

HIGH PERFORMANCE RESEARCH COMPUTING

ACES: NGS Metagenomics

HPRC Training
24 October 2023

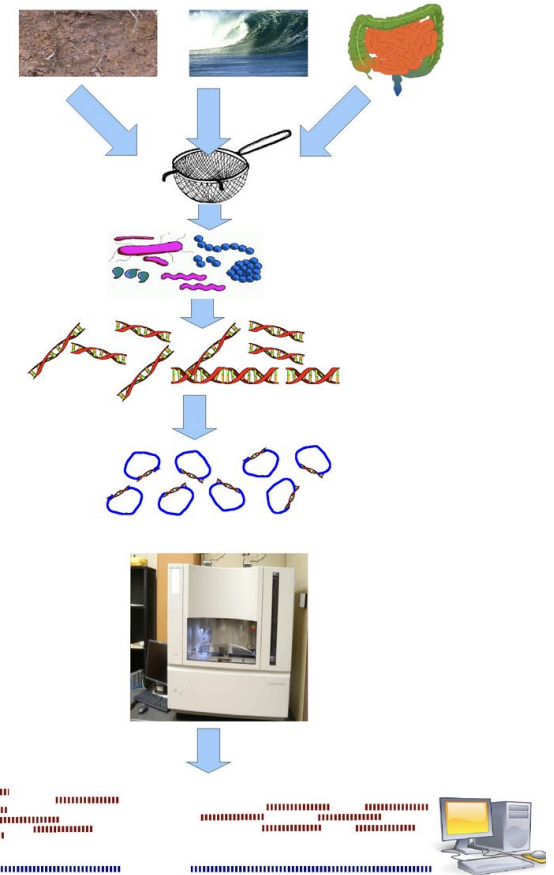


High Performance
Research Computing
DIVISION OF RESEARCH



Introduction to Metagenomics

- Sequencing of communities of microorganisms
- No need for isolation and lab cultivation
- High depth short-read sequencing (Next Generation Sequencing - NGS)
- Long-read (third generation) sequencing
 - PacBio
 - Oxford Nanopore Technologies



Wooley et al. 2010 - <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000667>

Sequencing Strategies

- Whole genome sequencing (WGS)
- Marker gene
 - 16S Ribosomal RNA (rRNA)
 - Bacteria
 - Archaea
 - 18S Ribosomal RNA
 - Fungus
 - Eukaryotes



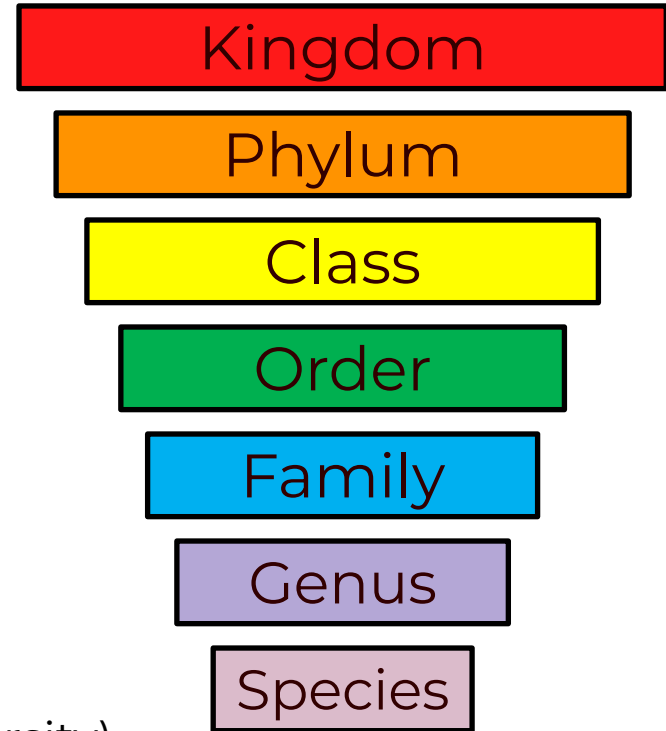
<https://www.illumina.com/systems/sequencing-platforms/miseq.html>

