Introduction to Bioinformatics on Unity III

Cecile Cres & Anna Schrecengost

May 6, 2024
Outline

- Microbiome research
- Amplicon sequencing overview
- Example of amplicon sequencing data analysis pipeline
  - Downloading Data
  - Data pre-processing (primer trimming)
  - Denoising
  - Taxonomic assignment
  - Phylogenetic placement & taxonomic assignment
A microbiome refers to the collection of genomes from all the microorganisms in a particular habitat as well as the structural elements, metabolites and environmental conditions.

The microbiota describes the microorganisms.

Example of microbiomes:
- Human microbiome
- Ocean microbiome

What are the microbial species and what are their function?

Microbiome research

Microbiome

Microbiota

- Bacteria
- Archaea
- Fungi
- Protists
- Algae

“Theatre of activity”

Microbial structural elements

- Proteins/peptides
- Lipids
- Polysaccharides
- Nucleic acids structural DNA/RNA
- Mobile genetic elements incl. viruses/phages relic DNA

Environmental conditions

Microbial metabolites

- Signalling molecules
- Toxins
- (An)organic molecules

Biome: a reasonably well defined habitat which has distinct bio-physio-chemical properties


Prokaryotes, eukaryotic microbes and viruses
- Biogeochemical cycling (CO2 capture, O2 generation and carbon removal)

Marine water, sediments, coral reefs, hydrothermal vents

Goal: improve our understanding of microorganisms and their roles in the ocean

Methods used in microbiome research

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culturome</td>
<td>• High-throughput</td>
<td>• Expensive</td>
</tr>
<tr>
<td></td>
<td>• Targeted selection</td>
<td>• Laborious</td>
</tr>
<tr>
<td></td>
<td>• Provides microbial isolates</td>
<td>• Influenced by media and the environment</td>
</tr>
<tr>
<td>Amplicon</td>
<td>• Quick analysis</td>
<td>• PCR and primer biases</td>
</tr>
<tr>
<td>(16S/18S/ITS)</td>
<td>• Low-biomass requirement</td>
<td>• Resolution limited to genus level</td>
</tr>
<tr>
<td></td>
<td>• Applicable to samples contaminated by host DNA</td>
<td>• False positive in low-biomass samples</td>
</tr>
<tr>
<td>Metagenome</td>
<td>• Taxonomic resolution to species or strain level</td>
<td>• Expensive</td>
</tr>
<tr>
<td></td>
<td>• Functional potential</td>
<td>• Time-consuming in analysis</td>
</tr>
<tr>
<td></td>
<td>• Uncultured microbial genome</td>
<td>• Host-derived contamination</td>
</tr>
<tr>
<td>Virome</td>
<td>• Can identify RNA and DNA viruses</td>
<td>• Most expensive</td>
</tr>
<tr>
<td></td>
<td>• Quick diagnosis</td>
<td>• Difficult to analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Severe host-derived contamination</td>
</tr>
<tr>
<td>Metatranscriptome</td>
<td>• Can identify live microbes</td>
<td>• Complex sample collection and analysis</td>
</tr>
<tr>
<td></td>
<td>• Can evaluate microbial activity</td>
<td>• Expensive and complex in sequencing</td>
</tr>
<tr>
<td></td>
<td>• Transcript-level responses</td>
<td>• Host mRNA and rRNA contamination</td>
</tr>
</tbody>
</table>

Amplicon sequencing

- Amplicon: the resulting sequence of a targeted amplification of genetic material
- Useful for detection of hotspot mutations, gene fusions and single-nucleotide polymorphisms (SNPs), taxonomic classification of microorganisms
- Targeted (use of primers) sequencing of marker genes
  - 16S ribosomal DNA in prokaryotes
  - 18S ribosomal DNA in eukaryotes

Workflow of amplicon sequencing:
Amplicon sequencing

Library preparation

- Two-step polymerase chain reactions (PCR):
  - First PCR reaction: the targeted DNA region is amplified using specific primers flanked by sequencing primers
  - Second PCR reaction: the sequencing primers allow for a second PCR reaction to add adapter sequences and indexes for sample multiplexing

Demultiplexing: step in high-throughput sequencing data analysis where sequences are sorted based on their sample of origin

Amplicon sequencing data analysis pipeline

1. Raw sequencing data
2. Evaluate data quality
   - Fastqc
3. Import into QIIME
   - QIIME2
4. Primer trimming
   - QIIME2 cutadapt
5. Quality filtering, denoising, merging, chimera removal
   - QIIME2 DADA2
6. Taxonomic assignment
   - QIIME2 classify-sklearn
   - PaPaRa, EPA-ng
7. Phylogenetic placement
8. Taxonomic assignment
   - gappa
SSU-rRNA Gene Sequencing Survey of Benthic Microbial Eukaryotes from Guaymas Basin Hydrothermal Vent

Alexis Pasulka\textsuperscript{a}, Sarah K. Hu\textsuperscript{b,1}, Peter D. Countway\textsuperscript{c}, Kathryn J. Coyne\textsuperscript{d}, Stephen C. Cary\textsuperscript{e}, Karla B. Heidelberg\textsuperscript{b} & David A. Caron\textsuperscript{b}

