in estimating the number of species present in the sample.
Rarefaction Curve: what is it?

original sample has 78,650 tags in 3,320 OTUs
Rarefaction Curve: what is it?

- Original sample has 78,650 tags in 3,320 OTUs.
- Create pseudo-replicate sample with n tags by removing 78,650 – n tags.
- Count how many OTUs have at least one tag.
- Repeat for n = 100, 200, 300 ... 78,500 or other interval.

18,000 tags, 1,604 OTUs still represented (1,635 s.d. 27)
Rarefaction Curve

High abundance OTUs always present in pseudo replicates

Many low-abundance OTUs always missing in pseudo replicates
Same tag data, but:
“sub-samples” of tags clustered into OTUs
rarefaction run on these new OTUs
Same tag data, but:
“sub-samples” of tags clustered into OTUs
rarefaction run on these new OTUs

This should not happen!
The solution is not to sub-sample to some common tag count to avoid the problem!
This “sub-sampling” phenomenon indicated something was not happening the way it was supposed to happen. There was a problem with the procedure. An improved clustering method generates OTUs that do not display the phenomenon.

Understand what the methods do so that you can recognize when there is a problem!
Tag Sequencing Metagenomics

7. Analyze your groups of sequences
   - Assumption: each read represents an individual
   - Assumption: steps 1-6 are effectively unbiased, or equally biased across samples
   - Quandary: what to do about low abundant groups (singletons)?
   - Quandary: how to compare samples of unequal size?

   Within a sample or combined samples
   - Richness, number of observed and estimated groups
   - Evenness, relative abundance of the groups

   Between samples or groups of samples
   - Beta diversity
   - Relating diversity to metadata
International Census of Marine Life
http://icomm.mbl.edu/

10.2 million bacterial sequence tags from 549 surveys
Species Abundance of 3% OTUs from the ICoMM Survey

48% singletons!

Global survey of crustaceans across 93 coral reef sites...
38% singletons!
Richness Distributions of 100 Most Abundant OTUs

OTUs

- Singletons
- Doubletons
- 3-5
- 6-10
- 11-50
- 51-100
- 101-500
- > 500

Bacteria102.03.otusizes.icomm
The most abundant OTUs are rare somewhere

- If an OTU appears with abundance greater than 100 in one dataset, there is a ~90% chance it appears as a singleton in another

Most singletons in a dataset are found elsewhere

- 79% of the singleton OTUs in a dataset are not unique
What is a singleton?

~$1 \times 10^9$ bacterial cells/liter epipelagic seawater

Based on average ICoMM dataset abundances:

- 1-3 OTUs represent >1% each, or more than 100,000,000 cells/L
- ~10 OTUs represent 0.1% – 1%, or 10,000 – 100,000,000 cells/L
- ~100 OTUs represent 0.01 – 0.001%, or 1,000 – 10,000,000 cells/L
- ~1,000 OTUs represent less than 0.001%, that’s still ~ 100,000 cells/L
Our Sampling is *Very Incomplete*, even with 20,000 Tags / Liter

![Graph showing rank abundance with a decreasing curve from high to low values.](image_url)
What is a Singleton?

If sampling 20,000 tags from 1 L at $10^9$ cells/L, a singleton likely represents an OTU with an abundance of 50,000 – 100,000 cells/L.
Is there a real difference in abundance between an OTU observed as singleton and as a doubleton?
Is there a real difference in abundance between an OTU observed as singleton and as a doubleton?
What’s Represented in 20,000 tags?  
(assuming sampling 20,000 tags from 1 liter at $10^9$ cells/liter)

- Singletons
- Doubletons
- Missing!
Understand the methods you are using; recognize when the answers don’t make sense

Appreciate the biology of your system; recognize when your results are in conflict and be suspicious of your results when this happens

Document everything you do