WGS Metagenomics

- Sequencing whole genomes of the microorganisms present in the sample
- Facilitates discovering gene functions and genome structures
- Steps involved:
 - Genome assembly (special software/considerations)
 - Binning
 - Predicting and annotating genes

Marker Gene Metagenomics

- Usually based on 16S rRNA
 - Conserved within species
 - Varies greatly between species
 - Widely used for microbial ecology - many resources available
- Needs a reference database to match the Operational Taxonomic Units (OTUs)
 - Silva
 - Greengenes
- Steps:
 - Preprocessing (removing noise, QC)
 - OTU clustering and taxonomic assignment
 - Alpha diversity analysis (within sample diversity)
 - Beta diversity analysis (between sample diversity)



Sequencing Platforms

- Illumina
 - Short reads (up to 300 bp)
 - Highly accurate
 - High depth sequencing
- PacBio
 - Long reads (10-25kb)
 - Accurate (99.5%)
- Oxford Nanopore
 - Long reads (10-30 kb)
 - Less accurate (~95%)
 - Portable
 - Affordable



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

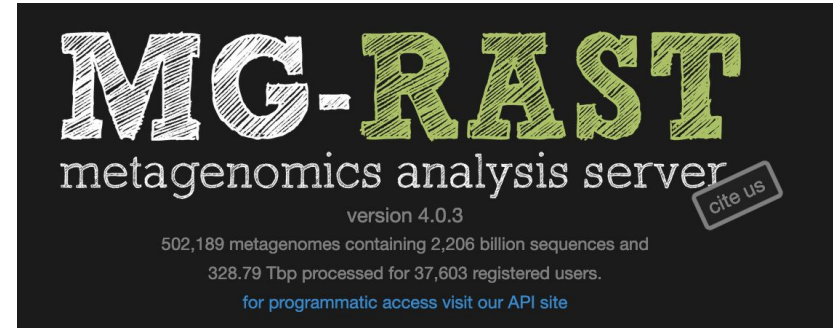
<https://nanoporetech.com>

Metagenomics Software

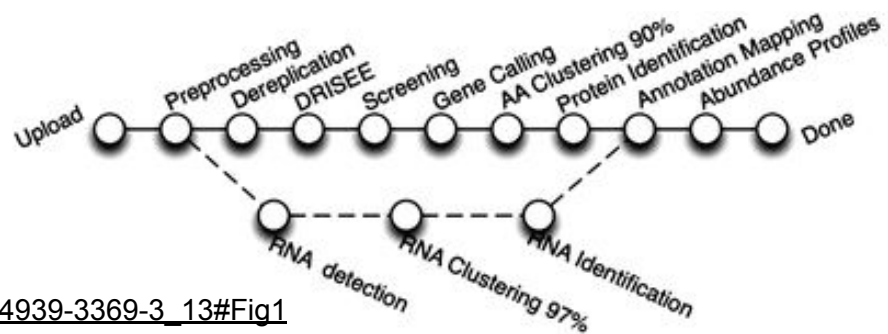
Commonly used software suites and pipelines

MG-RAST metagenomics analysis server

- Open-source web application for metagenomic analysis
- Large repository for metagenomic data
- Hosted by The University of Chicago and Argonne National Laboratory
- Full analysis pipeline



<https://www.mg-rast.org>



Keegan et al. 2016 -

https://link.springer.com/protocol/10.1007/978-1-4939-3369-3_13#Fig1

Bacterial and Viral Bioinformatics Resource Center

- Formerly known as PATRIC
- Designed to support research on bacterial and viral diseases
- Among other tools, provides some metagenomic functions:
 - Metagenomic read mapping
 - Taxonomic classification
 - Metagenomic binning



