Installing QIIME: Full install, virtual & cloud machines

Antonio González Peña
University of Colorado at Boulder
Knight Lab
Why we care?

- Microbes are ubiquitous

Image by Micah Hamady
Why we need QIIME?

Split libraries, pick OTU's, chimera-check, align, phylogenetic tree, taxonomic classification, diversity analyses, multivariate statistics, plots
Why we need QIIME?

Align, Chimera-check:
Upload to greengenes.lbl.gov server, 500 at a time
pick OTU's:
Dotur/Mothur through terminal on personal machine
phylogenetic tree:
ARB, upload to RAxML server (1 week for 2500 seqs)
taxonomic classification:
greengenes.lbl.gov server, 500 at a time or ARB (count by hand)
diversity analyses
DOTUR/MOTHUR, upload to UniFrac server distance matrices
upload to UniFrac server, vegan package in R, PrimerE
multivariate statistics, plots
vegan and MASS packages in R, MATLAB, PrimerE
QIIME

Extract DNA and PCR amplify with barcoded primer

Pool amplicons

Pyrosequence amplicons using 454's GS FLX instrument

Screen, assign sequences to samples using barcode

Cluster samples based on UniFrac distances

Fast UniFrac

tens of thousands of sequences, 100's of samples

Align to reference alignment (e.g., using NAST), infer phylogeny (e.g., using FastTree)

QIIME

- Demultiplex samples

Sequencing output
(454, Illumina, Sanger)
- fasta, qual, a/sf/trace files

Metadata
mapping file

Pre-processing
- e.g., remove primer(s), demultiplex, quality filter

Denoise 454 Data
PyroNoise, Denoiser

SRA Submission
Submit sequences and metadata to SRA

Pick OTUs and representative sequences

Reference based
BLAST, UCLUST

De novo
- e.g., UCLUST, CD-HIT, MOTHUR

Assign taxonomy
- BLAST, RDP Classifier

Align sequences
- e.g., PyNAST, INFERNAL, MUSCLE, MAFFT

Build OTU table
- i.e., per sample OTU counts

Build phylogenetic tree
- e.g., FastTree, RAxML, ClearCut

OTU Table
(i.e., per sample OTU counts)

Phylogenetic Tree
Evolutionary relationship between OTUs

β-diversity and rarefaction
- e.g., Weighted and unweighted UniFrac, Bray-Curtis, Jaccard

α-diversity and rarefaction
- e.g., Phylogenetic Diversity, Chao1, Observed Species

Visualization
- e.g., 2D and 3D PCoA plots, distance histograms, taxonomy pie charts, rarefaction plots, OTU network visualization, jackknifed hierarchical clustering.
QIIME

• Creating an OTU table

Sequencing output
(454, Illumina, Sanger)
fasta, qual, sff/trace files

Metadata
mapping file

Pre-processing
e.g., remove primer(s), demultiplex,
quality filter

Denoise 454 Data
PyroNoise, Denoiser

SRA Submission
Submit sequences and
metadata to SRA

Pick OTUs and representative sequences
Reference based
BLAST, UCLUST
De novo
e.g., UCLUST, CD-HIT,
MOTHUR

Assign taxonomy
BLAST, RDP
Classifier

Align sequences
e.g., PyNAST,
INFERNAL, MUSCLE,
MAFFT

Build ‘OTU table’
i.e., per sample OTU
counts

OTU Table
(i.e., per sample OTU
counts)

Phylogenetic Tree
Evolutionary relationship
between OTUs

α-diversity and rarefaction
e.g., Phylogenetic
Diversity, Chao1,
Observed Species

β-diversity and rarefaction
e.g., Weighted and
unweighted UniFrac, Bray-
Curtis, Jaccard

Visualization
e.g., 2D and 3D PCoA plots, distance histograms, taxonomy pie
charts, rarefaction plots, OTU network visualization, jackknifed
hierarchical clustering.
QIIME

• Alpha / Beta Diversity

Sequencing output
(454, Illumina, Sanger)
fasta, qual, sff, trace files

Metadata
mapping file

Pre-processing
e.g., remove primer(s), demultiplex, quality filter

Denoise 454 Data
PyroNoise, Denoiser

SRA Submission
Submit sequences and metadata to SRA

Pick OTUs and representative sequences
Reference based
BLAST, UCLUST
De novo
e.g., UCLUST, CD-HIT, MOTHUR

Assign taxonomy
BLAST, RDP Classifier

Align sequences
e.g., PyNAST, INFERNAL, MUSCLE, MAFFT

Build 'OTU table'
i.e., per sample OTU counts

Build phylogenetic tree
e.g., FastTree, RAxML, ClearCut

OTU Table
(i.e., per sample OTU counts)

