16SrRNA Intermediate Bioinformatics Online Course: Int_BT_2019

Module 4

16S rRNA Sequencing Bioinformatics Pipelines: the theory

Learning Objectives

1. To understand various 16S rRNA Sequencing platforms

2. Understand how to assess the quality of raw sequences

3. To know various Bioinformatics Pipelines used to process 16S rRNA Microbiome data

Learning Outcomes

1. Use Appropriate Microbiome Sequencing Tools

2. Acquire Knowledge to Assess Quality of Raw Sequence Reads

3. Knowledge on the Available 16S rRNA Bioinformatics Processing Pipelines

Outline

1. Background on 16S rRNA gene

2. Sequencing Tools

3. Quality Assessment of Raw Sequence reads

4. 16S rRNA Bioinformatic processing Pipeline

      i.   UCT-CBIO microbiome processing pipeline

      ii.   QIIME2 microbiome processing pipeline

      iii.   MOTHUR microbiome processing pipeline

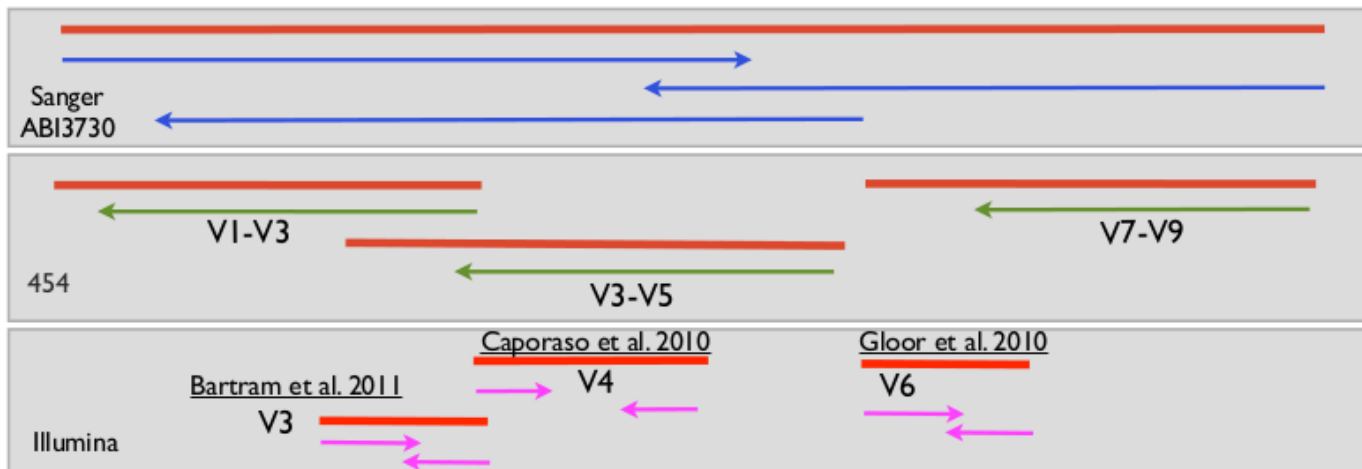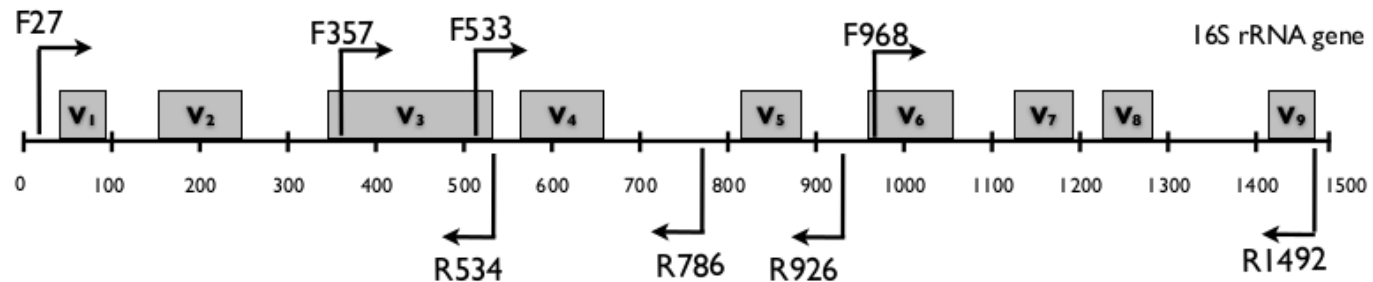   iv.   DADA2 microbiome processing pipeline

Module 4 : Part II

Background on 16S rRNA gene

&

Sequencing Platforms

Background on 16S rRNA

- The 16S rRNA is a gene which codes for the RNA component of the small subunit of ribosomes, and it is among the **most conserved** gene across all kingdoms of life.