<https://www.bv-brc.org>

Mothur

- Open-source software for microbial ecology
- Single piece of software - many functions/commands
- Example data and protocols available
- Contains accelerated versions of the DOTUR and SONS programs
- Highly cited (> 17,700 citations)
- Schloss et al. 2009:
<https://journals.asm.org/doi/10.1128/aem.01541-09>



<https://mothur.org>

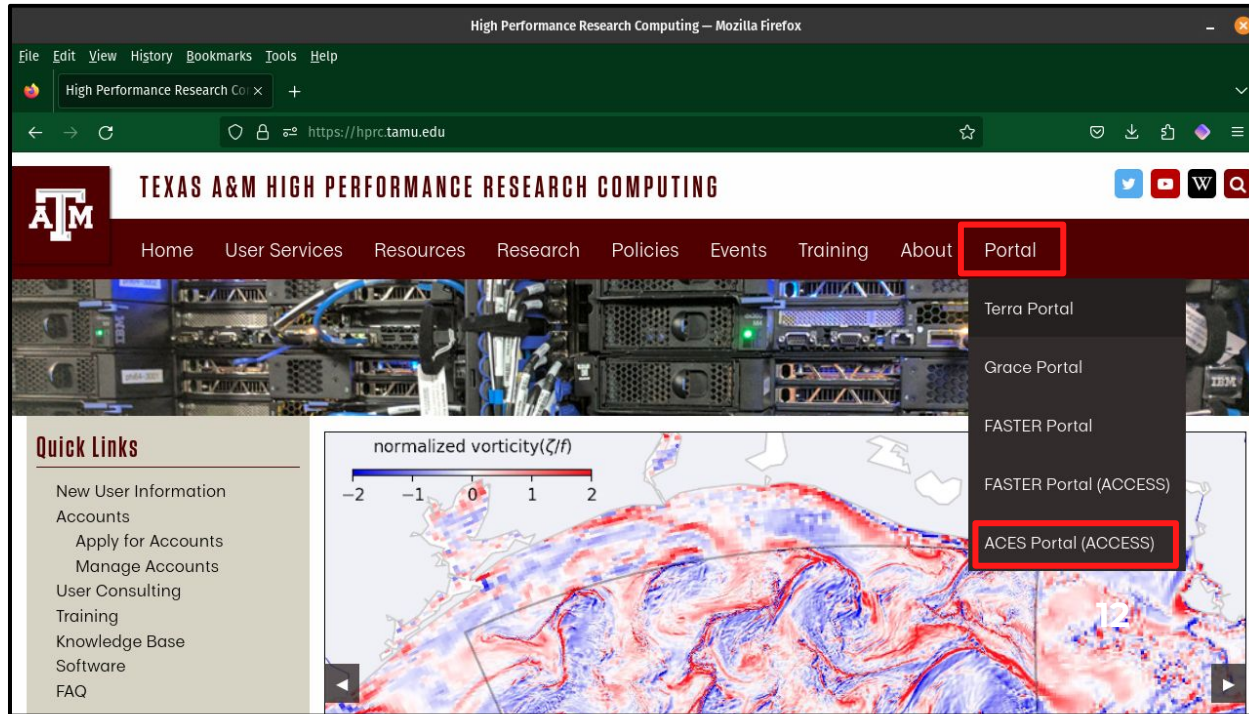
QIIME 2

- Open-source pipeline for microbiome analysis
- Takes raw fastq data (multiplexed or demultiplexed)
- From demultiplexing to publication-ready figures
- Incorporates many other software packages (e.g. Mothur, FastTree)