\textsuperscript{a} Biological Sciences Department, California Polytechnic State University, 1 Grand Avenue, San Luis Obispo, California, USA
\textsuperscript{b} Department of Biological Sciences, University of Southern California, 3616 Trousdale Parkway, AHF 301 Los Angeles, Los Angeles, California, USA
\textsuperscript{c} Bigelow Laboratory for Ocean Sciences, 60 Bigelow Drive, East Boothbay, Maine, USA
\textsuperscript{d} College of Earth, Ocean, and Environment, University of Delaware, 700 Pottstown Road, Lewes, Delaware, USA
\textsuperscript{e} Department of Biological Sciences, The University of Waikato, Private Bag 3105, Hamilton, New Zealand

- 18S rRNA gene high-throughput sequencing of the V4 region (expected amplicon size: \(\sim\)300 bp)
- Raw sequence data (Illumina MiSeq) available in the Sequence Read Archive (SRA) repository (BioProject accession ID: PRJNA391741)
- The following example for amplicon sequencing data analysis is done one one sample
Start interactive session on Unity command line

Requesting: # of cpu cores; amount of memory; time; Unity partition

```
salloc --cpus-per-task 8 --mem=8G --time 1:00:00 --partition cpu,uri-cpu,cpu-preempt
```
Download from NCBI Sequence Read Archive (SRA)

Need to use NCBI’s SRA toolkit to download data from the SRA

```
module load sratoolkit/3.0.7
module load entrezdirect/10.7.20190114
module load parallel/20210922

project='PRJNA391741'
esearch -db sra -query $project | efetch -format runinfo > runinfo.csv
cat runinfo.csv | cut -d "," -f 1 > SRR.numbers
sed -i '1d' SRR.numbers
cat SRR.number | parallel fastq-dump --split-files --origfmt --gzip

mkdir fastq/
cd fastq/
fastq-dump SRR5753741 --split-files --origfmt --gzip
```

- Fetches information about sequencing runs based on accession ID
- Pulls out sample IDs
- Downloads fastq files from all samples in parallel
- Download only one sample based on sample ID on SRA
Visualize quality of reads with fastqc

module load fastqc/0.11.9
module load MultiQC/1.12-foss-2021b

mkdir fastqc/
fastqc fastq/*_1* --outdir fastqc/
fastqc fastq/*_2* --outdir fastqc/

multiqc fastqc/*_1_fastqc.zip --filename forward_multiqc.html --outdir multiqc/
multiqc fastqc/*_2_fastqc.zip --filename reverse_multiqc.html --outdir multiqc/

Summarizes sequence quality for each fastq file (forward and reverse)

If you have many samples, summarizes the fastqc results into one file per read direction
Import into QIIME2

Note: this is an older version of qiime, you should install the latest version in a conda environment

We are importing into QIIME using a specific file name format (Casava), so we need to rename our files first. You can also use a manifest file to import fastq files, instructions here

Import fastq files into QIIME

module load uri/main QIIME2/2021.8

cd fastq/
rename 's/_00_L001_/g' *
rename 's/.fastq.gz/_001.fastq.gz/g' *
rename 's/_1/_R1/g' *
rename 's/_2/_R2/g' *

qiime tools import \n  --type 'SampleData[PairedEndSequencesWithQuality]' \n  --input-path fastq/ \n  --output-path work/demux_PE.qza \n  --input-format CasavaOneEightSingleLanePerSampleDirFmt
Pre-processing - trim primers

- Primers don’t always bind perfectly to the target sequence (sequence not identical to the target DNA sequence).
- Use Cutadapt to remove primers and any preceding bases.

- Primer: short nucleotide sequence complementary to the target sequence
- Index: short nucleotide sequence that serves as a unique identifier associated with a sample
- Adapter: short nucleotide sequence that allows the library to bind to the sequencing flow cell
Pre-processing - trim primers

```
qiime cutadapt trim-paired
   --i-demultiplexed-sequences work/demux_PE.qza
   --p-cores 8
   --p-front-f CCAGCASCYGCGGTAATTCC
   --p-front-r ACTTTCGTTCTTGATYRA
   --p-match-adapter-wildcards
   --p-match-read-wildcards
   --p-minimum-length 10
   --p-discard-untrimmed
   --verbose
   --o-trimmed-sequences work/demux_PE_trimmed.qza

qiime demux summarize
   --i-data work/demux_PE_trimmed.qza
   --o-visualization work/demux_PE_trimmed.qzv
```
Denoising - DADA2

**DADA2**: Filters based on quality score of bases, denoises sequences (models and corrects sequencing errors from Illumina sequencer), merges forward and reverse reads, and then filters out chimeras.

```bash
qiime dada2 denoise-paired
   --i-demultiplexed-seqs work/demux_PE_trimmed.qza
   --p-trunc-len-f 220
   --p-trunc-len-r 210
   --p-n-threads 8
   --verbose
   --o-table work/table.qza
   --o-representative-sequences work/rep-seqs.qza
   --o-denoising-stats work/DADA2-stats.qza

qiime metadata tabulate --m-input-file work/DADA2-stats.qza --o-visualization work/DADA2-stats.qzv
```

