Introduction to microbiota data analysis

Natalie Knox, PhD
Head Bacterial Genomics, Bioinformatics Core
National Microbiology Laboratory, Public Health Agency of Canada
3 Who I am

“Jack of all trades, master of none”

4 The NML bioinformatics’ toolbox
Microbial ecology is the study of microbes in the environment and their interactions with each other. Microbes are the tiniest creatures on Earth, yet despite their small size, they have a huge impact on us and on our environment.
Microbial ecology - the big 3

1. Who is there?
2. How many are there?
3. What do they do?

What are microbes?

- Bacteria
- Archaea
- Virus
- Parasite
  - Helminths (parasitic worms)
- Fungi
  - yeast, mold, mushrooms
- Protozoa
  - Ciliates
  - Amoebae
  - flagellates
Tree of life: Six Kingdom Classification

- **Eukaryotes**: unicellular/multicellular - Presence of nucleus and mitochondria
- **Prokaryotes**: unicellular - Absence of nucleus and mitochondria

Taxonomic Ranking System

- **Life**
- **Domain**
- **Kingdom**
- **Phylum**
- **Class**
- **Order**
- **Family**
- **Genus**
- **Species**

- **Bacteria**
- **Eubacteria**
- **Proteobacteria**
- **Gammaproteobacteria**
- **Enterobacteriales**
- **Enterobacteriaceae**
- **Escherichia**
- **E. coli**

- **Life**
- **Bacteria**
- **Animalia**
- **Chordata**
- **Mammalia**
- **Primates**
- **Hominidae**
- **Homo**
- **H. sapiens**
Scientific Nomenclature for Bacteria

binomial nomenclature

- **Italicization**
  - Family, genus, species, subspecies
  - Not strain or serovar
- **Capitalization**
  - Genus and above
  - Serovars
- **Abbreviations**
  - Genus after initial introduction (e.g. *E. coli*)
  - “sp.” & “spp.” (e.g. *Salmonella spp.)*
  - “gen. nov.” (*genus novum*) or “sp. nov.” (*species nova*)
- **Plurals vs singular**
  - Genus / genera
  - Species / species
  - Phylum / phyla

“**The Father of Taxonomy**”
Carl Linnaeus

Introduction to Microbiota Studies
Let’s start with some terminology

Microbiota
▷ Represents the assemblage of microorganisms present in a given environment

Microbiome
▷ Includes the microorganisms and their entire genetic content in a given environment

Metagenome
▷ Gene and genomes of a microbiota

DNA-based microbiota survey methods

Targeted amplicon sequencing

Shot-gun metagenomics
DNA-based microbiota survey methods

Targeted amplicon
- Requires taxonomically informative marker
- Ubiquitous to target group
  - Bacteria and archaea: 16S rRNA, cpn60
  - Fungi: ITS, 18S rRNA
- Easily amplified
- Highly curated and comprehensive reference database
- Amplification and primer bias/error
- Ideally single copy
- Low selective pressure

Shot-gun metagenomics
- Unrestricted sequencing of all DNA present in a sample
  - Eukaryotic, prokaryotic, virus
- Sampling all genomic content
  - Sequencing depth
  - Sample matrix
  - Low or high biomass sample
- Functional profiling
- Assemblies challenging

What approach is best?
## DNA-based microbiota study categories

<table>
<thead>
<tr>
<th></th>
<th>Targeted amplicon</th>
<th>Shotgun metagenomics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advantage</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microbial target</td>
<td>Narrow scope</td>
<td>Wide scope (unbiased*)</td>
</tr>
<tr>
<td>Cost</td>
<td>Relatively low</td>
<td>Can be costly</td>
</tr>
<tr>
<td>Computational</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>requirements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data analysis</td>
<td>Many workflows &amp; software available</td>
<td>Complex &amp; (few) research grade software</td>
</tr>
<tr>
<td></td>
<td>Relies on curated DB</td>
<td>Relies on curated DB</td>
</tr>
</tbody>
</table>

*a priori knowledge required

* unbiased sequencing (wet-lab requires targeted approaches)

---

### Bioinformatics rule of thumb

Using the right tool for the job

- **“Shot-gun metagenomics”**
- **“Targeted amplicon sequencing”**
Consider the following:

▶ What are your research questions
▶ How much money and time do you want to invest?
▶ Both methods require moderate to expert knowledge
▶ High performance computing environment
▶ Most open-source software available
  ▶ Not always microbiologist friendly
  ▶ Often based on the command-line
Targeted amplicon sequencing