<http://qiime.org>

Accessing the HPRC ACES Portal



The screenshot shows a Mozilla Firefox browser window displaying the Texas A&M High Performance Research Computing (HPRC) website. The browser's address bar shows the URL <https://hprc.tamu.edu>. The website header includes the Texas A&M logo and the text "TEXAS A&M HIGH PERFORMANCE RESEARCH COMPUTING". A navigation menu is visible with the following items: Home, User Services, Resources, Research, Policies, Events, Training, About, and Portal. The "Portal" item is highlighted with a red box. A dropdown menu is open from the "Portal" item, listing the following options: Terra Portal, Grace Portal, FASTER Portal, FASTER Portal (ACCESS), and ACES Portal (ACCESS). The "ACES Portal (ACCESS)" option is highlighted with a red box. Below the navigation menu, there is a banner image of server racks and a "Quick Links" section with the following items: New User Information, Accounts, Apply for Accounts, Manage Accounts, User Consulting, Training, Knowledge Base, Software, and FAQ. A map showing normalized vorticity (ζ/f) is also visible, with a color scale ranging from -2 to 2.

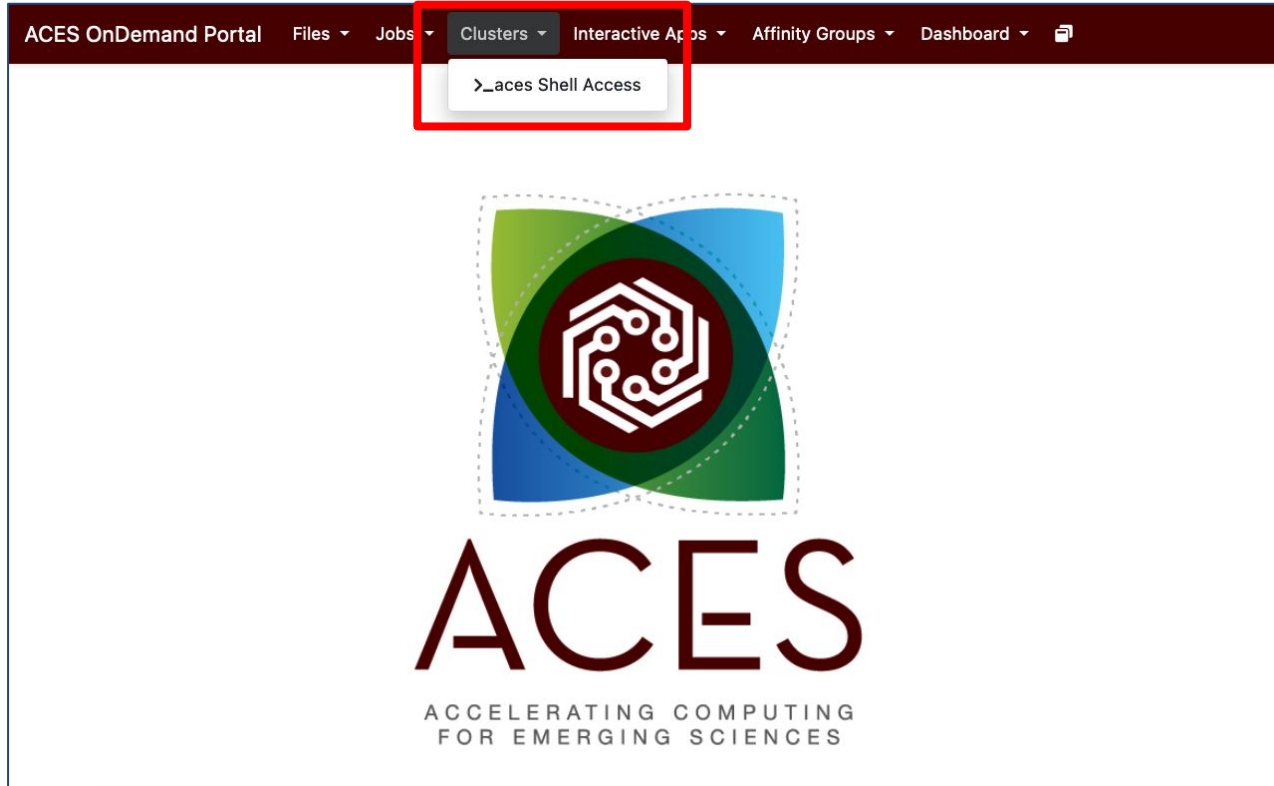
HPRC webpage: <https://hprc.tamu.edu>

Accessing ACES via the Portal (ACCESS)

Log-in using your ACCESS credentials.