Truncate sequences when they start to drop off in quality.
Amplicon denoising

- Sequence quality control step to remove sequence errors from amplicon reads and obtain Amplicon Sequence Variants (ASVs)
- Used to improve taxonomic assignment of amplicon reads
- Use DADA2 to perform denoising
  - DADA2 implements a ‘quality-aware model’ of sequencing errors and corrects the reads by removing noise related to the sequencing methodology

![Diagram showing PCR/sequencing and denoising steps]
Assign taxonomy

Download premade classifier from [QIIME2 website](https://qiime2.org)

Can also create your own (e.g. with another database like [PR2](https://pr2.org) - [instructions to train your own](https://qiime2.org)

```bash
qiime feature-classifier classify-sklearn
  --i-classifier work/silva-138-99-nb-classifier.qza
  --i-reads work/rep-seqs.qza
  --o-classification work/taxonomy.qza
```
Export QIIME artifacts

- Export count table (BIOM format)
  - Convert to readable format (.csv, .tsv) with tool like `biom-format`

- Export representative sequences of ASVs (.fasta file)

- Export taxonomic assignments of ASVs (.tsv file)

qiime tools export
   --input-path work/table.qza
   --output-path work/export/table

qiime tools export
   --input-path work/rep-seqs.qza
   --output-path work/export/rep-seqs

qiime tools export
   --input-path work/taxonomy.qza
   --output-path work/export/taxonomy
Phylogenetic placement

Place query sequences (ASVs) onto a reference phylogenetic tree in order to get a deeper understanding of the phylogenetic composition of your samples

- Capture diversity which is underrepresented in reference databases - do not need exact match, takes evolutionary history into account
- More accurate way to analyze phylogeny of your samples and conduct phylogenetically aware diversity analyses (versus de novo tree-building methods)
- Can use for taxonomic assignment, diversity quantification, sample comparison, correlation with environmental variables
Phylogenetic placement

Need a reference phylogenetic tree to place sequences onto:

1. Reference phylogenetic tree
2. Reference alignment
3. File describing taxonomy of each tip of phylogenetic tree

This reference tree should span the diversity of sequences that will be placed on the tree, and should contain (nearly) full-length, high quality, curated sequences from relevant gene (here, 18S rRNA)

Here, using eukaryotic tree of life from [this publication](#)
Phylogenetic placement - **PaPaRa**

Aligns ASVs to reference sequences so that they can be placed onto reference phylogenetic tree

```
module load papara_nt/2.5

cd work/phylo-placement

papara \
-t euk_tree.tree \ 
-s eukaryotic_reference_tree.phy \ 
-q ../export/rep-seqs/dna-sequences.fasta \ 
-j 8 \ 
-r
```

Reference tree
Reference alignment in phylip format
Query sequences (ASVs to be placed on reference tree)
Number of threads/cpu cores
Phylogenetic placement - **EPA-ng**

**Conda environment where installed**

- `epa-ng`, `raxml-ng`, and `gappa`

**Output from papara**

- `epa-ng --split` and `raxml-ng`: necessary steps to prepare for placement

**Reference tree**

**Reference alignment and ASVs (output from `epa-ng --split`)**

**Output from `raxml-ng`**

---

```
module load anaconda/2022.10
conda activate phylo-placement

epa-ng
  --split eukaryotic_reference.fasta
  papara_alignment.default

raxml-ng
  --evaluate
  --msa reference.fasta
  --tree euk_tree.tree
  --model GTR+G
  --threads 8

epa-ng
  --filter-acc-lwr 0.99
  --filter-max 70
  -t euk_tree.tree
  -s reference.fasta
  -q query.fasta
  --model reference.fasta.raxml.bestModel
```
Phylogenetic placement - **gappa**

Tools to visualize and analyze results from phylogenetic placement

- **examine heat-tree**:
  - `--jplace-path` `epa_result.jplace`
  - `--mass-norm` `absolute`
  - `--write-svg-tree`
  - `--write-newick-tree`
  - `--write-nexus-tree`

- **examine assign**:
  - `--jplace-path` `epa_result.jplace`
  - `--taxon-file` `eukaryotic_reference_tax.txt`
  - `--per-query-results`

- **examine lwr-list**:
  - `--jplace-path` `epa_result.jplace`

Provides you with reference tree annotated with ASV placements. Visualize with iTOL (create an account)

Output from placement

Figures out the best taxonomic assignment based on all the placements

File describing the taxonomy of tips of tree

Summary of likelihood ratios for all placements
Additional Resources

- Unity Onboarding video (Spring 2024)
- QIIME2 snakemake pipeline
- Snakemake workshop
- Unity community Slack
- More contact information
- URI AI Lab Workshop recordings (including this series)