QIIME Tutorial

Workshop on Comparative Genomics

July 2011

Tony Walters

William.a.walters@gmail.com

Knight Lab

Important Links

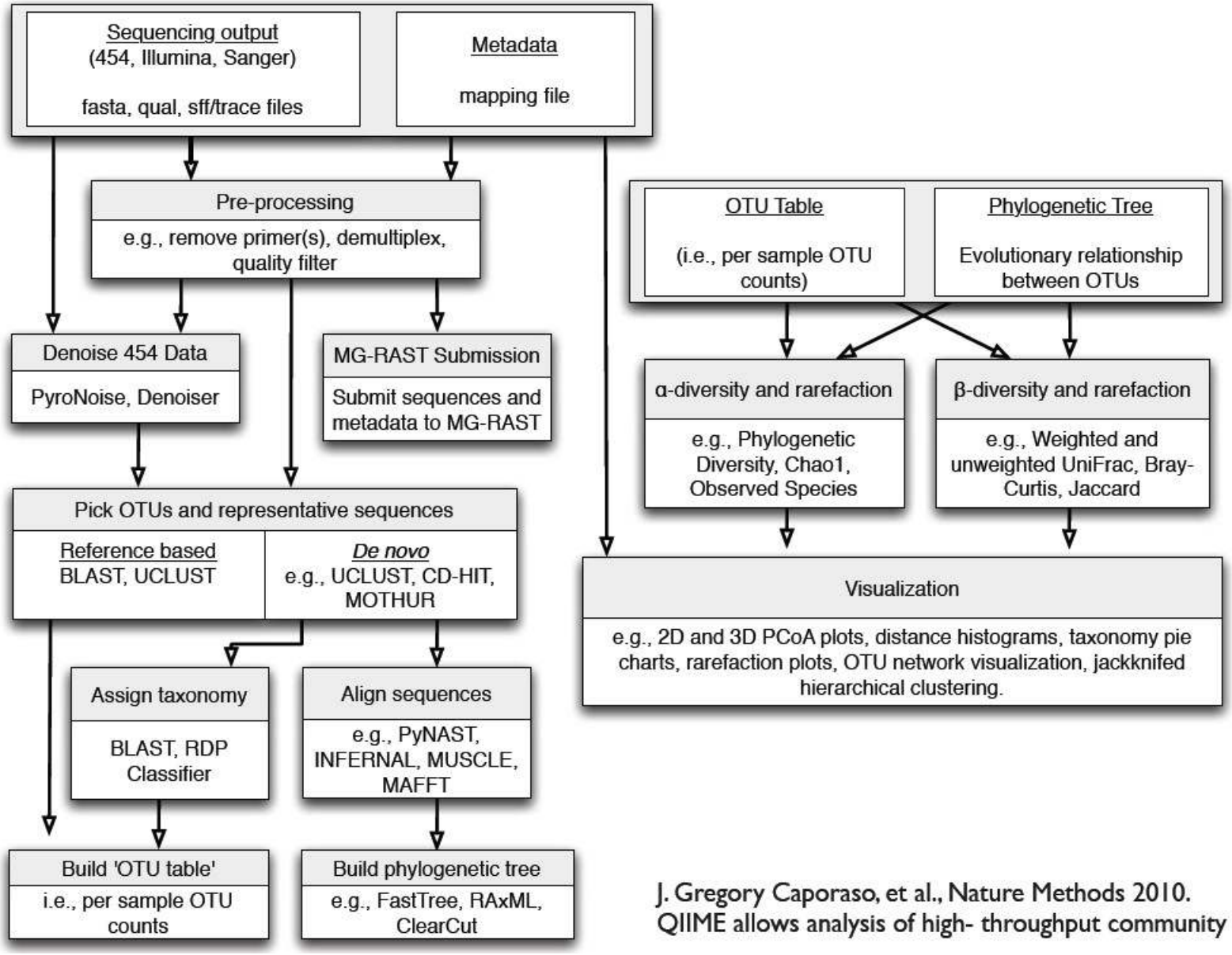
- Tutorial Files

  http://bit.ly/nXmmIL

- QIIME Overview Tutorial (for future reference)

  http://qiime.org/tutorials/tutorial.html

Quantitative Insights Into Microbial Ecology



J. Gregory Caporaso, et al., Nature Methods 2010.
QIIME allows analysis of high- throughput community sequencing data

Raw Sequencing Data

454 Data:

.fna sequence files in FASTA format

Fasting_Example.fna

.qual quality scores

Fasting_Example.qual

.sff flowgram files

Illumina Data:

.fastq both sequence and quality scores

Mapping File

Fasting_Map.txt

#SampleID	BarcodeSequence	LinkerPrimerSequence	Treatment	DOB	Description
#Example mapping file for the QIIME analysis package. These 9 samples are from a study of the effects of exercise and diet on mouse cardiac physiology (Crawford, et al, PNAS, 2009).					
PC.354	AGCACGAGCCTA	YATGCTGCCTCCCGTAGGAGT	Control	20061218	Control_mouse_I.D._354
PC.355	AACTCGTCGATG	YATGCTGCCTCCCGTAGGAGT	Control	20061218	Control_mouse_I.D._355
PC.356	ACAGACCACTCA	YATGCTGCCTCCCGTAGGAGT	Control	20061126	Control_mouse_I.D._356
PC.481	ACCAGCGACTAG	YATGCTGCCTCCCGTAGGAGT	Control	20070314	Control_mouse_I.D._481
PC.593	AGCAGCACTTGT	YATGCTGCCTCCCGTAGGAGT	Control	20071210	Control_mouse_I.D._593
PC.607	AACTGTGCGTAC	YATGCTGCCTCCCGTAGGAGT	Fast	20071112	Fasting_mouse_I.D._607
PC.634	ACAGAGTCGGCT	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._634
PC.635	ACCGCAGAGTCA	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._635
PC.636	ACGGTGAGTGTC	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._636

Verify the Mapping File

Command:

check_id_map.py –m Fasting_Map.txt –o check_mapping/

Result:

No errors or warnings for mapfile Fasting_Map.txt

Fasting_Map.log and Fasting_Map_corrected.txt files will be created in the check_mapping folder which state that no errors have been found.

Also, this is one of the only QIIME modules that prints a message about the input file by default.

Invalid Mapping File
Fasting_Map_bad.txt

#SampleID	BarcodeSequence	LinkerPrimerSequence	Treatment	DOB	Description
#Example mapping file for the QIIME analysis package. These 9 samples are from a study of the effects of exercise and diet on mouse cardiac physiology (Crawford, et al, PNAS, 2009).					
PC.354	AGCACGAGCCTA	YATGCTGCCTCCCGTAGGAGT	Control	20061218	Control_mouse_I.D._354
PC.354	AACTCGTCGATG	YATGCTGCCTCCCGTAGGAGT	Control	20061218	Control_mouse_I.D._355
PC.356	ACAGACCACTCA	YATGCTGCCTCCCGTAGGAGT	Control	20061126	Control_mouse_I.D._356
PC.481&&	ACCAGCGACTAG	YATGCTGCCTCCCGTAGGAGT	Control	20070314	Control_mouse_I.D._481
PC.593	AGCAGCACTTGT	YATGCTGCCTCCCGTAGGAGT	Control	20071210	Control_mouse_I.D._593
PC.607	AACTGTGCGTAC		Fast	20071112	Fasting_mouse_I.D._607
PC.634	ACAGAGTCGGCT	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._634
PC.635	ACCGCAGAGTCA	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._635
PC.636	ACGGTGAGTGTC	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._636

Verify the bad Mapping File

Command:

check_id_map.py –m Fasting_Map_bad.txt –o ../test_check_id_map/

Errors and/or warnings occurred, see log file ../test_check_id_map/Fasting_Map_bad.log

Contents of Fasting_Map_bad.log

#Listed locations of errors/warnings in row/column format have an index that is

#in reference to the beginning of the sample IDs and metadata.

#Location (0,0) is the first SampleID in a given data set.

