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 many 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
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.
Highly connected network of microbial communities across habitats

adapted from:
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
Comparative analysis of genome sequences from multiple strains of the same species allows for the creation of a tree showing lineage and relationships among the strains (e.g., strains of Yersinia pestis in the United States were probably introduced in San Francisco in 1899).

The determination of Bacillus anthracis genome sequences enables the development of tests specific to the Ames strains found in the 2001 mailing of anthrax-containing letters and greatly aids in the federal investigation.

Hospitals are now using real-time PCR assays to screen patients for colonization by Staphylococcus aureus, including methicillin-resistant strains.

Purified protein can be used to determine its 3D structure, and from the structure drugs that might inhibit the activity of the protein can be designed. Libraries of possible drug compounds can be screened to see whether they inhibit or bind the protein directly (e.g., new drugs for treating schistosomiasis).
METAGENOMICS: study microbial communities through direct DNA sequencing

Culture-Dependent Methods

Susceptibility Testing

Genome Sequencing

PCR-based Screens

Culture-Independent Methods

Direct DNA Sequencing
The Great Plate-Count Anomaly (Staley and Konopka, 1985)
Molecular Phylogeny: 16S/18S SSU rRNA

Tree of Life

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 31st, 2023 (GOLD, JGI):

(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

- **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
Sequencing: Relatively cheap nowadays

Required: MASSIVE COMPUTING RESOURCES
Sequence-based Metagenomics

16S rRNA gene-based surveys

Sample

Extract DNA

PCR Amplify (sample specific barcodes)

Sequence

Shotgun sequencing of metagenome/transcriptome

Total DNA/cDNA

Sequence

Who’s there?

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

Phylogenetic 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
Why use the gene for 16S rRNA?

- V1-V2 and V3-V5 are used frequently
- V4 region (~350bp product) perhaps most commonly used
- Full-length sequencing now also being applied for improved species-level resolution
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
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”
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
16S analysis pipelines

GS FLX → PGM → MiSeq

Raw sequences → Quality control → Trimmed and filtered reads → Combined fna files

Performed for each platform separately

Initial Processing

QIIME

QIIME1 → QIIME2 → QIIME3 → QIIME4

De novo OTU picking → Open reference OTU picking → Chimera detection → Taxonomy assignment

Chimera detection → Taxonomy assignment → Alpha and beta diversity analyses

UPARSE

UPARSE1 → UPARSE2

OTU picking → Chimera detection → Taxonomy assignment → Alpha and beta diversity analyses

DADA2 (Illumina only)

Error

Suppression/Dereplication → Taxonomy assignment → Alpha and beta diversity analyses

Allali, et al., BMC Microbiology. 2017
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

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

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

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.

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.

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
\( n_i \) 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
A human gut microbial gene catalogue established by metagenomic sequencing

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
Adult mice given 2 weeks of ampicillin, vancomycin, neomycin, metronidazole, and amphotericin B (AIMD-treated group) or water (vehicle-treated group)

AIMD depletes the gut microbiome. a Stool cultures from AIMD mice yielded fewer colonies, n = 5/group. For box plot, center is mean, box is 25th to 75th percentile, and whiskers are 5th to 95th percentile. Kruskal–Wallis test, *p < 0.05, **p < 0.01. b, c 16S results show that AIMD mice had a decrease in OTUs detected from the Firmicutes and Bacteroidetes phyla, and an increase in Proteobacteria phyla, n = 6–8/group, mean percent abundance (with SEM). These differences were significant as assessed by ANCOM. d 16S results showing the shifts in microbiome in each treatment condition. e Principal coordinate analysis of the gut microbiome, n = 6–8/group. Fecal specimens were collected approximately 2 weeks after intervention.

Zarrinpar, et al., Nat Comm. 2018
Supplementary Figure 1: AIMD changes the composition of the gut microbiome

(A) Sample stool cultures from vehicle-treated and AIMD mice.
(B) DNA extracted from stool of AIMD mice was significantly less than vehicle-treated mice (n = 6).
(C) Cecum from a vehicle-treated and AIMD mouse. White bar is 1 cm.
(D) Cecal weight, n = 9-10; and (E) Stool weight (SEM), n=12; between AIMD and vehicle-treated mice.
(F) 16S composition results from individual mice.
(G) Number of shared OTUs between conditions. Rarefaction plot (H) and PD-whole tree plot
(I) show that α-diversity (SEM) was decreased in AIMD mice.
(J) Unifrac distances showing β-diversity differences. The pre-treatment and vehicle treatment microbiomes were similar to each other based on UniFrac distance, whereas the AIMD microbiome had high UniFrac distances both within its own group and with the other conditions.
(K) PICRUSt results of lipid metabolism pathways in the gut microbiome of AIMD and Vehicle mice.
For boxplots, center is mean, box is 25th to 75th percentile, whiskers are 5th to 95th percentile. Mann-Whitney U test, except for (I) which was a Kruskal-Wallis test; * p<0.05, ** p<0.01, *** p<0.001.

Zarrinpar, et al., Nat Comm. 2018
Sequence-based Metagenomics: Shotgun environmental sequencing to characterize genetic diversity
Shotgun whole metagenome sequencing to estimate community functions by database comparison.

- Extract Total Metagenomic DNA From Sample 1
- Shear and Select for 400-500 bp fragments
- Barcode Selected Fragments
- Pool Selected Fragments
- Map Reads to Database
- Reads Binned by Barcode
- Sequencing

<table>
<thead>
<tr>
<th>Sample</th>
<th>gene A</th>
<th>gene B</th>
<th>gene C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>546</td>
<td>205</td>
<td>0</td>
</tr>
<tr>
<td>Sample 2</td>
<td>0</td>
<td>48</td>
<td>988</td>
</tr>
<tr>
<td>Sample n</td>
<td>0</td>
<td>700</td>
<td>0</td>
</tr>
</tbody>
</table>
NGS analysis of microbiome samples

Quince, et al., Nat Biotech. 2017
Assembly- and read-based profiling

Quince, et al., Nat Biotech. 2017
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


- 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 metagenomics

Xu, et al., Protein Cell. 2018
Example 4: Single-cell viral tagging – separating viromes and identifying phage-bacteria interactions in single cells

Modern/molecular Koch’s postulates

- (i) A nucleic acid sequence belonging to a putative pathogen should be present in most cases of an infectious disease.
- (ii) Fewer, or no, copy numbers of pathogen-associated nucleic acid sequences should occur in hosts or tissues without disease.
- (iii) With resolution of disease, the copy number of pathogen-associated nucleic acid sequences should decrease or become undetectable. With clinical relapse, the opposite should occur.
- (iv) When sequence detection predates disease, or sequence copy number correlates with severity of disease or pathology, the sequence-disease association is more likely to be a causal relationship.
- (v) The nature of the microorganism inferred from the available sequence should be consistent with the known biological characteristics of that group of organisms.
- (vi) Tissue-sequence correlates should be sought at the cellular level: efforts should be made to demonstrate specific in situ hybridization of microbial sequence to areas of tissue pathology.
- (vii) These sequence-based forms of evidence for microbial causation should be reproducible.

Fredricks & Relman, Clin Micro Rev. 1996
Questions??

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