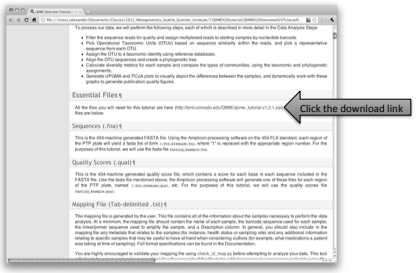
QIIME tutorial

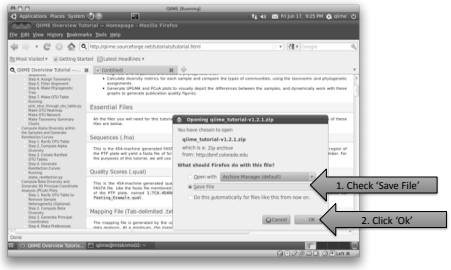
http://qiime.org/tutorials/tutorial.html

Download the necessary files



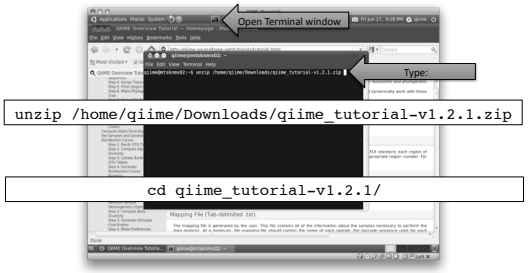
2

Save files in the virtual machine



3

Expand the tutorial files archive



4

Tutorial files

Fasting_Example.fna	Fasta sequences
Fasting_Example.qual	Quality scores
Fasting_Example.sff	Binary sequence/quality file from 454 machine
Fasting_Map.txt	Meta-data for the samples in the run
custom_parameters.txt	Parameters of QIIME analysis
qiime_tutorial_commands_serial.sh	Commands to execute QIIME pipeline

5

Quick glance at the files

```
> head Fasting_Example.fna
```

```
>FLP3FBN01ELBSX length=250 xy=1766_0111 region=1 …
ACAGAGTCGGCTCATGCTGCCTCCCGTAGGAGTCTGGGCCGTGTCTCAGT
CCCAATGTGGCCGTTTACCCTCTCAGGCCGGCTACGCATCATCGCCTTGG
TGGGCCGTTACCTCACCAACTAGCTAATGCGCCGCAGGTCCATCCATGTT
CACGCCTTGATGGGCGCTTTAATATACTGAGCATGCGCTCTGTATACCTA
TCCGGTTTTAGCTACCGTTTCCAGCAGTTATCCCGGACACATGGGCTAGG
>FLP3FBN01EG8AX length=276 xy=1719_1463 region=1 …
ACAGAGTCGGCTCATGCTGCCTCCCGTAGGAGTTTGGACCGTGTCTCAGT
TCCAATGTGGGGGCCTTCCTCTCAGAACCCCTATCCATCGAAGGCTTGGT
GGGCCGTTACCCCGCCAACAACCTAATGGAACGCATCCCCATCGATGACC
GAAGTTCTTTAATAGTTCTACCATGCGGAA
```

6

Quick glance at the files

```
> head Fasting_Example.fna
```

```
>FLP3FBN01ELBSX length=250 xy=1766_0111 region=1 …
37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37
36 36 33 33 33 36 37 37 37 37 37 37 40 40 40 39 39 38 40 40
40 40 40 40 40 37 37 37 37 37 35 35 35 37 37 37 37 37 35 35
35 31 31 23 23 23 31 21 21 21 35 35 37 37 37 36 36 36 36 36
36 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37
37 28 28 28 36 36 37 37 37 37 37 37 37 37 37 37 37 37 37 37
37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 36 36 36 37 37
37 37 37 37 37 37 37 37 37 37 37 36 36 36 37 37 37 37 37 37
37 37 37 37 37 37 37 37 37 37 37 35 32 32 32 32 35 37 37 37
37 37 37 37 37 37 37 37 37 37 37 37 37 37 36 32 32 32 36 37
35 32 32 32 32 32 32 32 32 36 37 37 37 37 36 36 31 31 32 32
36 36 36 36 36 36 36 36 36 36 36 28 27 27 27 26 26 26 30 29
30 29 24 24 24 21 15 15 13 13
```


Mapping file

```
#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
```


Sketch of the analysis

- Check the mapping file
- Separate the sequences by barcode
- Define OTUs
- Descriptive analyses
- Diversity analyses
- PCA analysis

Check mapping file

```
> check_id_map.py -m
  Fasting_Map.txt -o
  mapping_output -v
```

Output:

```
mapping_output/
   Fasting_Map_corrected.txt
      corrected mapping (if any errors found)
   Fasting_Map.log
      log of found errors
```