WARNINGS ----------

0: Removed bad chars from cell PC.481&& (now PC.481..) in sample id PC.481.., col #SampleID. Location (row, column): 3,0

1: These sample ids have duplicate ids:

PC.354,PC.354: PC.354

Row, column for all duplicate ids:

Location (row, column): 0,0

Location (row, column): 1,0

2: Missing primer. Location (row, column): 5,2

0: Removed bad chars from cell PC.481&& (now PC.481..) in sample id PC.481.., col #SampleID. Location (row, column): 3,0

1: These sample ids have duplicate ids:

PC.354,PC.354: PC.354

Row, column for all duplicate ids:

Location (row, column): 0,0

Location (row, column): 1,0

2: Missing primer. Location (row, column): 5,2

#SampleID	BarcodeSequence	LinkerPrimerSequence	Treatment	DOB	Description
#Example mapping file for the QIIME analysis package. These 9 samples are from a study of the effects of exercise and diet on mouse cardiac physiology (Crawford, et al, PNAS, 2009).					
PC.354	AGCACGAGCCTA	YATGCTGCCTCCCGTAGGAGT	Control	20061218	Control_mouse_I.D._354
PC.354	AACTCGTCGATG	YATGCTGCCTCCCGTAGGAGT	Control	20061218	Control_mouse_I.D._355
PC.356	ACAGACCACTCA	YATGCTGCCTCCCGTAGGAGT	Control	20061126	Control_mouse_I.D._356
PC.481&&	ACCAGCGACTAG	YATGCTGCCTCCCGTAGGAGT	Control	20070314	Control_mouse_I.D._481
PC.593	AGCAGCACTTGT	YATGCTGCCTCCCGTAGGAGT	Control	20071210	Control_mouse_I.D._593
PC.607	AACTGTGCGTAC		Fast	20071112	Fasting_mouse_I.D._607
PC.634	ACAGAGTCGGCT	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._634
PC.635	ACCGCAGAGTCA	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._635
PC.636	ACGGTGAGTGTC	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._636

Corrected Version (save as Fasting_Map_bad_fixed.txt)

#SampleID	BarcodeSequence	LinkerPrimerSequence	Treatment	DOB	Description
#Example mapping file for the QIIME analysis package. These 9 samples are from a study of the effects of exercise and diet on mouse cardiac physiology (Crawford, et al, PNAS, 2009).					
PC.354	AGCACGAGCCTA	YATGCTGCCTCCCGTAGGAGT	Control	20061218	Control_mouse_I.D._354
PC.355	AACTCGTCGATG	YATGCTGCCTCCCGTAGGAGT	Control	20061218	Control_mouse_I.D._355
PC.356	ACAGACCACTCA	YATGCTGCCTCCCGTAGGAGT	Control	20061126	Control_mouse_I.D._356
PC.481	ACCAGCGACTAG	YATGCTGCCTCCCGTAGGAGT	Control	20070314	Control_mouse_I.D._481
PC.593	AGCAGCACTTGT	YATGCTGCCTCCCGTAGGAGT	Control	20071210	Control_mouse_I.D._593
PC.607	AACTGTGCGTAC	YATGCTGCCTCCCGTAGGAGT	Fast	20071112	Fasting_mouse_I.D._607
PC.634	ACAGAGTCGGCT	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._634
PC.635	ACCGCAGAGTCA	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._635
PC.636	ACGGTGAGTGTC	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._636

Now test the corrected file…

Command:

check_id_map.py -m Fasting_Map_bad_fixed.txt -o ../test_check_id_map

Output:

No errors or warnings for mapfile Fasting_Map_bad_fixed.txt

Check_id_map.py will also create a _corrected.txt file, but this will only replace bad characters and will not fix other problems.

Assigning Sequences to Samples split_libraries.py

Barcodes are used to demultiplex reads according to SampleID in the mapping file.

A large number of quality filters are used, including quality scores, sequence length, barcode/primer mismatches, ambiguous bases, homopolymers, etc.

