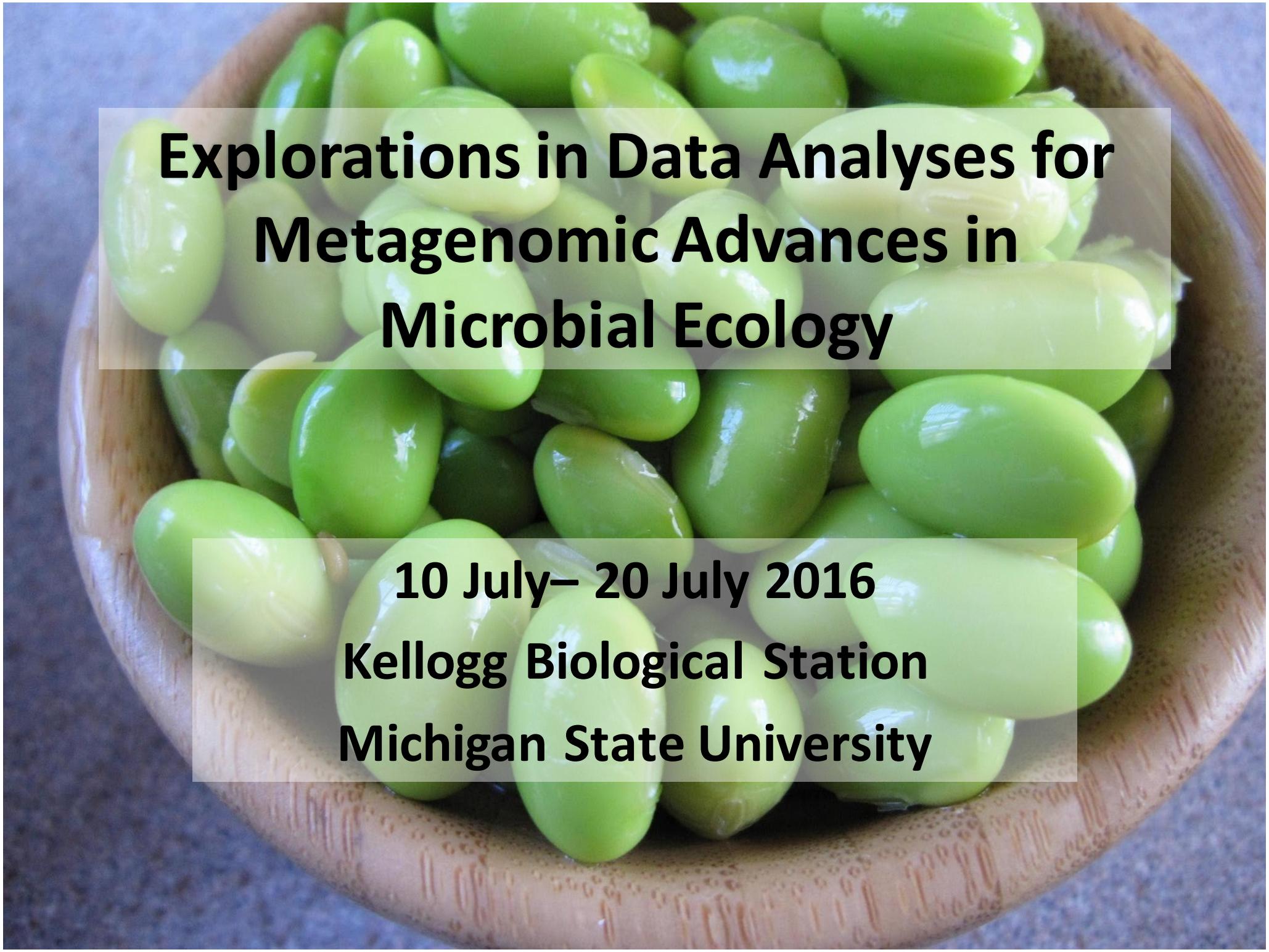


Explorations in Data Analyses for Metagenomic Advances in Microbial Ecology

10 July– 20 July 2016

Kellogg Biological Station

Michigan State University

GOOD MORNING!