Separate the sequences by barcode

```
> split_libraries.py …
```

Output:

```
split_library_output/
   histograms.txt
      histogram of sequence lengths before and after
splitting
   seqs.fna
      sequences separated by barcode and re-labled
according to sample
   split_library_log.txt
      summary of demultiplexing performed
```


Split library log

```
Number raw input seqs   1339
Length outside bounds of 200 and
1000   0

Num ambiguous bases exceeds limit of
0  4

Missing Qual Score      0

Mean qual score below minimum of 25
1

Max homopolymer run exceeds limit of
6  0

Num mismatches in primer exceeds
limit of 0: 1
```

```
Num Samples     9
Sample ct min/max/mean: 146 /
150 / 148.11

Sample  Sequence Count  Barcode
PC.634  150     ACAGAGTCGGCT
PC.356  150     ACAGACCACTCA
PC.593  149     AGCAGCACTTGT
PC.607  149     AACTGTGCGTAC
PC.354  148     AGCACGAGCCTA
PC.635  148     ACCGCAGAGTCA
PC.636  147     ACGGTGAGTGTC
PC.481  146     ACCAGCGACTAG
PC.355  146     AACTCGTCGATG

Total number seqs written
1333
```

Sequences separated by barcode are re-labeled with new identifiers

```
>PC.634_1 FLP3FBN01ELBSX orig_bc=ACAGAGTCGGCT
new_bc=ACAGAGTCGGCT bc_diffs=0
CTGGGCCG …
>PC.634_2 FLP3FBN01EG8AX orig_bc=ACAGAGTCGGCT
new_bc=ACAGAGTCGGCT bc_diffs=0
TTGGACCG …
>PC.354_3 FLP3FBN01EEWKD orig_bc=AGCACGAGCCTA
new_bc=AGCACGAGCCTA bc_diffs=0
TTGGGCCG …
>PC.481_4 FLP3FBN01DEHK3 orig_bc=ACCAGCGACTAG
new_bc=ACCAGCGACTAG bc_diffs=0
CTGGGCCGTG …
```

13

Define OTUs

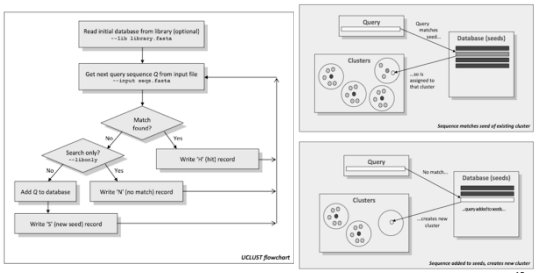
```
> pick_otus_through_otu_table.py …
```

1. Pick OTUs with uclust at similarity of 0.97;
2. Pick a representative set with the first method;
3. Align the representative set with PyNAST
4. Assign taxonomy with RDP classifier;
5. Filter the alignment prior to tree building - remove positions which are all gaps, and specified as 0 in the lanemask
6. Build a phylogenetic tree with FastTree;
7. Build an OTU table.

14

uclust

- http://drive5.com/usearch/usearch3.0.html

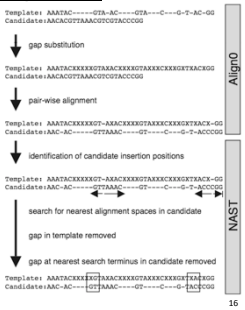


15

PyNAST (Nearest Alignment Space Termination)

- http://pynast.sourceforge.net/

PyNAST: a flexible tool for aligning sequences to a template alignment. Caporaso, JG.; Bittinger, K; Bushman, FD.; DeSantis, TZ., Andersen, GL.; Knight, R. January 15, 2010, DOI 10.1093/bioinformatics/btp636. Bioinformatics 26: 266-267.



16

RDP classifier

- http://rdp.cme.msu.edu/classifier/

Classes: Well-defined fully taxonomies of bacteria
Features: frequencies of all 4^8 = 65,536 possible 8-base
Assignment: According to highest posterior probability
Bootstrap significance: Randomize input sequence 8-mers
Under conditional independence assumption the classification model is:

$$P(C \mid F_1,\ldots,F_n) = \frac{P(C)P(F_1,\ldots,F \mid C)}{P(F_1,\ldots,F)} \propto P(C)P(F_1,\ldots,F \mid C)$$

$$= P(C)P(F_1 \mid C) \times P(F_2 \mid C,F_1) \times \ldots \times P(F_n \mid C,F_1,\ldots,F_{n-1})$$

$$= P(C)P(F_1 \mid C) \times P(F_2 \mid C) \times \ldots \times P(F_n \mid C) = P(C)\prod_{j=1}^{n} P(F_j \mid C)$$

Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 73(16):5261-7.

