Baldridge Metagenomics Lectures

• Originally scheduled for March 23rd and 25th
• Important disclaimer: in these chaotic times, going for “good enough”...
• An enormous thank-you to Gautam Dantas!
METAGENOMICS: The Genomic Study of Uncultured Organisms

A research field
&
A set of research techniques

MERRIAM-WEBSTER
Main Entry: meta-
Function: prefix
Meaning: more comprehensive : transcending
—usually used with the name of a discipline to designate a new but related discipline designed to deal critically with the original one
Metagenomics: A Set of Research Techniques

• Experimental and computational methods
• Overcome diversity and unculturability of most microbes
• Maximize understanding of genetic composition and activities of communities
• Representatively sample complexity that cannot be exhaustively characterized
Metagenomics: A Research Field

- Understand biology at the aggregate level, wherein multiple organisms likely contribute to a particular function, effect or phenotype
- Transcends the individual organism
- Describe community function
- Understand networks of genes working across organisms to serve collective functions
- Systems biology for communities and the biosphere
The Tyranny of Koch’s Postulates (1884): Proof of Microbial Causation

• The microorganism must be found in abundance in all organisms suffering from the disease, but should not be found in healthy animals.
• The microorganism must be isolated from a diseased organism and grown in pure culture.
• The cultured microorganism should cause disease when introduced into a healthy organism.
• The microorganism must be re-isolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.
The Pure-Culture Paradigm

• **Axenic cultures** (free from living organisms other than the species required) are the gold standard for good microbiological practice
• Multiple species in culture = contamination
• “Planktonic” growth in rich media = “natural”
• Slimy complex biofilm = “accident” and “unworthy of scientific study”
• Genomics: reinforced use/need for pure culture
Microbes are traditionally studied by culture-dependent methods.

99 - 99.9% of bacteria in most habitats are not easily cultured!

“...animalcules were in such enormous numbers, that all the water...seemed to be alive.” — van Leeuwenhoek (1683)
Microbial communities are highly diverse, abundant, and serve important functions in almost every environment.
soil

as
treatment

human

agr
culture

food
animals

manure

runoff

rainfall

sludge

wastewater
treatment

vegetables
and fruit

direct contact
and meat

drinking water
and swimming

fish

industry and
households

Highly connected network of microbial communities across habitats

adapted from: Dantas and Sommer, American Scientist (2014)
In addition to bacteria, metagenomic communities include fungi, helminths, eukaryotic and prokaryotic viruses, etc.

- Bacteria are by far the best-studied component of microbial/metagenomic communities
- Depending on methods used, your analysis technique may or may not be powered to look at other types of microbes

Rowan-Nash et al., *Microbiol Mol Biol Rev.* 2019
METAGENOMICS: study microbial communities through direct DNA sequencing

- Susceptibility Testing
- Genome Sequencing
- PCR-based Screens
- Direct DNA Sequencing

Culture-Dependent Methods

Culture-Independent Methods
The Great Plate-Count Anomaly (Staley and Konopka, 1985)

1. Collect organism sample from water or soil.
2. View in microscope.
3. Culture sample on agar.
5. Count cells.
Molecular Phylogeny: 16S/18S SSU rRNA

Most Life and Most Biological Evolution is Microbial

- Multicellular lineages (red) rare, not diverse as measured by SSU rRNA

- Most molecular diversity can be found in microbes

- Most (>99%) microbes can’t be cultured: known only from sequences

Figure adapted from Pace 1997 Science 276:734-740.
It’s a Prokaryotic World!!!

- Number of Species: Millions?
- Number of Individuals/Cells: $5 \times 10^{30}$
- Largest reservoirs of biomass:
  - 350-550 Petagrams of Carbon ($1 \text{ Pg} = 10^{15} \text{ g}$)
  - 85-130 Pg of Nitrogen
  - 9-14 Pg of Phosphorous
- Diversity of Habitats: 340°C, radioactive, 6km underground, soil, ocean, terrestrial subsurface
- Diversity of Cellular Chemistries
- Environmental Impact: remediation, disease, commensals, energy, food, photosynthesis
Prokaryotic Genomes Are Rapidly Being Sequenced!

