ECOGEO Workshop 2: Introduction to Env ‘Oomics

Amplicon Analysis - Oligotyping
A. Murat Eren (Meren)
http://merenlab.org
“An arbitrary amount of similarity between 16S reads is enough to group sequences into ecologically relevant units”

Gardnerella vaginalis story.
(genus level bacterial community composition of three healthy women)  

(genus level bacterial community composition of three women with BV)
Reads that are binned into a ‘genus’, ‘species’, or a 3% OTU, usually have very subtle, yet systematical variation.
Shannon entropy results for *Gardnerella vaginalis* alignments:
**Oligotypes in Samples:**

<table>
<thead>
<tr>
<th>Oligotype</th>
<th>P01</th>
<th>P02</th>
<th>P03</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCCACAGAT</td>
<td>169</td>
<td>220</td>
<td>172</td>
</tr>
<tr>
<td>CCCTCGATA</td>
<td>88</td>
<td>26</td>
<td>85</td>
</tr>
<tr>
<td>TTTTCGAAT</td>
<td>13</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>TTTACGAGG</td>
<td>9</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>CCTACGAGA</td>
<td>6</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>TTTACGAGA</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>TTCATGAGG</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>CCCACGGGA</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>CCCACGGAT</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

(...)

(Additional oligotypes are not shown in this view.)
http://merenlab.org/software/oligotyping

Carl Woese. Photo courtesy of Don Hamerman (and Robin Tecon, who published it on his blog).
A Phylogenetic Definition of the Major Eubacterial Taxa

C. R. WOENSE1, E. STACKEBRANDT2, T. J. MACKE3, and G. E. FOX4

1 Department of Genetics and Development, University of Illinois, Urbana, Illinois 61801, USA
2 Institut für Allgemeine Mikrobiologie, Christian-Albrechts-Universität, 2300 Kiel, Federal German Republic
3 Department of Computer Science, University of Illinois, Urbana, Illinois 61801, USA
4 Department of Bioophysical Sciences, University of Houston, Houston, TX 77004, USA

Received January 30, 1985

Summary

Through oligonucleotide signature analysis of 16S ribosomal RNAs, it is possible to define ten major groups of eubacteria. These are:
1. the Gram positive bacteria,
2. the purple photosynthetic bacteria and their relatives,
3. the spirochetes and their relatives,
4. the sulfur-dependent eubacteria and their relatives,
5. the bacterioles, flavobacteria and cytophagas and their relatives,
6. the cyanobacteria,
7. the green sulfur bacteria,
8. the green non-sulfur bacteria and their relatives,
9. the radio-resistant micrococi, and
10. the planctomyces and their relatives.

Although no consensus exists as regards the taxonomic terminology, these ten groupings are appropriately termed eubacterial Phyla or Divisions. The major subdivisions of those Phyla or Divisions that have been extensively characterized can also be defined by characteristic oligonucleotide signatures.

Key words: Eubacteria – Phylum – Division – Phylogeny – Taxonomy – Evolution – Oligonucleotide Signature – Ribosomal RNA
Certain regions in the rRNA molecule are highly conserved in sequence – some positions therein showing no variation whatever among all organisms, other positions varying in composition, but infrequently. For reasons that are not completely understood (Woese et al., 1985b) changes involving these regions of conserved sequence tend to occur mainly during the evolution of major phylogenetic groups. Because of this, sequence in these regions tends to be characteristic of the major taxonomic groups, and so is useful in defining them. Families of sequence variation occurring at these phylogenetically significant locales are shown in Table 1. A set of oligonucleotides that unequivocally defines each of the eubacterial “phyla” can be extracted from these data.
(6) The Cyanobacteria

The grouping is well defined by oligonucleotides at positions 365, 795, 1210 and 1240. Too few species of this group have been characterized by cataloging, however, to define its internal structure well (Doolittle et al. 1975).

(8) The Green Non-sulfur Bacteria and Relatives

This “phylum” presents one of the more distinctive oligonucleotide signatures; see positions 50, 315, 910 and 1225. Whether this means that Chloroflexus and its relatives represent a particularly deep branching in the

(10) The Planctomyces Group

Of all the eubacterial “phyla” this one is the most distinct in terms of its oligonucleotide signature – e.g. positions 340, 570, 935, 960, 985 and 1240 – a uniqueness also reflected in the exceptionally low Sab values planctomyces show with other eubacterial rRNAs.
Couple 01, Vaginal and Penile Sample
Couple 02, Vaginal and Penile Sample
Couple 03, Vaginal and Penile Sample
Couple 04, Vaginal and Penile Sample
Couple 05, Vaginal and Penile Sample
Couple 06, Vaginal and Penile Sample
Couple 07, Vaginal and Penile Sample
Couple 08, Vaginal and Penile Sample
Hmm.
(as in “I have mixed feelings about this”)
More examples from the literature if you would like to read more:

Original Article
A single genus in the gut microbiome reflects host preference and specificity
A. Murat Eren1, Mitchell L. Sogin1, Hilary G. Morrison1, Joseph H. Vineis1, Janny C. Park1, Ryan J. Newton2 and Sandra L. McLellan2
1Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, USA and 2School of Physiology, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Oligotyping analysis of the human oral microbiome
A. Murat Eren1, Gary G. Borisy3,4, Susan M. Huse4, and Jessica L. Mark Welch4,5
1Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, USA; 2Department of Microbiology, The Forsyth Institute, Cambridge, MA 02142, and 3Department of Pathology and Laboratory Medicine, Brown University, Providence, RI 02912

Research Article
Sewage Reflects the Microbiomes of Human Populations
Ryan J. Newton, Sandra L. McLellan, Deborah K. Dills, Joseph H. Vineis, Hilary G. Morrison
School of Freshwater Sciences, University of Wisconsin—Milwaukee, Milwaukee, Wisconsin, USA; Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, USA

Original Article
Diverse, rare microbial taxa responded to the Deepwater Horizon deep-sea hydrocarbon plume
Sara Kleindienst1,4, Sharon Grim2,3, Mitchell Sogin2, Annalisa Bracco1, Melitza Crespo-Medina1,4 and Samantha B Joye3
1Department of Marine Sciences, University of Georgia, Athens, GA, USA; 2Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, USA and 3School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA, USA

But I will give you my favorite.
Eren et al. (2014). ISMEJ.

Hexadella dedritifera  Hexadella cf. dedritifera  Water Column

With Taxa

Gammaproteobacteria
Hexadella dedritifera  Hexadella cf. dedritifera  Water Column

With 3% OTUs

OTU 0

Eren et al. (2014). ISMEJ.
OK. One more.
Alluvial diagram of the most abundant 100 MED nodes in oral microbiome dataset (%83 of the entire dataset).

Eren et al. (2015). ISMEJ.
mothur and QIIME are awesome.
Oligotyping and MED are not the only algorithms you can achieve single-nucleotide resolution.
Distribution Based Clustering

Combines ecological insights obtained from the distribution patterns of sequences across samples with the genetic distance of sequences to identify ecologically distinct groups of reads regardless of sequence similarity.

Paper :: http://aem.asm.org/content/79/21/6593.full
Code :: https://github.com/spacocha/dbOTUcaller
SWARM

An agglomerative clustering algorithm that uses pairwise sequence similarities to form single-linkage clusters of reads that occur in the same $d$-neighborhood in a hypothetical sequence-distance space, where $d$ represents the number of nucleotides that can differ between two reads for them to still be considered in the same cluster.

Paper :: http://aem.asm.org/content/79/21/6593.full
Code :: https://github.com/spacocha/dbOTUcaller
CFF (Cluster-free Filtering)

A cross-sample correlation analysis to collapse error clouds around candidate sequences of individually de-noised samples.

Paper :: http://nature.com/ismej/journal/v9/n1/full/ismej2014117a.html
Code :: https://github.com/hepcat72/CFF
DADA2

A divisive partitioning algorithm that identifies real sequences in a dataset by modeling and correcting sequencing errors.

Paper :: http://www.nature.com/nmeth/journal/v13/n7/full/nmeth.3869.html
Code :: https://github.com/benjjneb/dada2

phloseq! https://joey711.github.io/phyloseq/
My time is probably up at this point, so please go through the rest of the presentation yourselves.

Thank you very much for your time and attention!
Meren’s two cents
“High-resolution methods assume sequencing errors are distributed across reads randomly, but sequencing errors are not random”.

They aren’t? Well that is great then! Because in that case we can model them, and factor them in in our analyses.
How about PCR errors?

There is so little we can do about them bioinformatically. But we can’t keep using 3% OTUs simply because they mitigate these errors if the damage they cause is more than the benefits.
Is there such a thing as “too sensitive”? How do we manage ‘neutral variation’?

“Neutral” is a misleading word. I think this is a much better way to say what most people use “neutral variation” for: “Variation across microbial community structures that we can’t explain using any of the measurements available to us”. It is real, and scary. But is it not something we need to understand, and not something to hide behind arbitrary similarity thresholds? Since we’re not .. engineers?
What about the rare biosphere?

So far the rare biosphere has been either captured and was hidden within 3% OTUs, or was mixed with PCR and sequencing errors in the long tail.

Studying rare biosphere requires much more than how we have been studying microbial community structures.
All this effort goes into 16S… In the era of metagenomics, is it really good for anything?

This was my 2015 response: Yes, when it is used right, it is indispensable.

This is my response now: Well, definitely good for your budget.