- Please make a nametag (back table)
- Take 1 **blue** and 1 **pink** stickie (back table)

Outline

- Overview of a general amplicon analysis workflow
- What is an OTU and how do we pick them?
- What are the features and formats of an OTU table?
- Assigning taxonomy: databases

Ecological traits of microbial communities

Understand the Nature of the Beast. Microbial community data are:

- "Species" rich
- Depend on operational taxonomic unit (OTU) definitions
- Dynamic : sensitive to environmental changes
- Distinctive: even very similar habitats "house" distinct microbial communities (e.g., every human has her own gut community)
- Influenced by dispersal?
- Influenced by gene-swapping (phage, HGT)
- Large proportion of dormant members
- Large proportion of rare members

(A beast, hyperboleandahalf.blogspot.com)

Many many options

- What sequencer?
- If amplicon, which gene? Which variable region?
- What quality control options?
- Defining OTUs
- Describing communities
- Testing hypotheses
- Visualizing results

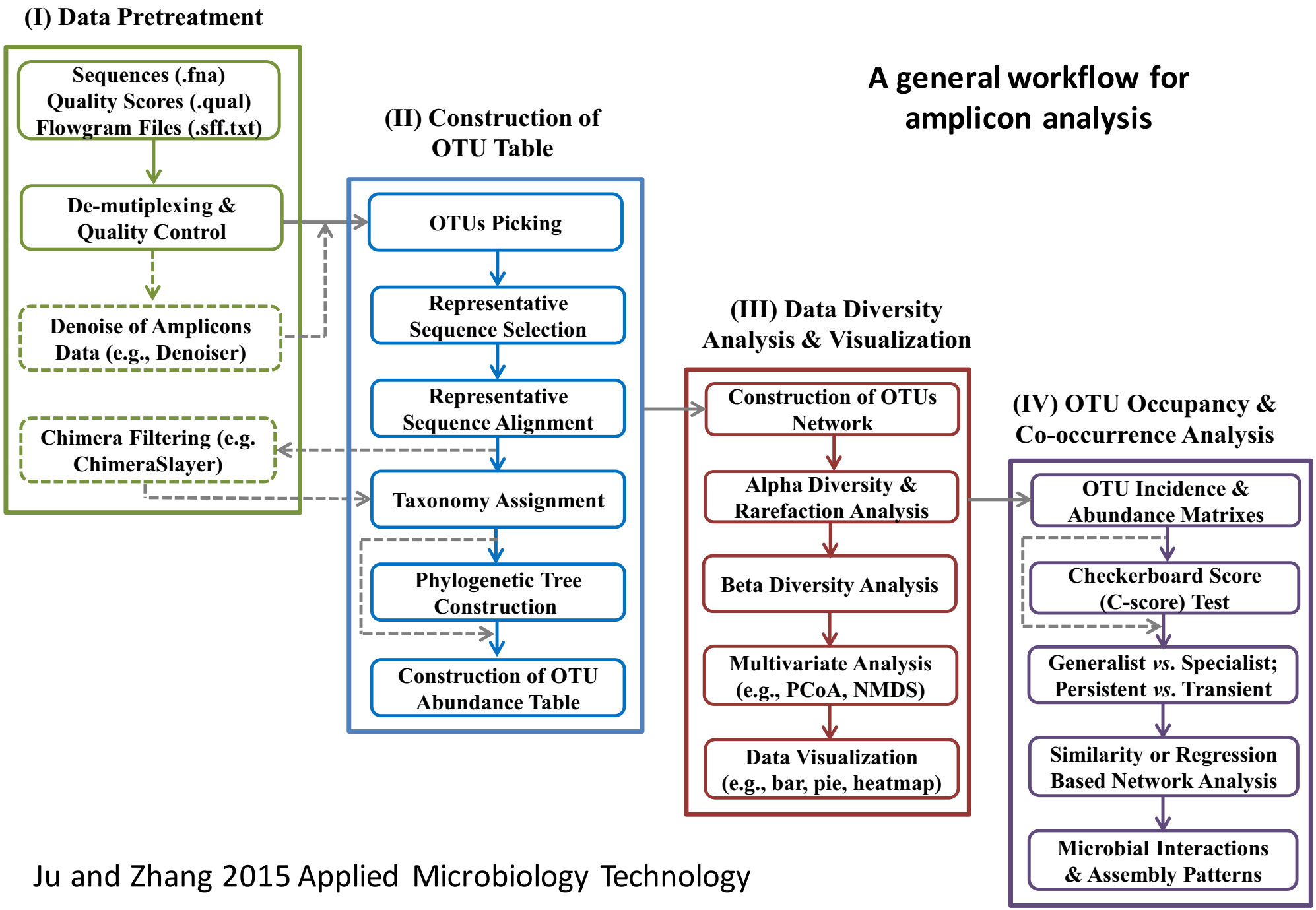
Outline

- Overview of a general amplicon analysis workflow
- What is an OTU and how do we pick them?
- What are the features and formats of an OTU table?
- Assigning taxonomy: databases