Phylogenetic Tree
Evolutionary relationship between OTUs

α-diversity and rarefaction
e.g., Phylogenetic Diversity, Chao1, Observed Species

β-diversity and rarefaction
e.g., Weighted and unweighted UniFrac, Bray-Curtis, Jaccard

Visualization
e.g., 2D and 3D PCoA plots, distance histograms, taxonomy pie charts, rarefaction plots, OTU network visualization, jackknifed hierarchical clustering.
QIIME

• Visualizations

- Sequencing output (454, Illumina, Sanger):
  - fasta, qual, sff/trace files
- Metadata:
  - mapping file

  Pre-processing
  - e.g., remove primer(s), demultiplex, quality filter

  Denoise 454 Data
  - PyroNoise, Denoiser

  SRA Submission
  - Submit sequences and metadata to SRA

  Pick OTUs and representative sequences
  - Reference based
    - BLAST, UCLUST
  - De novo
    - e.g., UCLUST, CD-HIT, MOTHUR

  Assign taxonomy
  - BLAST, RDP Classifier

  Align sequences
  - e.g., PyNAST, INFERNAL, MUSCLE, MAFFT

  Build 'OTU table'
  - i.e., per sample OTU counts

  Build phylogenetic tree
  - e.g., FastTree, RAxML, ClearCut

- OTU Table (i.e., per sample OTU counts)
  - α-diversity and rarefaction
    - e.g., Phylogenetic Diversity, Chao1, Observed Species
  - β-diversity and rarefaction
    - e.g., Weighted and unweighted UniFrac, Bray-Curtis, Jaccard

- Phylogenetic Tree
  - Evolutionary relationship between OTUs

  Visualization
  - e.g., 2D and 3D PCoA plots, distance histograms, taxonomy pie charts, rarefaction plots, OTU network visualization, jackknifed hierarchical clustering.
What do I need to run QIIME?

1. Python-2.6.4
2. Qiime-1.3.0
3. setuptools-0.6c11
4. MySQL-python-1.2.3
5. SQLAlchemy-0.7.1
6. PyCogent-1.5.1
7. PyNAST-1.1
8. numpy-1.3.0
9. matplotlib-0.98.5.3
10. mpi4py-1.2.2
11. lxml-2.2.7
12. Sphinx-1.0.4
13. RAxML-7.0.3.tar.bz2
14. FastTree-2.1.3.c
15. cdbfasta

16. cd-hit-2007-0131
17. rdp_classifier_2.2
18. blast-2.2.22-x64-linux
19. muscle3.6_linux_ia32
20. infernal-1.0.2
21. cytoscapeSource-v2.7.0
22. Clearcut.source
23. Mothur.1.6.0
24. uclustq1.2.22_i86linux64
25. R-2.12.0
26. AmpliconNoiseV1.21
27. ViennaRNA-1.8.4
28. pprospector-1.0.1
29. microbiomeutil_2010-04-29
QIIME installs

• Install every single dependency by hand
• Use the app-deploy.py script
• Download the Virtual Machine QIIME hard drive
• Use the latest EC2 QIIME release
QIIME installs

✗ Install every single dependency by hand
  • Use the app-deploy.py script
  • Download the Virtual Machine QIIME hard drive
  • Use the latest EC2 QIIME release
app-deploy.py

- Only works in Linux 64bit machines
- `python app-deploy.py [install dir] -f [conf file]`
- Multi-process install
  
  - `wget http://bmf.colorado.edu/QIIME/app-deploy-qiiime-1.3.0.tgz`
  - `tar zxvf app-deploy-qiiime-1.3.0.tgz`
  - `cd app-deploy-qiiime-1.3.0`
  - `python app-deploy.py /software/ -f etc/qiime_1.3.0.conf`
  - `exit`
app-deploy.py

- Only works in Linux 64bit machines
- `python app-deploy.py [install dir] -f [conf file]
- Multi-process install

  - `wget http://bmf.colorado.edu/QIIME/app-deploy-qiime-1.3.0.tgz`
  - `tar zxvf app-deploy-qiime-1.3.0.tgz`
  - `cd app-deploy-qiime-1.3.0`
  - `python app-deploy.py /software/ -f etc/qiime_1.3.0.conf`
  - `exit`
INFO Launched initial deploy,
INFO Deploying app: qiime
INFO Making deploy directory:
INFO Downloading qiime
INFO Building qiime
INFO Installing qiime
INFO Successfully built qiime.
INFO Deploy of qiime
INFO Generating new ~/.qiime_config
INFO Backing up /var/lib/hudson/.qiime_config
INFO Generated new ~/.qiime_config
Virtual Machines
Virtualization Software Packages