Select the Identity Provider appropriate for your account.

Accessing ACES shell in OOD Portal



The screenshot displays the ACES OnDemand Portal interface. The top navigation bar includes links for Files, Jobs, Clusters, Interactive Apps, Affinity Groups, and Dashboard. A red box highlights the Clusters dropdown menu, which contains a link labeled '>_aces Shell Access'. Below the navigation bar is the ACES logo, which consists of a stylized circuit board design within a square frame, and the text 'ACES' in large letters, with the tagline 'ACCELERATING COMPUTING FOR EMERGING SCIENCES' underneath.

Getting started with QIIME 2 - Terminology

- Artifacts
 - Contain data and metadata
 - .qza file extension
 - Allows QIIME to track type, format, and provenance of the data
- Visualization
 - Terminal output of an analysis (e.g. tables, graphs)
 - Also contain data and metadata
 - Can be viewed at <https://view.qiime2.org>

Getting started with QIIME 2 - Terminology

- Semantic types
 - Essentially the classification of an artifact
 - Helps users avoid using incorrect inputs for analyses

Common semantic types

Unless otherwise noted the following semantic types are defined by, and importable from, the `q2-types` plugin. It is also possible to define semantic types in any plugin, so the available semantic types are not limited to those defined in `q2-types`. Instructions will be added soon for how to accomplish this. In the meantime, you can refer to the `q2-dummy-types` repository for annotated examples.

`FeatureTable[Frequency]` : A feature table (e.g., samples by OTUs) where each value indicates the frequency of an OTU in the corresponding sample expressed as raw counts.

`FeatureTable[RelativeFrequency]` : A feature table (e.g., samples by OTUs) where each value indicates the relative abundance of an OTU in the corresponding sample such that the values for each sample will sum to 1.0.

`FeatureTable[PresenceAbsence]` : a feature table (e.g., samples by OTUs) where each value indicates whether an OTU is present or absent in the corresponding sample.

`FeatureTable[Composition]` : A feature table (e.g., samples by OTUs) where each value indicates the frequency of an OTU in the corresponding sample, and all frequencies are greater than zero.

`PhyLogeny[Rooted]` : A rooted phylogenetic tree.

`PhyLogeny[Unrooted]` : An unrooted phylogenetic tree.

`DistanceMatrix` : A distance matrix.

<https://docs.qiime2.org/2022.2/semantic-types>

Getting started with QIIME 2 - Terminology

- Plugins
 - Software packages that perform specific analyses (e.g. q2-demux, q2-diversity)
 - Can be written by third-party developers
 - Plugins available for all steps necessary for complete pipeline

Available plugins

QIIME 2 microbiome analysis functionality is made available to users via plugins. The following official plugins are currently included in QIIME 2 train releases:

- alignment: Plugin for generating and manipulating alignments.
 - Methods
 - mafft: De novo multiple sequence alignment with MAFFT
 - mafft-add: Add sequences to multiple sequence alignment with MAFFT.
 - mask: Positional conservation and gap filtering.
- composition: Plugin for compositional data analysis.
 - Methods
 - add-pseudocount: Add pseudocount to table
 - Visualizers
 - ancom: Apply ANCOM to identify features that differ in abundance.