The famous 16S rRNA gene

The Bacterial Ribosome
Ribosome fun facts

- Prokaryotes and Eukaryotes have ribosomes
- The ribosome is composed of 2 major components:
  - Small subunit (SSU): reads the RNA
  - Large subunit (LSU): join amino acids to form a polypeptide chain
- Unit of measurement is Svedberg unit
  - Measure of the rate of sedimentation in centrifugation
  - Prokaryotes have a 70S ribosome
    - SSU = 30S (16S rRNA)
    - LSU = 50S (5S and 23S rRNA)
  - Eukaryotes have an 80S ribosome
    - SSU = 40S (18S rRNA)
    - LSU = 60S (5S, 5.8S, and 28S rRNA)

Let’s talk about 16S rRNA

- Bacteria and archaea specific (mitochondria, chloroplasts)
- Multiple copy numbers (anywhere from x to z)
  - copies of the ribosome gene operon dispersed throughout the genome
    - e.g. *E. coli* has ~ 7 copies of same operon (rrnA, rrnB, etc.)
  - Eukaryotic ribosome gene operons typically occur in tandem arrays
Getting to know the 16S rRNA gene in more depth

- Part of the 30S SSU of prokaryotic ribosome that binds to the RBS (Shine-Dalgarno sequence/Ribosome Binding Site)
  - Length = 1540 nucleotides bound to 21 proteins
  - 3’ end codes for the anti-RBS
  - Helps bind SSU and LSU together (interacts with 23S in LSU)

Getting to know the 16S rRNA gene intimately

- Most widely used taxonomic marker
- Conserved secondary structure acts as an anchor

**CONSERVED REGIONS:** unspecific applications

**VARIABLE REGIONS:** group or species-specific applications
Why is the 16S rRNA gene ideal to characterize microbiota in a given environment?

- Universally present in bacteria
- Low selective pressure
- Easily amplified
- Conserved with variable regions
- Adequate discriminatory power to genus level
- Variable regions can be adequately sequenced with most next generation sequencing technologies

Other taxon markers

- ITS: Internal transcribed spacers (ITS1 and ITS2) for Fungi
  - Variable lengths: ~ 360/232 bp each (600-700 bp)
- cpn60: chaperonin 60 (cpn60 group I) ~550 bp
- rpoB: beta subunit of DNA polymerase ~370 bp
- 23S rRNA ~ 2300 bp
Targeted amplicon sequencing

Overall workflow of targeted-amplicon sequencing approach:
- gDNA extraction
- 16s rRNA variable region amplification
- High-throughput sequencing
- Data analysis: QC and filtering building OTUs
- Downstream analytics & statistics
16S rRNA sequencing methodology

Illumina sequencing methodology
Illumina sequencing methodology

Illumina clusters and base calling

Base calling from fluorescent image
* This is a fastq file. Remember, you will have 2 fastq files for each sample sequenced (forward and reverse).

### The basics of a fastq file

```
@FORJUSP02A3W1
CGTCAATTCATTTAAGTTTAACCTTGGGGCGTACTCCAGGCCGTT
+
AAAAAAAAAAAAAAA:???:??:??:??????:????:AAAAACCAA:.:B?:?::?
```

**Q scores (as ASCII char)**

Base=T, Q=\*1=25
Phred scores

A quality score is a prediction of the probability of an error in sequencing base calling.
Wet-lab and sequencing aspects to consider while planning your project...

- Sequencing platform
- Chemistry
- Read length
- Primers
- Sample matrix and expected microbial community
- DNA extraction protocol

Sequencing concerns

- Low base diversity libraries
  - PhiX incorporation (~10-50%)
- Cluster density
- Amplicon length
  - Sequencing overlap
Data analysis

General overview and software

Data analysis software

- Pat Schloss (University of Michigan)
  - Open-source
  - written in C and C++
  - Few dependencies / easy to install
  - platform independent
    - Windows-compatible executable
    - source code:
    - Unix/Linux or Mac OS X environments

- Rob Knight (University of California San Diego)
  - Open-source
  - Backbone written in python
  - Available as a virtual machine or in packaged container (miniconda)
  - Many dependencies / install challenging
  - Linux or MAC supported
  - Container for different analysis software
  - More visualizations
  - Written in python
  - Open-source
Data analysis workflow