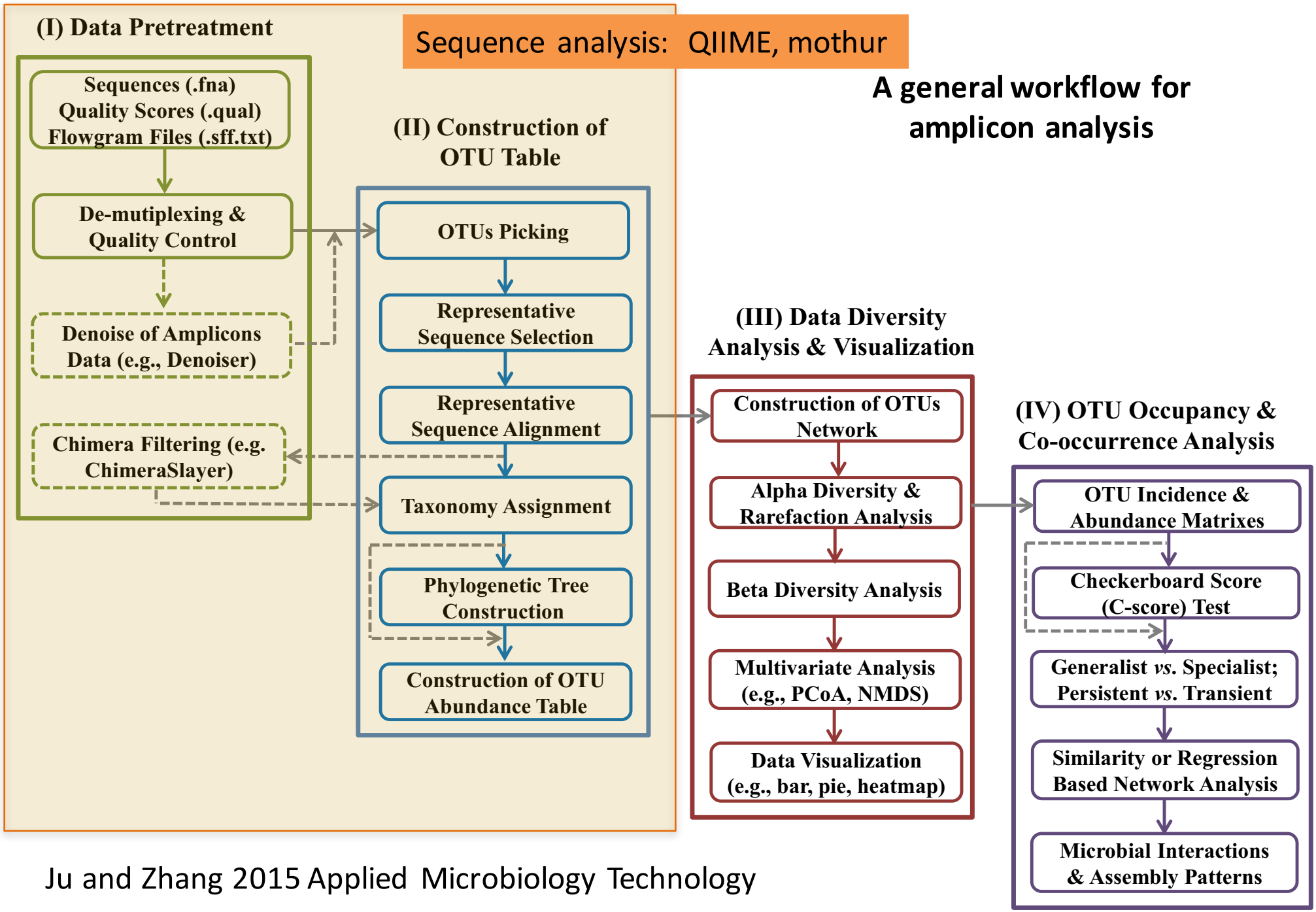
General amplicon workflow

Where to we start?