- Mar 11th, 2020 (GOLD, JGI):
  - ~163,000 prokaryotic genomes completed or in progress

(https://gold.jgi.doe.gov/)
Two Means of Bacterial Evolution

Widespread HGT Can Obscure Phylogenetic Inference


conserved nitrogenase operon
It’s WAY more diverse than we thought

"What's in a name? That which we call a rose
By any other name would smell as sweet.”

*Romeo and Juliet (II, ii, 1-2), W. Shakespeare*

- **K12**: Harmless lab strain
- **CTF073**: Uropathogenic strain, causes 70–90% of the 7 million cases of acute cystitis and 250,000 cases of pyelonephritis annually in US
- **O157:H7**: Enterohemorrhagic, 3-5% fatality
- Only 39.2% of their combined (nonredundant) set of proteins actually are common to all three strains!
- **SAME SPECIES BY 16S rRNA!**
Sequence-based Metagenomics
DNA Sequencing Keeps Getting Cheaper

DNA sequencing costs continue to drop much faster than Moore’s Law.

Note *useful* costs (e.g. finishing a genome) are higher than *raw* costs.

NHGRI; [http://www.genome.gov/sequencingcostsdata/](http://www.genome.gov/sequencingcostsdata/)
Required:

MASSIVE COMPUTING RESOURCES
Sequence-based Metagenomics

Sample

Extract DNA

PCR Amplify (sample specific barcodes)

Sequence

16S rRNA gene-based surveys

Shotgun sequencing of metagenome/transcriptome

Total DNA/cDNA

Sequence

Who’s there?

What are they doing?
Sequence-based Metagenomics: 16S rRNA based Characterization of Bacterial Communities
16S sequencing redefined the tree of life

16S rRNA / SSU sequencing

• 16S rRNA forms part of ALL bacterial ribosomes.

• Contains regions of highly conserved and highly variable sequence.

• **Variable sequence** can be thought of as a molecular “fingerprint”—can be used to identify bacterial genera and species.

• Large **public databases available for comparison**.—Ribosomal Database Project contains >3.3 million rRNA sequences.

• **Conserved regions** can be targeted to amplify broad range of bacteria from environmental samples.

• **Not quantitative** due to copy number variation

*Circumvents the need to culture*

*Erlandsen S L et al. J Histochem Cytochem 2005;53:917-927*

*Alan Walker, Sanger*
Specific hyper-variable regions (**tags**) are sequenced

<table>
<thead>
<tr>
<th>Region</th>
<th>Position</th>
<th># b.p.</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>69-99</td>
<td>30</td>
</tr>
<tr>
<td>V2</td>
<td>137-242</td>
<td>105</td>
</tr>
<tr>
<td>V3</td>
<td>338-533</td>
<td>195</td>
</tr>
<tr>
<td>V4</td>
<td>576-682</td>
<td>106</td>
</tr>
<tr>
<td>V5</td>
<td>822-879</td>
<td>57</td>
</tr>
<tr>
<td>V6</td>
<td>967-1046</td>
<td>79</td>
</tr>
<tr>
<td>V7</td>
<td>1117-1173</td>
<td>56</td>
</tr>
<tr>
<td>V8</td>
<td>1243-1294</td>
<td>51</td>
</tr>
<tr>
<td>V9</td>
<td>1435-1465</td>
<td>30</td>
</tr>
</tbody>
</table>


16S by NGS experimental workflow

Extract DNA and amplify marker gene tag with barcoded primers

Pool amplicons and sequence tags

Visualize and compare community relationships

Assign millions of tags to “SPECIES” (OTUs/ASVs)
How do we define a species?

“No single definition has satisfied all naturalists; yet every naturalist knows vaguely what [s]he means when [s]he speaks of a species”
How do we define a species for tag/amplicon data?