Demultiplexing of Tutorial Data

Command:

split_libraries.py -m Fasting_Map.txt -f Fasting_Example.fna -q Fasting_Example.qual -o split_library_output

Three files will be created in the split_library_output folder:

histograms.txt, seqs.fna, split_library_log.txt

Output from split_libraries.py

histograms.txt – shows counts and lengths of sequences before and after processing

split_library_log.txt – shows details about number of sequences written for each SampleID, and counts of sequences discarded for each quality filter.

seqs.fna file

>PC.634_1 FLP3FBN01ELBSX orig_bc=ACAGAGTCGGCT new_bc=ACAGAGTCGGCT bc_diffs=0

CTGGGCCGTGTCTCAGTCCCAATGTGGCCGTTTACCCTCTCAGGCC
GGCTACGCATCATCGCCTTGGTGGGCCGTTACCTCACCAACTAG
CTAATGCGCCGCAGGTCCATCCATGTTCACGCCTTGATGGGCGC
TTTAATATACTGAGCATGCGCTCTGTATACCTATCCGGTTTTAGCT
ACCGTTTCCAGCAGTTATCCCGGACACATGGGCTAGG

>PC.634_2 FLP3FBN01EG8AX orig_bc=ACAGAGTCGGCT new_bc=ACAGAGTCGGCT bc_diffs=0

TTGGACCGTGTCTCAGTTCCAATGTGGGGGCCTTCCTCTCAGAACC
CCTATCCATCGAAGGCTTGGTGGGCCGTTACCCCGCCA.....

PC.634	ACAGAGTCGGCT	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._634

Modifying qiime_parameters.txt

qiime_parameters.txt

WARNING: DO NOT EDIT OR DELETE Qiime/qiime_parameters.txt. Users should copy this file and edit copies of it.

split libraries parameters

split_libraries:min-seq-length

split_libraries:max-seq-length

split_libraries:trim-seq-length

split_libraries:min-qual-score

split_libraries:keep-primer

split_libraries:keep-barcode

split_libraries:max-ambig

split_libraries:max-homopolymer

split_libraries:max-primer-mismatch

……..

Modifying qiime_parameters.txt

```
# Multiple sequence alignment parameters
align_seqs:template_fp      /home/tony/Desktop/greengenes_core_set/
    core_set_aligned.fasta
align_seqs:alignment_method pynast
align_seqs:pairwise_alignment_method        uclust
align_seqs:blast_db
align_seqs:min_length       150
align_seqs:min_percent_id       75.0

# Alignment filtering (prior to tree-building) parameters
filter_alignment:lane_mask_fp /home/tony/Desktop/greengenes_core_set/
    lanemask_in_1s_and_0s
filter_alignment:allowed_gap_frac   0.999999
filter_alignment:remove_outliers    False
filter_alignment:threshold 3.0
```

Workflow Script

`pick_otus_through_otu_table.py`

Input:
- sequences
- output directory
- parameters file (optional)