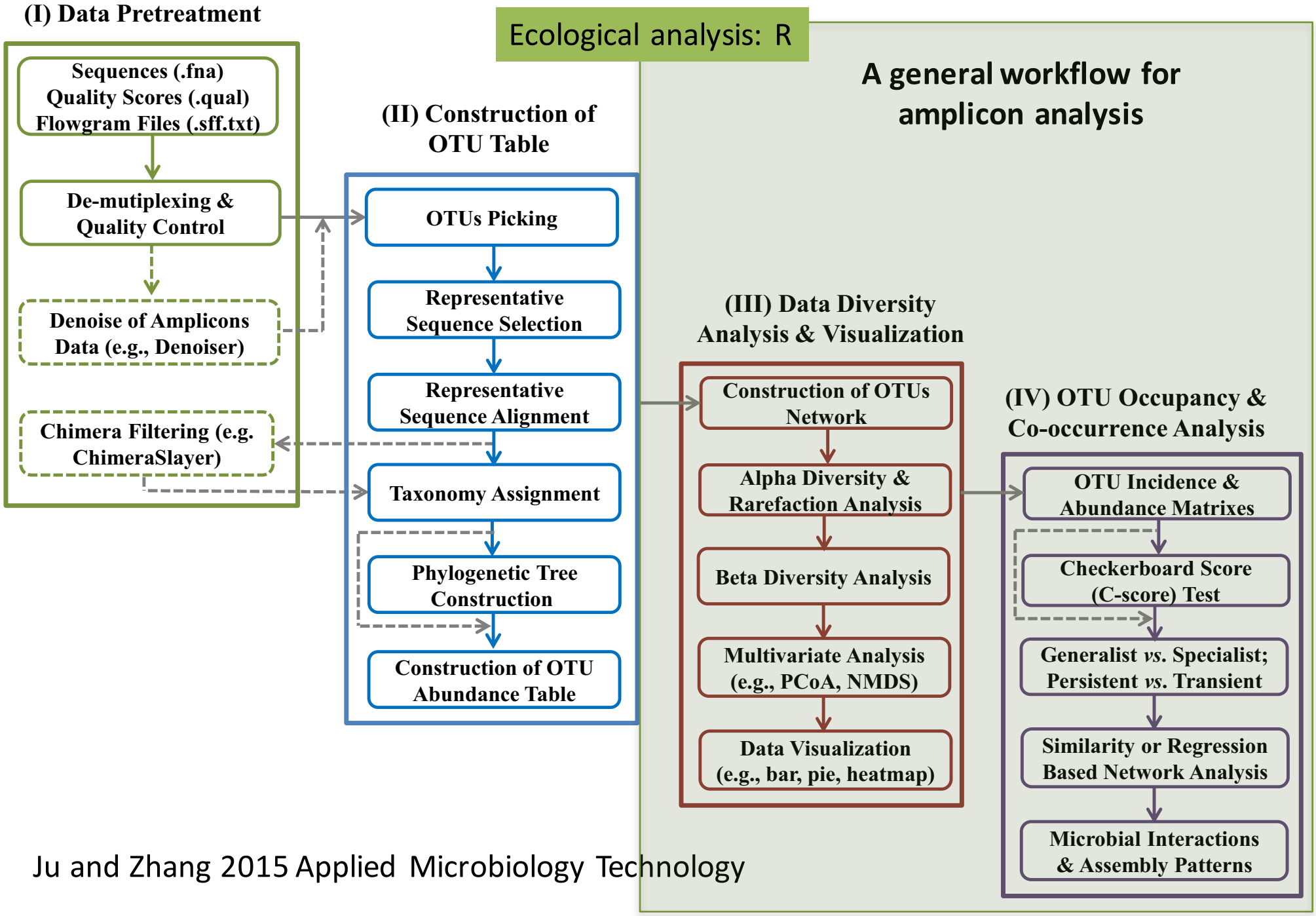
- Question: what are the steps in amplicon sequence analysis? To the white board!