Species concept works for sexually reproducing organisms

- Breaks down when applied to bacteria and fungi
  - Plasmids
  - Horizontal gene transfer
  - Transposons/Viruses

- Attempt to bin amplicon sequences to estimate relative abundance of different taxa in a sample, using OTU or ASV definitions
OTUs and ASVs

- **Operational Taxonomic Unit (OTU)**
  - An arbitrary definition of a taxonomic unit based on sequence divergence
  - OTU definitions matter
  - QIIME-uclust, MOTHUR, and USEARCH-UPARSE

- **Amplicon Sequence Variant (ASV)**
  - Attempts to reconstruct the exact biological sequences present in the sample
  - “Zero-noise” OTUs
  - DADA2, Qiime2-Deblur, and USEARCH-UNOISE3

Caron, Applied & Env Micro 2009, 75(18):5797-808

Prodan, Plos One 2020, 15(1):e0227434
How do we define a species for tag data?

• Search for sequence similarity between 16S/18S variable regions (e.g. V3-V4) or particular genes (e.g. rpoB)

• These genes are ‘housekeeping’ genes which are less likely to be involved in horizontal transfer

• However, note that 16S/18S sequences are known to have variable copy numbers which can bias results in “counts”

www.mlst.net
Binning tags

Tags may be analysed in one of two ways:

- **Composition-based binning**
  - Most helpful if analysing a sample in which expect novel species
  - Relies on comparisons of gross-features to species/genus/families which share these features
    - GC content
    - Di/Tri/Tetra/... nucleotide composition (kmer-based frequency comparison)
    - Codon usage statistics

- **Similarity-based binning**
  - Requires that most sequences in a sample are present in a reference database
    - Direct comparison of OTU sequence to a reference database
    - Identity cut-off varies depending on resolution required (though inexact)
      - “Family” - 80%
      - “Genus” - 90%
      - “Species” - 97%
    - Multiple marker genes used for finer sub-strain identification (MLST)
    - Too stringent cut-off selection will lead to excessive diversity being reported
      - Sequencing errors
      - Sample prep issues
Extract DNA and amplify marker gene with barcoded primers

Pool amplicons and sequence

Assign millions of sequences from thousands of communities to OTUs/ASVs

Assign reads to communities

Visualize and compare community relationships

Your favorite bioinformatic pipeline
QIIME workflow schematic

Sequencing output
(454, Illumina, Sanger)
fasta, qual, sff/trace files

Metadata
mapping file

Pre-processing
e.g., remove primer(s), demultiplex,
quality filter

Denoise 454 Data
PyroNoise, Denoiser

MG-RAST Submission
Submit sequences and
metadata to MG-RAST

OTU Table
(i.e., per sample OTU
counts)

Phylogenetic Tree
Evolutionary relationship
between OTUs

α-diversity and rarefaction
e.g., Phylogenetic
Diversity, Chao1,
Observed Species

β-diversity and rarefaction
e.g., Weighted and
unweighted UniFrac, Bray-
Curtis, Jaccard

Visualization
e.g., 2D and 3D PCoA plots, distance histograms, taxonomy pie charts, rarefaction plots, OTU network visualization, jackknifed hierarchical clustering.

Pick OTUs and representative sequences
Reference based
BLAST, UCLUST

De novo
e.g., UCLUST, CD-HIT, MOTHUR

Assign taxonomy
BLAST, RDP
Classifier

Align sequences
e.g., PyNAST,
INFERNAL, MUSCLE,
MAFFT

Build ‘OTU table’
i.e., per sample OTU
counts

Build phylogenetic tree
e.g., FastTree, RAxML,
ClearCut

beta_diversity_through_plots.py

$ beta_diversity_through_plots.py -i otu_table.txt -m Fasting_Map.txt -o betadivtree -t rep_set.tre
dada2: ASVs instead of OTUs

Advantages

Resolution: ASVs resolve biological differences of even 1 or 2 nucleotides.
Accuracy: fewer false positive sequence variants than other methods report false OTUs.
Comparability: ASVs can be directly compared between studies, without reprocessing.
Computational Scaling: Compute time scales linearly with sample number

https://benjjneb.github.io/dada2/index.html
Measuring diversity of OTUs/ASVs

Two primary measures for sequence based studies:

• Alpha diversity
  – What is there? How much is there?
  – Diversity *within* a sample

• Beta diversity
  – How similar are two samples?
  – Diversity *between* samples
Measuring diversity: alpha

Alpha diversity

• Diversity *within* a sample
• Simpson’s diversity index (also Shannon, Chao indexes)
• Gives less weight to rarest species

\[
D = 1 - \frac{\sum_{i=1}^{S} n_i(n_i - 1)}{N(N - 1)}
\]

*S* is the number of species

*N* is the total number of organisms

*n*<sub>i</sub> is the number of organisms of species *i*

Measuring diversity: beta

Beta diversity

- Diversity *between* samples
- Sorensen’s index

\[ \beta = \frac{2c}{S_1 + S_2} \]

*S*₁ is the number of species in sample 1
*S*₂ is the number of species in sample 2
*c* is the number of species present in both samples

Beta diversity: UniFrac

Beta diversity
• Diversity *between* samples
• UniFrac distance (*shared branch length on phylo tree*)
• Percentage observed branch length unique to either sample

Clustering by Unifrac distance

Distance Matrix

PCoA

Hierarchical Cluster
An early example of 16S community analysis (by 454)
Mammalian Microbiomes Cluster by Diet

- Circles are different mammals, colors different categories
- Grey rounded squares are shared OTUs
- Thinner edge = fewer sequences
- Thicker edge = more sequences (shared OTUs only)
- Larger nodes = more connections
- Each line indicates that an OTU was found in a given mammal
- Diamonds contain OTUs unique to individual samples (bigger = more OTUs)

Ley et al. 2008 Science 320:1647

Herbivore  
Carnivore  
Omnivore
...with the carnivores clustering together...
...two types of herbivores, foregut and hindgut fermenters...
...and one group of ominovores...
...that includes humans despite our diverse habitat and diet
A human gut microbial gene catalogue established by metagenomic sequencing

Cohabiting family members share microbiota with one another and with their dogs

Human gut microbiome viewed across age and geography

Figure 4 | Bacterial species abundance differentiates IBD patients and healthy individuals. Principal component analysis with health status as

Figure 1 | Differences in the fecal microbial communities of Malawians, Amerindians and US children and adults. a, UniFrac distances between
Sequence-based Metagenomics: Shotgun environmental sequencing to characterize genetic diversity: 4+ Examples
Shotgun whole metagenome sequencing to estimate community functions by database comparison

Sample 1
Extract Total Metagenomic DNA From Sample 1
Shear and Select for 400-500 bp fragments
Barcode Selected Fragments
Pool Selected Fragments

Sample 2

Sample n

Mapped reads per gene

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene A</td>
<td>546</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gene B</td>
<td>205</td>
<td>48</td>
<td>700</td>
</tr>
<tr>
<td>Gene C</td>
<td>0</td>
<td>988</td>
<td>0</td>
</tr>
</tbody>
</table>
Example 1: Massive Metagenomic Diversity

Environmental Genome Shotgun Sequencing of the Sargasso Sea

- 1.045 Gb sequenced
- 1.2 million new genes found
- 148 new 16S rRNA sequences
- Sampled ~1800 different species

The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families

- 6.12 million predicted proteins
- **Doubled** total number of known proteins
- Covered ~all prokaryotic protein families
- 1,700 clusters with no homology to known families
- 6,000 previous ORFans have homologs in GOS data
Example 2: Reduced Metagenomic Diversity

Community structure and metabolism through reconstruction of microbial genomes from the environment

Gene W. Tyson¹, Jarrod Chapman¹, Phillip Hugenholtz¹, Eric E. Allen¹, Rachna J. Ram¹, Paul M. Richardson¹, Victor V. Solovyov¹, Edward M. Rubin¹, Daniel S. Rokhsar¹,² & Jillian F. Banfield¹,²


- Acid-mine drainage biofilm
- 76.2 Mb shotgun sequence
- Bimodal GC distribution
- 16S rRNA: 3 bacterial and 3 archaeal lineages
- Assembled genomes for 5 lineages
- Some lineages are clonal
- Some lineages show extensive recombination
- Harsh environment (acid mine) reduces diversity
- Insights into survival strategies in an extreme environment
Example 3: Single cell sequencing of uncultured organisms