17

FastTree

- http://www.microbesonline.org/fasttree/

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2 -- Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE, 5(3):e9490. doi:10.1371/journal.pone.0009490.

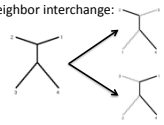
Heuristic phylogenetic tree reconstruction method consisting of the following steps:

1. Heuristic neighbor-joining
2. Reducing the length of the tree:
   - Nearest-neighbor interchanges
   - Subtree-prune-regraft moves
   - Distance model moves
3. Maximizing the tree's likelihood with NNIs
4. Local support values

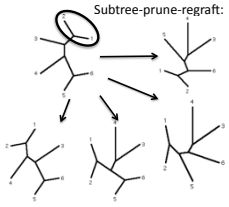
18

FastTree

Nearest-neighbor interchange:



Subtree-prune-regraft:



Distance model moves:
Update the branch lengths using Jukes-Cantor substitution distance:
-0.75*log(1 - 4/3 d),
where d is the proportion of positions that differ

19

(some) Output files

```
wf_da/
 log_[DATETIME].txt – summary of parameters and commands
 uclust_picked_otus/
  seqs_otus.txt – listing of sequences belonging to OTUs
  rep_set/
   seqs_rep_set.fasta – representative sequences for each OTU
   pynast_aligned_seqs/
     fasttree_phylogeny/
      seqs_rep_set.tre – NEWICK format phylogeny reconstructed
     seqs_rep_set_aligned.fasta – aligned representative sequences
   rdp_assigned_taxonomy/
     otu_table/
      seqs_otu_table.txt – OTU table
     seqs_rep_set_tax_assignments.txt – taxonomic assignment of OTUs
```

20

Descriptive analyses

- Heatmaps
- Co-occurrence network
- OTU abundance charts

21

Descriptive analysis output

- `wf_da/uclust_picked_otus/rep_set/rdp_assigned_taxonomy/otu_table/`
  - `OTU_Heatmap`
  - `OTU_Network`
  - `Taxa_Charts`

22

Diversity analyses

- `alpha_rarefaction.py …`

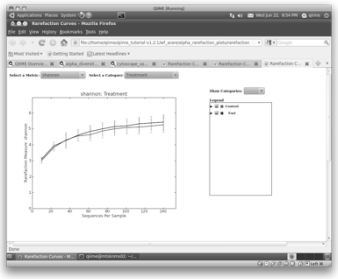
Rarefaction:
- Normalization to equal number of sequences – shows the trends and if the depth is adequate

(some) Diversity metrics:
- chao1: $S_{obs} + n_1(n_1\text{-}1)/(2n_2\text{-}2)$
- dominance (probability of randomly drawing two individuals of the same species): $\Sigma\ (S_i(S_i\text{-}1))/N(N\text{-}1)$
- fisher_alpha: solution to $S=a*\ln(1+n/a)$, where S is number of species
- observed_species: total number of unique species (richness)
- shannon: (evenness) $-\Sigma\ p_i \log(p_i)$
- simpson: 1 - dominance
- singles: number of species represented by a single individual
- PD_whole_tree: sum of branch lengths between all representatives

23

Alpha rarefaction results

```
> firefox wf_arare/alpha_rarefaction_plots/rarefaction_plots.html
```



24

Beta diversity

- Compares the diversity (similarity) between two samples
- e.g. Sørensen's similarity index $\beta = \frac{2C}{S_1 + S_2}$
- Corresponding dissimilarity (distance) is Bray-Curtis distance D = 1-β
- Unifrac is another popular diversity metric defined as the sum of branch lengths of phylogenetic trees that are in common between two samples:

Lozupone and Knight, UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Appl Environ Microbiol. 2005 December; 71(12): 8228–8235.

25

Beta diversity & PCA analysis

- beta_diversity_through_3d_plots.py …
- make_2d_plots.py …
- make_distance_histograms.py …

26

Beta diversity & PCA output

- `wf_bdiv_even146/`
  - `Distance_Histograms/`
  - PCoA plots
    - `unweighted_unifrac_3d_continuous/`
    - `weighted_unifrac_3d_continuous/`
    - `unweighted_unifrac_3d_discrete/`
    - `weighted_unifrac_3d_discrete/`
  - eigenvectors
    - `weighted_unifrac_pc.txt`
    - `unweighted_unifrac_pc.txt`
  - Unifrac distances
    - `weighted_unifrac_seqs_otu_table_even146.txt`
    - `unweighted_unifrac_seqs_otu_table_even146.txt`

27