Ju and Zhang 2015 Applied Microbiology Technology



Ju and Zhang 2015 Applied Microbiology Technology



Ju and Zhang 2015 Applied Microbiology Technology

Intro to amplicon sequence analysis

- QIIME:
  - Merging paired-end Illumina Reads
  - Open-reference clustering sequences by 97% identity
  - Picking representative sequences
  - assigning taxonomy to sequences
  - Building and alignment and phylogenetic tree
  - Building an “even” OTU table : equal No. sequences per sample so that comparisons can be made
  - Calculating within-sample (alpha) diversity and comparative (beta diversity)

Intro to amplicon sequence analysis

- mothur:
  - Merging paired-end Illumina Reads
  - Quality filtering based on trees/alignment
  - OTUs based on alignment to high-quality reference alignment
  - Create a sequence x sequence distance matrix based on alignment
  - Cluster sequences into OTUs based on percent similarity
  - Create an OTU table
  - Building an “even” OTU table : equal No. sequences per sample so that comparisons can be made
  - Assign taxonomy
  - Calculating within-sample (alpha) diversity and comparative (beta) diversity)

QIIME and mothur

- QIIME: Representative sequences, open-reference/de novo OTU options
- Mothur: high-quality alignment based OTUs

- See Pat's discussion here:
  - http://blog.mothur.org/2016/01/12/mothur-and-qiime/

- There are other workflows! RDP and usearch

Outline

- Overview of a general amplicon analysis workflow
- What is an OTU and how do we pick them?
- What are the features and formats of an OTU table?
- Assigning taxonomy: databases

OTUs

The "OTU"
operational taxonomic unit

- Species = basic unit of classification
- Defined somewhat arbitrarily
- Typical = 97% sequence identity
  - Originally, identity based on *full length* 16S rRNA gene
  - roughly equivalent to genus level
  - Does not well-distinguish "taxa" for all bacteria (*e.g., Streptomyces*)
- Different methods of defining OTUs will result in different numbers of taxa! Different numbers of taxa mean different perspectives of diversity!