1. Build contigs
2. Trim and filter sequences
3. Align sequences
4. Build OTUs
5. Taxonomic classification
6. Downstream analysis

The OTU
Operational Taxonomic Unit

A bin containing sequences of X % similarity - a sorting process

OTU1  OTU2  OTU3  OTU4
Major steps to building an OTU

1. Alignment
2. Distance Matrix
3. Clustering
Data analysis

Detailed workflow

16S rRNA data analysis token - OTUs

- Goal: High quality and stable OTUs
  - Are they the same?
- Considerations for creating OTUs
  - clustering method
  - Aligner
  - database
Building contigs: sample output

This is a fasta file

```
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
MB1049 261 00000000 ATGCA 1 1101 1818
```

Trim and filter sequences

- Pre-processing step
  - trim off primer sequences and barcodes
  - cull sequences based on:
    - quality scores
    - sequence length (min. and max.)
    - presence of ambiguous bases and homopolymers (sequencer dependent)
- Create non-redundant set of sequences
  - redundant sequences are indexed for use in downstream analysis
Removing sequencing noise and chimeras

▷ Pre-clustering sequences (Huse et al. 2010)
  ▷ assumes “rare” sequences which are highly similar to abundant sequences are an artefact of pyrosequencing error
  ▷ default: difference=1 base mismatch
    
    \[
    AGTCCTG == AGGCCCTG
    \]

▷ Identifying and removing chimeras
  ▷ chimera slayer (Broad Institute) and Uchime (Rob Knight’s Lab)
  ▷ cpu intensive

Pitfalls of not removing PCR and sequencing artefacts

▷ Inaccurate results
  ▷ Falsely inflated diversity
  ▷ Mis-classified OTUs or presence of “novel” OTUs
▷ Excessively large distance matrix
  ▷ High memory usage and CPU time
Removing other non-prokaryotic “junk”

▸ Other sources of contaminating DNA
  ▪ Mitochondrial
  ▪ Chloroplastic rRNA
The “kitome”: the reagent microbiome

- Lab reagent microbiome
  - e.g. DNA extraction reagents
- DNA is everywhere!
  - Autoclaving ≠ DNA-free
  - Bleach and filter tips (a must!)
  - Separate your DNA extraction and PCR setup stations
  - Sample-to-sample contamination
  - Some reagents produced by bacteria
- Sequence a negative control
  - Even if no band present on gel!

Some considerations...

1. Sequence positive and negative control at every step
   a. Mock community - sequencing error rate
   b. Abundant donor - sanity check
   c. Negative control (extraction blank) - contaminant check
2. Low biomass samples more susceptible
   a. Aim for starting sample $>10^3$-$10^4$ cells
3. Careful sample collection
4. Random order processing (different kits for replicates)
5. Documentation is key (eg. lot numbers)
6. Critical evaluation of results
Mock community

- Known community structure
- No PCR inhibitors
- Used to calculate sequencing error rate
- Create your own or use HMP’s

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

<table>
<thead>
<tr>
<th>Phylum</th>
<th>List of constituent contaminant genera</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alpha-proteobacteria:</strong></td>
<td>Alpha eikenella, Rhodobacter, Enterobacter, Escherichia, Pelobacter, Flavobacterium</td>
</tr>
<tr>
<td><strong>Beta-proteobacteria:</strong></td>
<td>Acinetobacter, Brevundimonas, Pseudomonas, Rhodococcus, Nocardioides</td>
</tr>
<tr>
<td><strong>Gamma-proteobacteria:</strong></td>
<td>Enterobacter, Citrobacter, Providencia, Serratia, Salmonella</td>
</tr>
<tr>
<td><strong>Proteobacteria:</strong></td>
<td>Shewanella, Pseudomonas, Vibrio, Flavobacterium, Necrophorus</td>
</tr>
<tr>
<td><strong>Actinobacteria:</strong></td>
<td>Rothia, Corynebacterium, Mycobacterium, Actinomyces, Streptomyces</td>
</tr>
<tr>
<td><strong>Firmicutes:</strong></td>
<td>Staphylococcus, Clostridium, Bacillus, Enterococcus, Lactococcus</td>
</tr>
<tr>
<td><strong>Bacteroidetes:</strong></td>
<td>Prevotella, Bacteroides, Parabacteroides, Ruminococcus</td>
</tr>
<tr>
<td><strong>Deinococcus-Thermus:</strong></td>
<td>Deinococcus</td>
</tr>
<tr>
<td><strong>Acidobacteria:</strong></td>
<td>Predominantly unclassified Acidobacteria, Gp2 organisms</td>
</tr>
</tbody>
</table>
Phylotype vs. OTU-based approach