<https://docs.qiime2.org/2022.2/plugins/available>

Example data

MOLECULAR ECOLOGY

Special Issue: Nature's Microbiome |  Open Access

Convergence of gut microbiomes in myrmecophagous mammals

Frédéric Delsuc, Jessica L. Metcalf, Laura Wegener Parfrey, Se Jin Song, Antonio González, Rob Knight 

First published: 29 August 2013 | <https://doi.org/10.1111/mec.12501> | Citations: 28

<https://onlinelibrary.wiley.com/doi/10.1111/mec.12501>

Getting started with QIIME 2 on ACES

Set up the environment and working directory:

```
$ module purge
```

```
$ module load QIIME2/2022.11
```

```
$ mkdir $SCRATCH/metagenomics
```

```
$ cd $SCRATCH/metagenomics
```

```
$ cp -r /scratch/training/metagenomics/* .
```

Import fastq reads to QIIME artifact

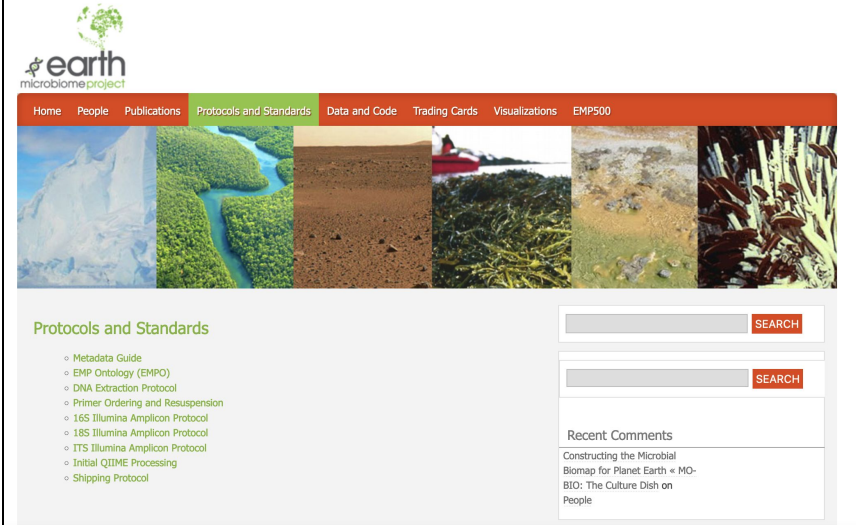
```
$ qiime tools import \  
  --type EMPSSingleEndSequences \  
  --input-path fastqs \  
  --output-path reads.qza
```

Importing Data

- QIIME 2 can import many types of data:
 - Fastq (single and paired-end)
 - Fasta
 - Feature tables
 - Phylogenetic trees
- Importing data creates QIIME 2 artifacts (with specific semantic types)
- Semantic types for raw fastq files:
 - EMPSingleEndSequences
 - EMPPairedEndSequences
 - MultiplexedSingleEndBarcodeInSequence
 - MultiplexedPairedEndBarcodeInSequence ...

Importing Data

- Example data is in the EMP single end format
- Data is still multiplexed (single fastq file)
- Directory with two fastq files
 - Sequences
 - Barcodes



The screenshot shows the Earth Microbiome Project website. At the top left is the logo with a tree icon and the text "earth microbiome project". A navigation bar contains the following tabs: Home, People, Publications, Protocols and Standards (highlighted in green), Data and Code, Trading Cards, Visualizations, and EMP500. Below the navigation bar is a row of six images: a desert landscape, a mangrove forest, a soil cross-section, a boat on water, a microbial culture plate, and a close-up of plant roots. The main content area is titled "Protocols and Standards" and lists several items with small tree icons:

- Metadata Guide
- EMP Ontology (EMPO)
- DNA Extraction Protocol
- Primer Ordering and Resuspension
- 16S Illumina Amplicon Protocol
- 16S Illumina Amplicon Protocol
- ITS Illumina Amplicon Protocol
- Initial QIIME Processing
- Shipping Protocol

On the right side of the page, there are two search boxes, each with a "SEARCH" button. Below them is a "Recent Comments" section with the following text:

Recent Comments
Constructing the Microbial
Biomap for Planet Earth ◀ MO-
BIO: The Culture Dish on
People

Demultiplexing the example data

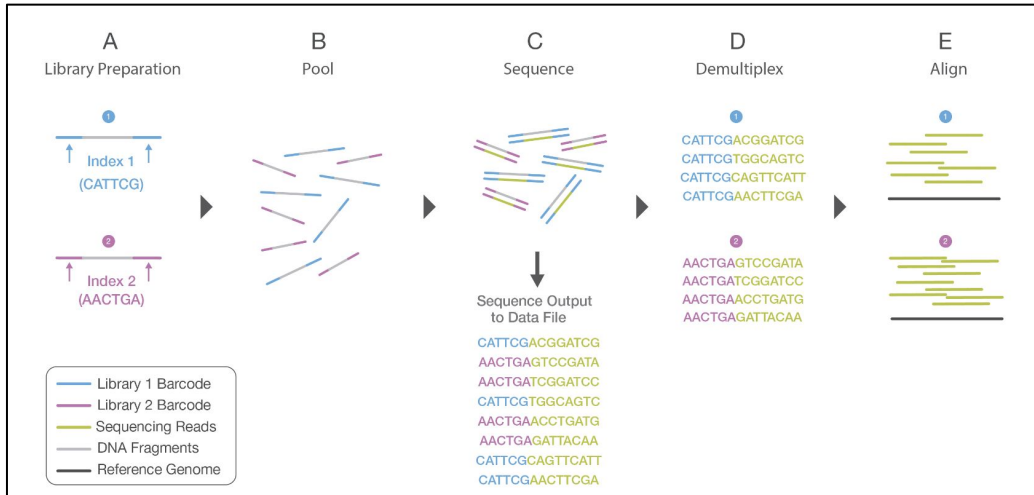
```
$ qiime demux emp-single --i-seqs reads.qza \  
  --m-barcodes-file myrme-sample-data.txt \  
  --m-barcodes-column barcode-sequence \  
  --o-per-sample-sequences demux.qza \  
  --o-error-correction-details demux-details.qza \  
  --p-rev-comp-mapping-barcodes
```

```
$ qiime demux summarize --i-data demux.qza \  
  --o-visualization demux.qzv
```

Upload and view demux.qzv file here: <https://view.qiime2.org>

Demultiplexing and Quality Control

- Example data (and most amplicon/targeted metagenomics datasets) are pooled for sequencing
- Unique barcodes (short oligos) applied to each sample
- Barcodes used to sort reads after sequencing



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Preprocessing Sequence Data

- Use the following commands to denoise and filter the example data
- Remember the interactive quality plot to choose the right values for the trimming options

```
$ qiime dada2 denoise-single \  
  --i-demultiplexed-seqs demux.qza \  
  --p-trim-left 8 --p-trunc-len 148 \  
  --o-representative-sequences rep-seqs.qza \  
  --o-table table.qza --o-denoising-stats stats.qza \  
  --p-n-threads 2
```

Denoising and Filtering

- Identify and correct sequenced amplicons
- Filter chimeric reads
- Filter PhiX reads
- Multiple options for denoising and filtering in QIIME 2
 - DADA2
 - Deblur

Preprocessing Sequence Data

- Generate feature/sequence tables for visualization:

```
$ qiime metadata tabulate --m-input-file stats.qza \  
  --o-visualization stats.qzv
```

```
$ qiime feature-table summarize --i-table table.qza \  
  --o-visualization table.qzv \  
  --m-sample-metadata-file myrme-sample-data.txt
```

```
$ qiime feature-table tabulate-seqs \  
  --i-data rep-seqs.qza --o-visualization rep-seqs.qzv
```