Approaches to Picking OTUs

- Reference based : percent identity to defined taxa populating in a reference database
  - Pros: you know the taxa are "real"!
  - Cons: Weird environments don't have many representatives in databases, only as good as your database, could end up throwing out a lot of real stuff

- De novo : percent identity to other sequences in the dataset; taxonomic assignment to the OTUs happens afterwards
  - Pros: Good for weird environments with low representation in databases
  - Cons: Computationally expensive, "greedy" algorithms can artificially inflate diversity

- Open reference : cluster against a reference db first, and anything that doesn't hit gets clustered de novo
  - Best of both worlds? Now can optimized so that new de novo OTUs are added to the original database and used subsequently in "reference" clustering
  - See Rideout et al. 2014 PeerJ

Outline

- Overview of a general amplicon analysis workflow
- What is an OTU and how do we pick them?
- What are the features and formats of an OTU table?
- Assigning taxonomy: databases

OTU Tables

The OTU Table

- An **OTU table** is the input file for community analyses. It contains information about the abundance of each OTU within every sample. OTU tables can be (**classic, .txt**) or (.**biom**) format.
- The OTU table is an **output** of the amplicon sequence analysis pipeline, and can be moved into other software for ecological analysis (e.g., R)
- An OTU table can be made for units that aren't amplicon sequences, e.g. metagenome COGs, bird counts – it is borrowed from ecology as a way to summarize species in samples. Thus, many different types of data can be analyzed using the same methods if you can generate an OTU table from the data.

Information in an OTU table

- Number of occurrences (per sample and for the whole dataset)
- Total no. OTUs observed in the dataset
- Average abundance of OTUs
- Richness (no. OTUs per sample, mean, max, min, range)
- Number of singletons (OTUs detected only once in a dataset)
- Calculate: Diversity, Evenness (equitability of OTU abundances, including rarity and dominance)
- Number of samples (communities) in your dataset

- Dimensions of an OTU table: rows (taxa) x columns (samples/communities)

Common features of microbial OTU tables

- Redundant: more than one taxa has the exact same pattern
- Unknown underlying distribution
- Contain many “zeros”
- Many samples and OTUs; computationally large

(A beast, hyperboleandahalf.blogspot.com)

What does an OTU table look like, data-style

Raw
Straight up counts*

	Soil 1	Soil 2	Soil 3
OTU 1	0	3000	23
OTU 2	1	5	5
OTU 3	20	100	100

**note: data must be subsampled to an even sequencing effort!*

Relative
Percent or proportion

	Soil 1	Soil 2	Soil 3
OTU 1	0	0.966	0.179
OTU 2	0.047	0.002	0.039
OTU 3	0.953	0.032	0.782

Binary
Presence/absence

	Soil 1	Soil 2	Soil 3
OTU 1	0	1	1
OTU 2	1	1	1
OTU 3	1	1	1

Be ware the .Biom table my friend

- Biom table – more concise & faster computing for extra large datasets – more about that in a bit
  - Newer formats : “biom2”
  - See McDonald et al. 2012. “The Biological Observation Matrix…” *GigaScience*.
  - Biom-format.org

Biom formatted OTU tables

- .biom format

Link:

http://biom-format.org

This is all changing very often!! Biom formats are constantly improved, keep up with when changes are anticipated

A dense representation of an OTU table:

```
OTU ID PC.354  PC.355  PC.356
OTU0    0    0    4
OTU1    6    0    0
OTU2    1    0    7
OTU3    0    0    3
```

Traditional OTU table - microbial communities have lots of 0's

A sparse representation of an OTU table:

```
PC.354 OTU1 6
PC.354 OTU2 1
PC.356 OTU0 4
PC.356 OTU2 7
PC.356 OTU3 3
```

.biom formatted – only list present taxa

Assigning taxonomy

- You can only do as well as your database
- Databases can have errors
  - 16S: greengenes, Silva, RDP
  - Fungal ITS: ITS2, ITS1
  - 18S: Silva
  - Viral: PHAST

- We'll explore some of these databases and other sequence repositories on the last day

Questions?