Genome sequencing on single *Prochlorococcus* cells

<table>
<thead>
<tr>
<th></th>
<th>JGI</th>
<th>Agencourt</th>
</tr>
</thead>
<tbody>
<tr>
<td># good seq reads</td>
<td>7,166</td>
<td>10,660</td>
</tr>
<tr>
<td>Average length (bp)</td>
<td>769.4</td>
<td>676.6</td>
</tr>
<tr>
<td>Total length (bp)</td>
<td>5,513,520</td>
<td>7,212,556</td>
</tr>
<tr>
<td># unknown seqs</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td># vectors</td>
<td>23</td>
<td>44</td>
</tr>
<tr>
<td># other seqs</td>
<td>74 (MED4)</td>
<td>2 (MED4)</td>
</tr>
<tr>
<td>% genome sampled</td>
<td>62.47%</td>
<td>66.53%</td>
</tr>
</tbody>
</table>

- “Unsampled” genomic regions were detected by PCR amplification from plones
- Primers can be designed from sequenced plones and metagenomic data

Insights into the phylogeny and coding potential of microbial dark matter

Christian Rinke¹, Patrick Schwientek¹, Alexander Sczyrba¹, Natalia N. Ivanova¹, Iain J. Anderson¹,³, Jan–Fang Cheng¹, Aaron Darling¹,⁴, Stephanie Malfatti¹, Brandon K. Swan⁵, Esther A. Gies⁶, Jeremy A. Dodsworth⁷, Brian P. Hedlund⁷, George Tsiamis⁸, Stefan M. Sievert⁹, Wen–Tso Liu¹⁰, Jonathan A. Eisen³, Steven J. Hallam⁶, Nikos C. Kyprides¹, Ramunus Stepanauskas⁹, Edward M. Rubin¹, Philip Hugenholtz¹¹ & Tanja Woyke¹

apply single-cell genomics to target and sequence 201 uncultivated archaeal and bacterial cells from nine diverse habitats belonging to 29 major mostly uncharted branches of the tree of life, so-called 'microbial dark matter'. With this
Single-cell viral tagging – separating viromes and identifying phage-bacteria interactions in single cells

Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw

Rachel Mackelprang³,⁴, Mark P. Waldrop³, Kristen M. DeAngelis⁴, Maude M. David⁴, Krystle L. Chavarria⁴, Steven J. B Edward M. Rubin⁵,⁶ & Janet K. Jansson⁵,⁶


Permafrost contains an estimated 1672 Pg carbon (C), an amount roughly equivalent to the total currently contained within land plants and the atmosphere¹–³

Sequencing yielded a total of 176 million reads and 39.8 Gb of raw sequence.

transition from a frozen to a thawed state there are rapid shifts in many microbial, phylogenetic and functional gene abundances and

We also construct the first draft genome from a complex soil metagenome, which corresponds to a novel methanogen.

Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen

Matthias Hess,¹,², Alexander Sczyrba,¹,²,⁴ Rob Egan,¹,⁵,⁶ Tae-Wan Kim,⁷ Harshal Chokhawala,³ Gary Schrotth,⁴ Shujun Luo,⁴ Douglas S. Clark,¹,²,⁵ Feng Chen,¹,² Tao Zhang,¹,² Roderick I. Mackie,⁶ Len A. Pennacchio,¹,²,⁵ Susannah G. Tringe,¹,²,⁵ Axel Visel,¹,² Tanja Woyke,¹,² Zhong Wang,¹,² Edward M. Rubin¹,²,†

Science Vol 331 28 January 2011

A

pre-incubation

B

d post-incubation

From these data, we identified 27,755 putative carbohydrate-active genes and expressed 90 candidate proteins, of which 57% were enzymatically active

We also assembled 15 uncultured microbial genomes,

Example 4+: Cheaper sequencing – more info and more complete genomes from metagenomes
Questions??

Email: mbaldridge@wustl.edu, or ask your excellent TAs!