Primary output:
- OTU table
- phylogenetic tree

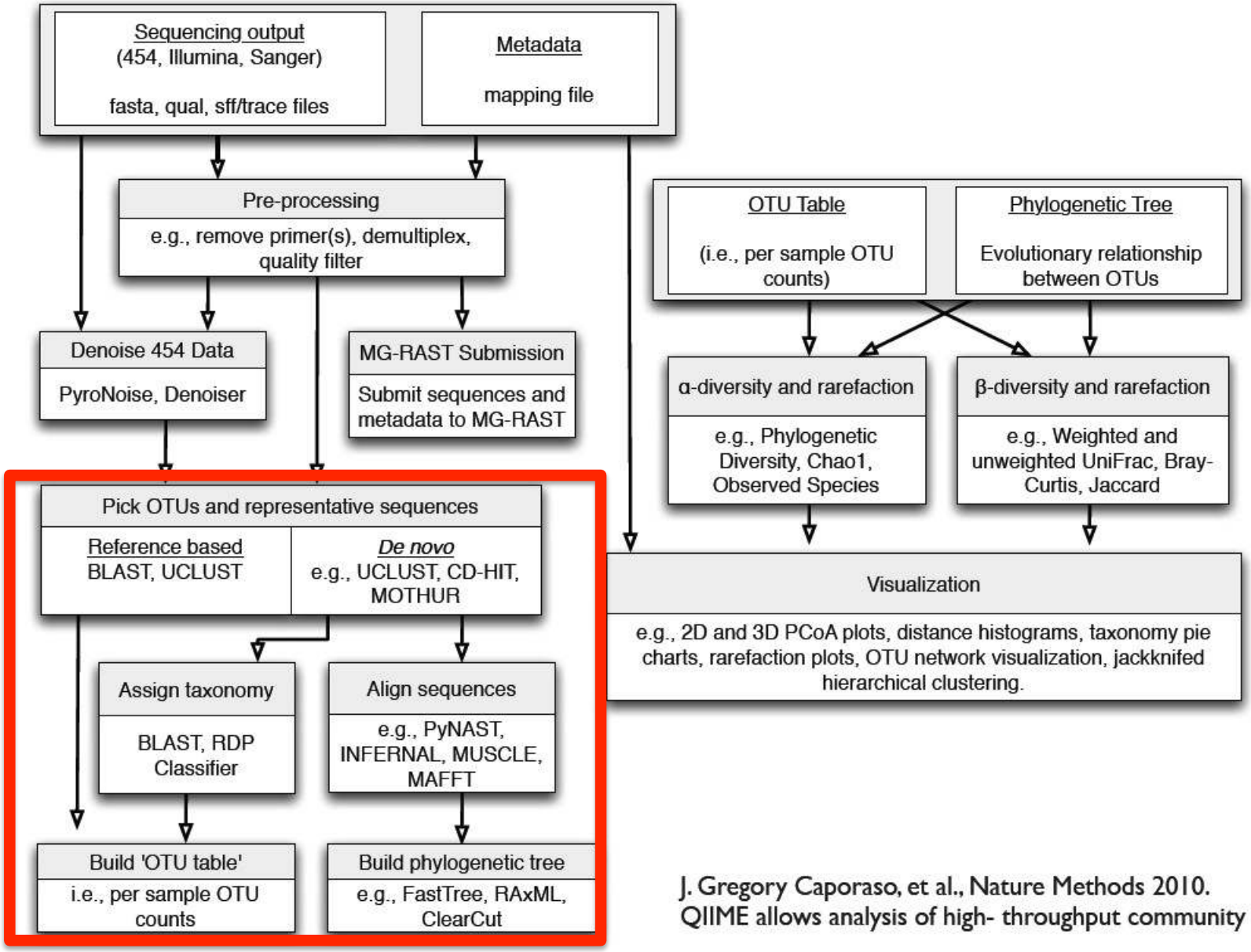
Additional output:
- OTU map (i.e., OTU id to seq id(s))
- representative sequences fasta (aligned and unaligned)
- taxonomy assignment (i.e., OTU id to taxonomy string)

Running the Workflow

Command:

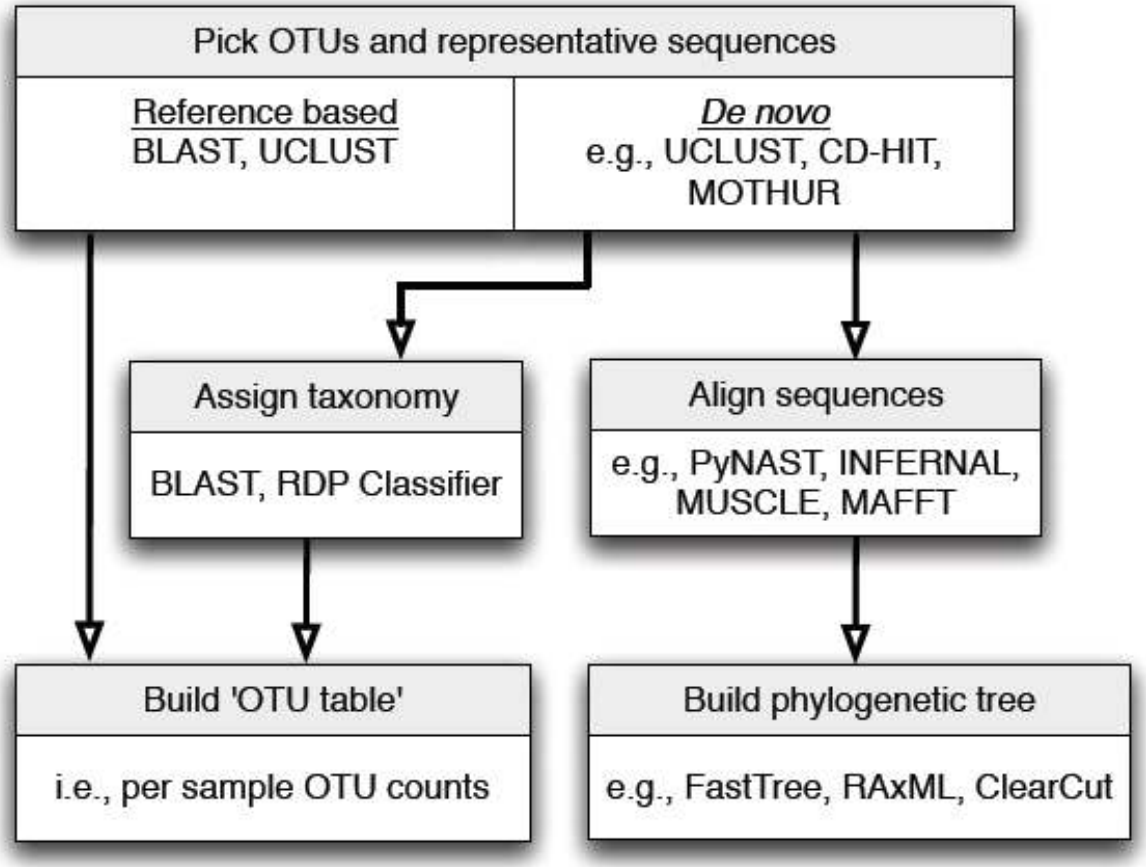
pick_otus_through_otu_table.py -i split_library_output /seqs.fna -o split_library_output /wf_output/ -p qiime_parameters.txt –f –p