Phylotype
- No MSA
- Sequences assigned to bins based on similarity to database
- inaccurate

OTU-based approach
- Based on MSA
  - Secondary structure should be considered to conserve positional homology
  - More CPU intensive
    - E.g. “typical user” could not align more than ~10,000 full-length sequences - lack of RAM

Phylotype-based analysis

**Phylotype-based analysis:**
- Sequences assigned to bin based on similarity to database sequence
- *Limitations*
  - Marginally similar sequences may be assigned the same taxonomy
  - Novel sequences may be incorrectly classified
  - Are taxonomy dependent – classification to species/genus level inaccurate
MSA: aligners

- Why can’t we use generic aligners?
  - Don’t scale well
  - Data stored in RAM
  - They don’t consider secondary structure of the 16S rRNA

- Why should we care about secondary structure?
  - Conserves positional homology
  - Increases robustness of alignment (less sensitive to user supplied parameters)

MSA: alignment strategy for 16S rRNA

Not using a typical *de novo* MSA approach
MSA: aligners for 16S rRNA analysis

▸ What are these “special” aligners?
  ▸ Profile-based aligners
    ▸ scale linearly in time and smaller memory footprint

▸ Considerations for choosing a 16S rRNA profile-based aligner
  ▸ Alignment quality, scalability for large datasets, speed, database, open-source

---

MSA: 16S rRNA aligners and databases

RDP (ribosomal database project)
  ▸ aligner uses secondary structure models (HMM)
  ▸ BUT...variable regions poorly aligned (limited number of models)
  ▸ Alignment length not fixed

SILVA database project (spin-off of ARB)
  ▸ uses SINA aligner (SILVA Incremental Aligner) – k-mer search/partial order alignment (POA)
  ▸ their alignment is 50,000 columns long (archaea and euk. 18S)
  ▸ Better alignment of variable regions

Greengenes
  ▸ aligner puts less emphasis on secondary structure but their database alignment does
  ▸ uses kmer search (k=7mers) to id closest match in ref DB followed by blastn for pairwise alignment
  ▸ use NAST (nearest alignment space termination) to re-insert gaps in candidate sequence – positional homology retained
    ▸ allows for fixed column count in MSA
  ▸ reference alignment 7,682 columns (much shorter than SILVA)
MSA: comparison of alignments to different databases

RDP
E. coli (U00906)  A070000AAGAAAGGTTGAGGGTTTGTGAGG
E. casso.Reus (Y18161)  A070000AAGAAAGGTTGAGGGTTTGTGAGG
E. termito (AM036966)  A070000AAGAAAGGTTGAGGGTTTGTGAGG
Uncotted (AY830399)  A070000AAGAAAGGTTGAGGGTTTGTGAGG
Uncotted (EF510700)

greengenes
E. coli (U00906)  A070000AAGAAAGGTTGAGGGTTTGTGAGG
E. casso.Reus (Y18161)  A070000AAGAAAGGTTGAGGGTTTGTGAGG
E. termito (AM036966)  A070000AAGAAAGGTTGAGGGTTTGTGAGG
Uncotted (AY830399)  A070000AAGAAAGGTTGAGGGTTTGTGAGG
Uncotted (EF510700)

SILVA
E. coli (U00906)  A070000AAGAAAGGTTGAGGGTTTGTGAGG
E. casso.Reus (Y18161)  A070000AAGAAAGGTTGAGGGTTTGTGAGG
E. termito (AM036966)  A070000AAGAAAGGTTGAGGGTTTGTGAGG
Uncotted (AY830399)  A070000AAGAAAGGTTGAGGGTTTGTGAGG
Uncotted (EF510700)

MSA: more on sequence alignments

Bottle neck of the analysis & many options!