- VirtualBox
- Windows Virtual PC
- Xen
- VMware
- Many, many more ...
QIIME Virtual Machine

- VirtualBox
- QIIME virtual hard drive

- http://qiime.org/install/virtual_box.html
- http://videos.qiime.org/
Cloud Computing

Image from: http://www.techwarelabs.com/articles/editorials/virtual_double/
Cloud computing options

• Amazon Elastic Compute Cloud (EC2)
• GoGrid

• Magellan – Argonne's DOE Cloud Computing
• Data Intensive Academic Grid (DIAG) – Institute for Genome Sciences (IGS), University of Maryland School of Medicine (UMSOM)
EC2 demo

http://aws.amazon.com
### EC2 – Region $$

#### Instance

<table>
<thead>
<tr>
<th>Region: US East (Virginia)</th>
<th>Linux/UNIX Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard On-Demand Instances</strong></td>
<td></td>
</tr>
<tr>
<td>Small (Default)</td>
<td>$0.085 per hour</td>
</tr>
<tr>
<td>Large</td>
<td>$0.34 per hour</td>
</tr>
<tr>
<td>Extra Large</td>
<td>$0.68 per hour</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region: Asia Pacific (Tokyo)</th>
<th>Linux/UNIX Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard On-Demand Instances</strong></td>
<td></td>
</tr>
<tr>
<td>Small (Default)</td>
<td>$0.10 per hour</td>
</tr>
<tr>
<td>Large</td>
<td>$0.40 per hour</td>
</tr>
<tr>
<td>Extra Large</td>
<td>$0.80 per hour</td>
</tr>
</tbody>
</table>
### EC2 – Region $$

#### Instance

<table>
<thead>
<tr>
<th>Region</th>
<th>Linux/UNIX Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>US East (Virginia)</strong></td>
<td><strong>Standard On-Demand Instances</strong></td>
</tr>
<tr>
<td>Small (Default)</td>
<td>$0.085 per hour</td>
</tr>
<tr>
<td>Large</td>
<td>$0.34 per hour</td>
</tr>
<tr>
<td>Extra Large</td>
<td>$0.68 per hour</td>
</tr>
</tbody>
</table>

#### Spot Instance

<table>
<thead>
<tr>
<th>Region</th>
<th>Linux/UNIX Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>US East (Virginia)</strong></td>
<td><strong>Standard Spot Instances</strong></td>
</tr>
<tr>
<td>Small (Default)</td>
<td>$0.03 per hour</td>
</tr>
<tr>
<td>Large</td>
<td>$0.12 per hour</td>
</tr>
<tr>
<td>Extra Large</td>
<td>$0.24 per hour</td>
</tr>
<tr>
<td>Instance Type</td>
<td>Linux/UNIX Usage</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td><strong>Standard On-Demand Instances</strong></td>
<td></td>
</tr>
<tr>
<td>Small (Default)</td>
<td>$0.085 per hour</td>
</tr>
<tr>
<td>Large</td>
<td>$0.34 per hour</td>
</tr>
<tr>
<td>Extra Large</td>
<td>$0.68 per hour</td>
</tr>
<tr>
<td><strong>Micro On-Demand Instances</strong></td>
<td></td>
</tr>
<tr>
<td>Micro</td>
<td>$0.02 per hour</td>
</tr>
<tr>
<td><strong>Hi-Memory On-Demand Instances</strong></td>
<td></td>
</tr>
<tr>
<td>Extra Large</td>
<td>$0.50 per hour</td>
</tr>
<tr>
<td>Double Extra Large</td>
<td>$1.00 per hour</td>
</tr>
<tr>
<td>Quadruple Extra Large</td>
<td>$2.00 per hour</td>
</tr>
<tr>
<td><strong>Hi-CPU On-Demand Instances</strong></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>$0.17 per hour</td>
</tr>
<tr>
<td>Extra Large</td>
<td>$0.68 per hour</td>
</tr>
<tr>
<td><strong>Cluster Compute Instances</strong></td>
<td></td>
</tr>
<tr>
<td>Quadruple Extra Large</td>
<td>$1.60 per hour</td>
</tr>
<tr>
<td><strong>Cluster GPU Instances</strong></td>
<td></td>
</tr>
<tr>
<td>Quadruple Extra Large</td>
<td>$2.10 per hour</td>
</tr>
</tbody>
</table>
AWS in Education (Grants)

- http://aws.amazon.com/education/
QIIME

• The code is tested (properly)
• The documentation is updated constantly based on users suggestions
• The help in the QIIME-forum has a collaborative spirit (developers & users sharing their research experiences)
Thanks

http://qiime.org

http://forum.qiime.org

http://videos.qiime.org

Antonio Gonzalez Peña
antgonza@gmail.com