Quantitative Insights Into Microbial Ecology



J. Gregory Caporaso, et al., Nature Methods 2010.
QIIME allows analysis of high- throughput community sequencing data

Getting from sequences to an OTU table



OTU Table

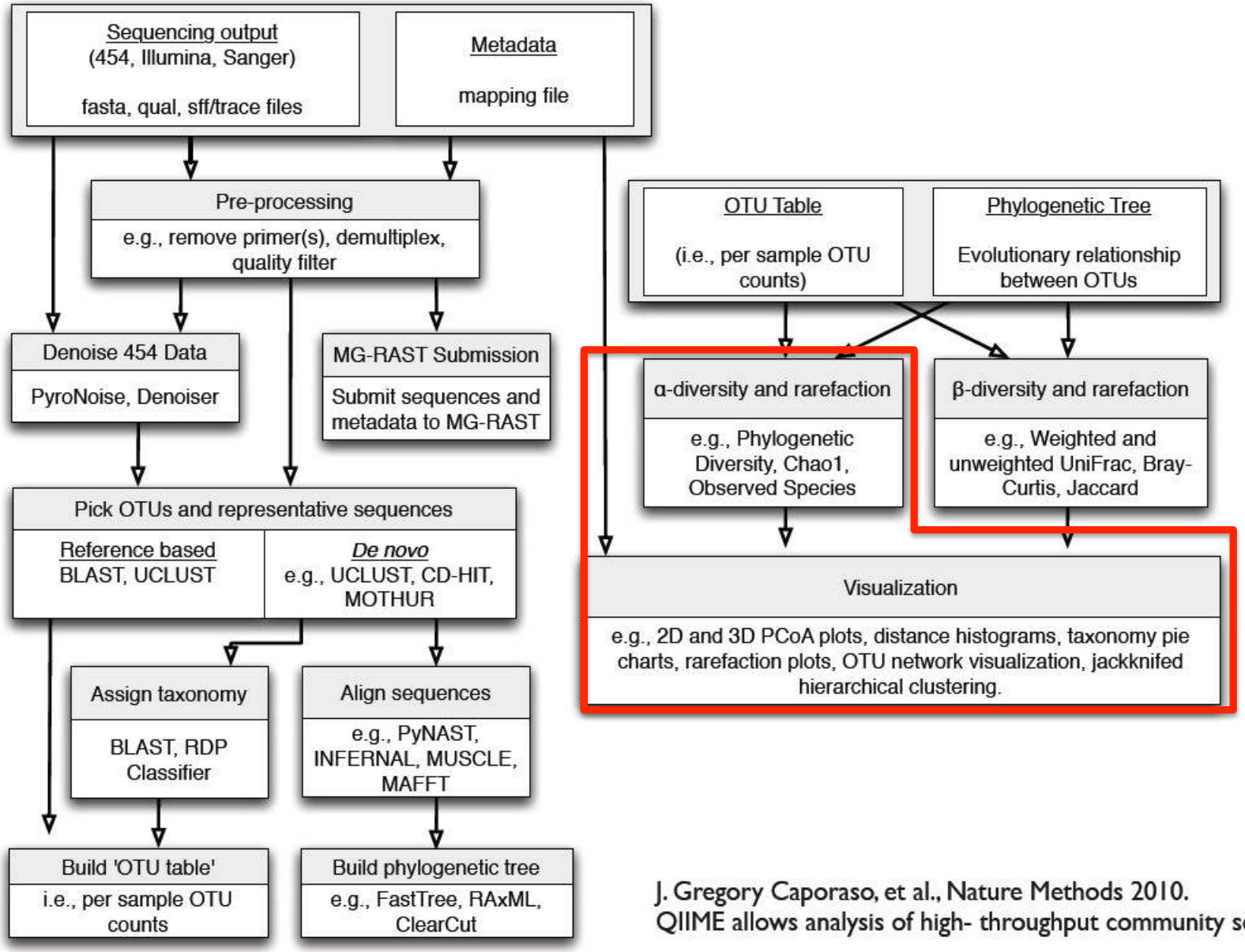
# QIIME v1.3.0-dev OTU table										
#OTU ID	PC.354	PC.355	PC.356	PC.481	PC.593	PC.607	PC.634	PC.635	PC.636	Consensus Lineage
0	0	0	0	0	0	0	2	1	0	Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides
1	0	0	0	0	0	0	0	0	1	Bacteria;Bacteroidetes
2	0	0	0	0	0	0	0	0	1	Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Staphylococcus
3	0	0	0	1	0	0	0	0	0	Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae
4	1	0	0	0	0	0	0	0	0	Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae
5	0	1	1	0	1	0	0	0	0	Bacteria;Firmicutes;Clostridia;Clostridiales
6	0	0	0	1	0	0	0	1	0	Bacteria;Firmicutes;Clostridia;Clostridiales

Alpha Diversity

Command:

alpha_rarefaction.py -m Fasting_Map.txt -o split_library_output /wf_output/alpha_div/ -i split_library_output /wf_output/otu_table.txt -t split_library_output /wf_output/rep_set.tre –f -v

Quantitative Insights Into Microbial Ecology



J. Gregory Caporaso, et al., Nature Methods 2010.
QIIME allows analysis of high- throughput community sequencing data

Alpha Diversity

Input:

- - OTU table
- - parameters file (optional)
- - output directory
- - mapping file
- - phylogenetic tree (optional, for calculating phylogenetic diversity metrics)

Primary output:

- - rarefaction plots

Additional output:

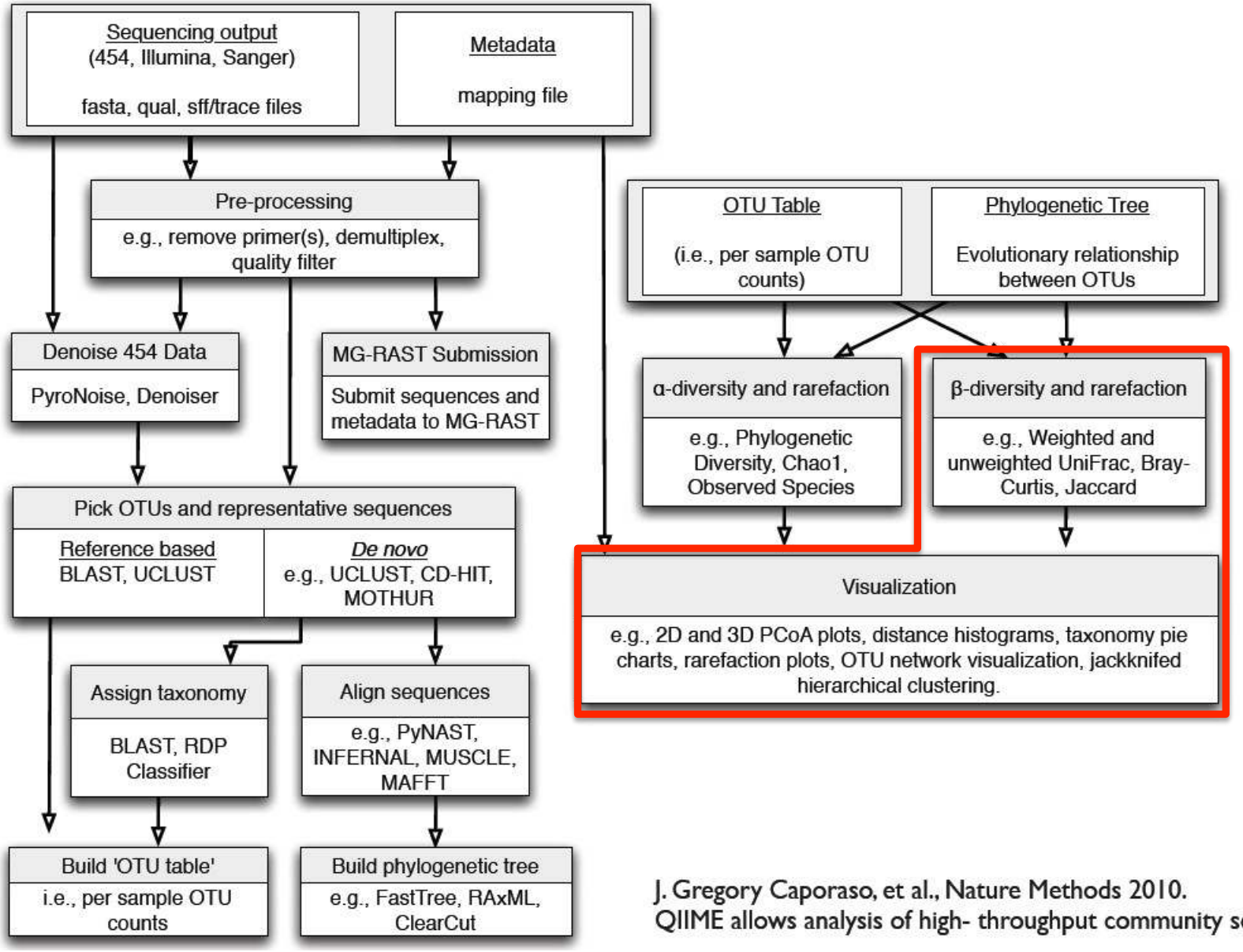
- - alpha diversity values at varied sequencing depths

Beta Diversity

Command:

beta_diversity_through_plots.py -i split_library_output /wf_output/otu_table.txt -m Fasting_Map.txt -o split_library_output / wf_output/beta_div/ -t split_library_output / wf_output/rep_set.tre -e 146 –f -v

Quantitative Insights Into Microbial Ecology



J. Gregory Caporaso, et al., Nature Methods 2010.
QIIME allows analysis of high- throughput community sequencing data

Beta Diversity

Input:

- - OTU table
- - parameters file (optional)
- - output directory
- - mapping file
- - phylogenetic tree (optional, for calculating phylogenetic diversity metrics)

Primary output:

- - 3D plots
- - 2D plots
- - distance histograms

Additional output:

- - between-sample distance matrices
- - principal coordinate matrices