1. **Find closest template for each candidate using:**
   - kmer searching
   - blastn (alignment dependent)
   - suffix tree searching
2. **Make pairwise alignment between candidate and de-gapped template sequences using one of the following algorithms:**
   - Needleman-Wunsch (fast; global alignment)
   - Gotoh (slowest; global alignment)
   - blastn (Smith-Waterman; local alignment)
3. **Re-insert gaps using NAST algorithm**
   - candidate sequence alignment and original template alignment same length
MSA and databases: Key message

Quality of candidate sequence alignment is dependent on reference alignment

Clustering

- **Build distance matrix**
  - calculate pairwise distances
- **Cluster**
  - cluster sequences into OTUs
  - Three clustering methods:
    - Nearest neighbour:
      Each of the sequences within an OTU are at most X% distant from its most similar sequence in the OTU.
    - Furthest neighbour
      All of the sequences within an OTU are at most X% distant from all of the other sequences within the OTU.
    - Average neighbour
      happy medium between NN and FN
The distance matrix

A massive file!

<table>
<thead>
<tr>
<th>OTU</th>
<th>Size</th>
<th>Taxonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

OTU classification sample output
### Classification and abundance count per sample

<table>
<thead>
<tr>
<th>taxonomy</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acidobacteria</td>
<td>100</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>10</td>
</tr>
<tr>
<td>Armamotoanaetes</td>
<td>1</td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>5</td>
</tr>
<tr>
<td>Deinococcus-Thermus</td>
<td>1</td>
</tr>
<tr>
<td>Fusobacteria</td>
<td>1</td>
</tr>
<tr>
<td>Lentisphaere</td>
<td>1</td>
</tr>
<tr>
<td>Planctomycetes</td>
<td>1</td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>6</td>
</tr>
<tr>
<td>Sphingobacteria</td>
<td>1</td>
</tr>
<tr>
<td>Synergistates</td>
<td>5</td>
</tr>
<tr>
<td>Tenericutes</td>
<td>1</td>
</tr>
<tr>
<td>Verrucomicrobia</td>
<td>1</td>
</tr>
<tr>
<td>Bacteria_unclassified</td>
<td>1</td>
</tr>
</tbody>
</table>

### Evaluation of output

<table>
<thead>
<tr>
<th>taxonomy</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acidobacteria</td>
<td>100</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>10</td>
</tr>
<tr>
<td>Armamotoanaetes</td>
<td>1</td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>5</td>
</tr>
<tr>
<td>Deinococcus-Thermus</td>
<td>1</td>
</tr>
<tr>
<td>Fusobacteria</td>
<td>1</td>
</tr>
<tr>
<td>Lentisphaere</td>
<td>1</td>
</tr>
<tr>
<td>Planctomycetes</td>
<td>1</td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>6</td>
</tr>
<tr>
<td>Sphingobacteria</td>
<td>1</td>
</tr>
<tr>
<td>Synergistates</td>
<td>5</td>
</tr>
<tr>
<td>Tenericutes</td>
<td>1</td>
</tr>
<tr>
<td>Verrucomicrobia</td>
<td>1</td>
</tr>
<tr>
<td>Bacteria_unclassified</td>
<td>1</td>
</tr>
</tbody>
</table>

---

**Note:**
- The data in the tables represent the classification and abundance count per sample.
- The evaluation of output shows the same distribution across different phyla and classes.
Understanding diversity indices

Alpha diversity
- how many taxa are in a sample
- rarefaction curves
- Richness (i.e. Chao1 and ACE) and diversity (Shannon and Simpson) indices

Beta diversity
- how many taxa are shared between samples
- PCoA

Characterizing the microbiota

- Classify sequences
- Measure alpha and beta-diversity
  - Alpha diversity:
    - how many taxa are in a sample
    - rarefaction curves
    - Richness (i.e. Chao1 and ACE) and diversity (Shannon and Simpson) indices
  - Beta diversity:
    - how many taxa are shared between samples
    - PCoA
- heat maps
- Unifrac
- venn diagrams
- phylogenetic trees
- And much more...
Data exploration

Downstream analysis

- Sample diversity
- Similarity of samples, associations with metadata
  - Procrustes analysis
  - PCoA
  - ANOSIM
  - Metastats
  - LEfSe
- Functional prediction
  - PICRUSt
Final thoughts

▷ No standardized approach!
▷ Garbage in = garbage out
  ▷ Data QC, filtering, and de-noising critical
▷ Increasing sequencing throughput will inevitably increase cpu performance requirement
▷ Does it make sense biologically?

Any questions?
You can find me at:
▷ natalie.knox@canada.ca

Special credits to:
▷ Gary Van Domselaar
▷ Morag Graham
▷ Jessica Forbes

THANKS!