- Download and view the qzv files at <https://view.qiime2.org>

Diversity Analysis

- Alpha Diversity - within sample
 - Shannon's Diversity Index
 - Observed Features
 - Faith's Phylogenetic Diversity
 - Evenness
- Beta Diversity - between samples
 - Jaccard Distance
 - Bray-Curtis Distance
 - Unweighted UniFrac Distance
 - Weighted UniFrac Distance

Diversity Analysis

- Use the following command to generate a phylogenetic tree for the diversity analyses

```
$ qiime phylogeny align-to-tree-mafft-fasttree \  
  --i-sequences rep-seqs.qza \  
  --o-alignment aligned-rep-seqs.qza \  
  --o-masked-alignment masked-aligned-reps-seqs.qza \  
  --o-tree unrooted-tree.qza \  
  --o-rooted-tree rooted-tree.qza
```

Diversity Analysis

- Use the following command to generate the alpha and beta diversity metrics, as well as the PCA plots for the beta diversity metrics

```
$ qiime diversity core-metrics-phylogenetic \  
  --i-phylogeny rooted-tree.qza --i-table table.qza \  
  --p-sampling-depth 10590 \  
  --m-metadata-file myrme-sample-data.txt \  
  --output-dir diversity-analysis
```

- Upload and view the beta diversity qzv files at <https://view.qiime2.org>

Diversity Analysis

- Use the following commands to test for significance of alpha-level diversity

```
$ qiime diversity alpha-group-significance \  
  --i-alpha-diversity diversity-analysis/faith_pd_vector.qza \  
  --m-metadata-file myrme-sample-data.txt \  
  --o-visualization diversity-analysis/faith_pd_vector-sig.qzv
```

```
$ qiime diversity alpha-group-significance \  
  --i-alpha-diversity diversity-analysis/evenness_vector.qza \  
  --m-metadata-file myrme-sample-data.txt \  
  --o-visualization diversity-analysis/evenness_vector-sig.qzv
```

- Upload and view the alpha diversity qzv files at <https://view.qiime2.org>

Diversity Analysis

- Use the following commands to test for significance of beta-level diversity

```
$ qiime diversity beta-group-significance \  
  --i-distance-matrix diversity-analysis/unweighted_unifrac_distance_matrix.qza \  
  --m-metadata-file myrme-sample-data.txt --m-metadata-column diet \  
  --o-visualization diversity-analysis/unweighted_unifrac_diet-significance.qzv \  
  --p-pairwise
```

```
$ qiime feature-table filter-samples --i-table table.qza \  
  --m-metadata-file myrme-sample-data.txt \  
  --p-where "[species]!='panda' AND [species]!='pink-fairy-armadillo'" \  
  --o-filtered-table only-multi-species.qza
```

```
$ qiime diversity core-metrics-phylogenetic --i-phylogeny rooted-tree.qza \  
  --i-table only-multi-species.qza --p-sampling-depth 10590 \  
  --m-metadata-file myrme-sample-data.txt \  
  --output-dir diversity-analysis-only-multi-species
```

Diversity Analysis

- Use the following commands to test for significance of beta-level diversity

```
$ qiime diversity beta-group-significance \  
  --i-distance-matrix \  
  diversity-analysis-only-multi-species/unweighted_unifrac_distance_matrix.qza \  
  --m-metadata-file myrme-sample-data.txt --m-metadata-column species \  
  --o-visualization \  
  diversity-analysis-only-multi-species/unweighted_unifrac_species-sig.qzv \  
  --p-pairwise
```

- Upload and view the qzv file at <https://view.qiime2.org>

Thank You!

Need Help? Contact the HPRC Helpdesk

Website: hprc.tamu.edu

Email: help@hprc.tamu.edu

Phone: (979) 845-0219

Help us help you -- we need more info

- Which Cluster (ACES, FASTER, Terra, Grace)
- NetID (NOT your UIN)
- Job id(s) if any
- Location of your jobfile, input/output files
- Application used, if any
- Module(s) loaded, if any
- Error messages
- Steps you have taken, so we can reproduce the problem

35