- They contain **regions** that are less evolutionarily constrained and whose sequences are indicative of their phylogeny.

- **Amplification** of these genomic regions by PCR and subsequent **sequencing** of a sufficiently large number of individual **amplicons** enables the analysis of the **diversity** of organism in the sample and a rough estimate of their **relative abundance.**

H3ABioNet
Pan African Bioinformatics Network for H3Africa

16S rRNA gene region

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Evolution of Sequencing platforms

H3ABioNet
Pan African Bioinformatics Network for H3Africa

1. Sanger Sequencing

- Good quality

- Long reads, up to 1000 bp

- By comparison very little data
  - Up to 384 sequences read in parallel

- Expensive per base pair

- Developed by Frederick Sanger, 1977

H3ABioNet
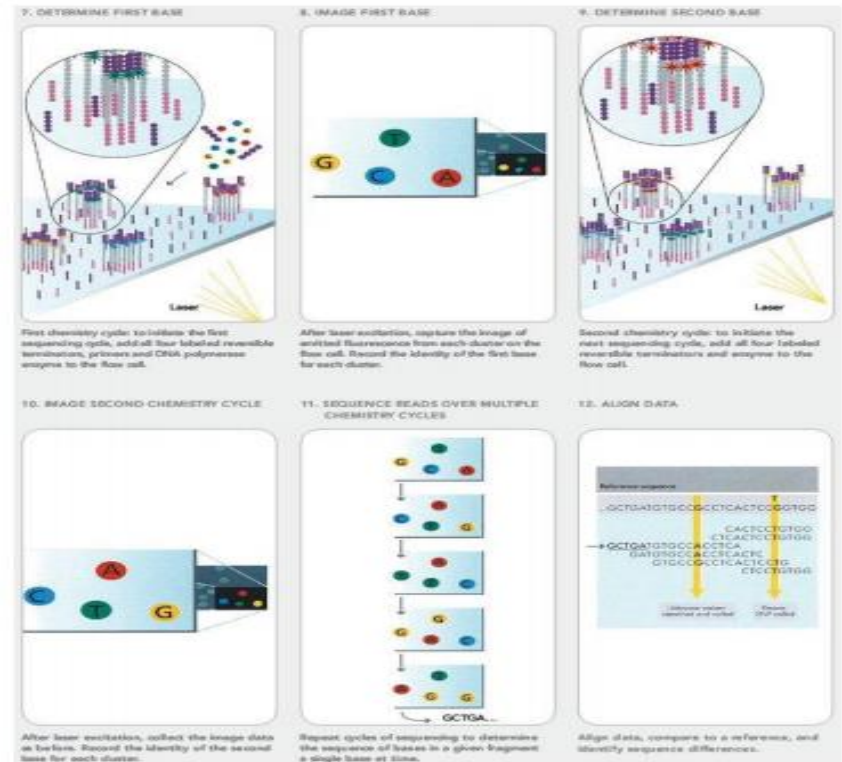Pan African Bioinformatics Network for H3Africa
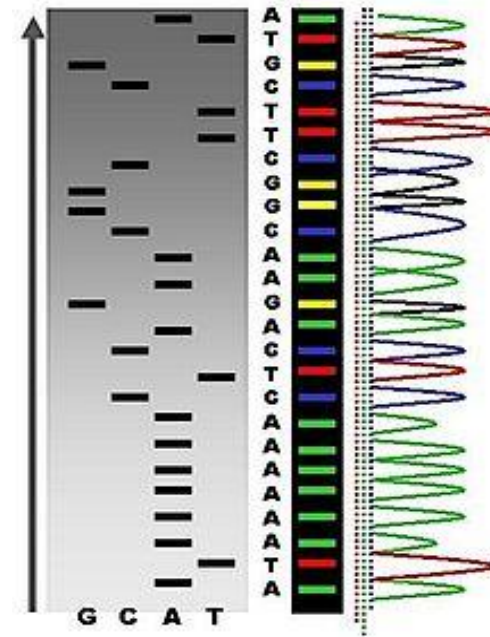
2. 454 and Ion Torrent

- Based on pyrosequencing

- About 1 million sequences in parallel

- 450bp, 700bp or longer

- Expensive chemicals

- More errors compared with Sanger
  - Homopolymers

- Ion Torrent is similar
  - Shorter reads (100bp-400bp)

- Much less expensive
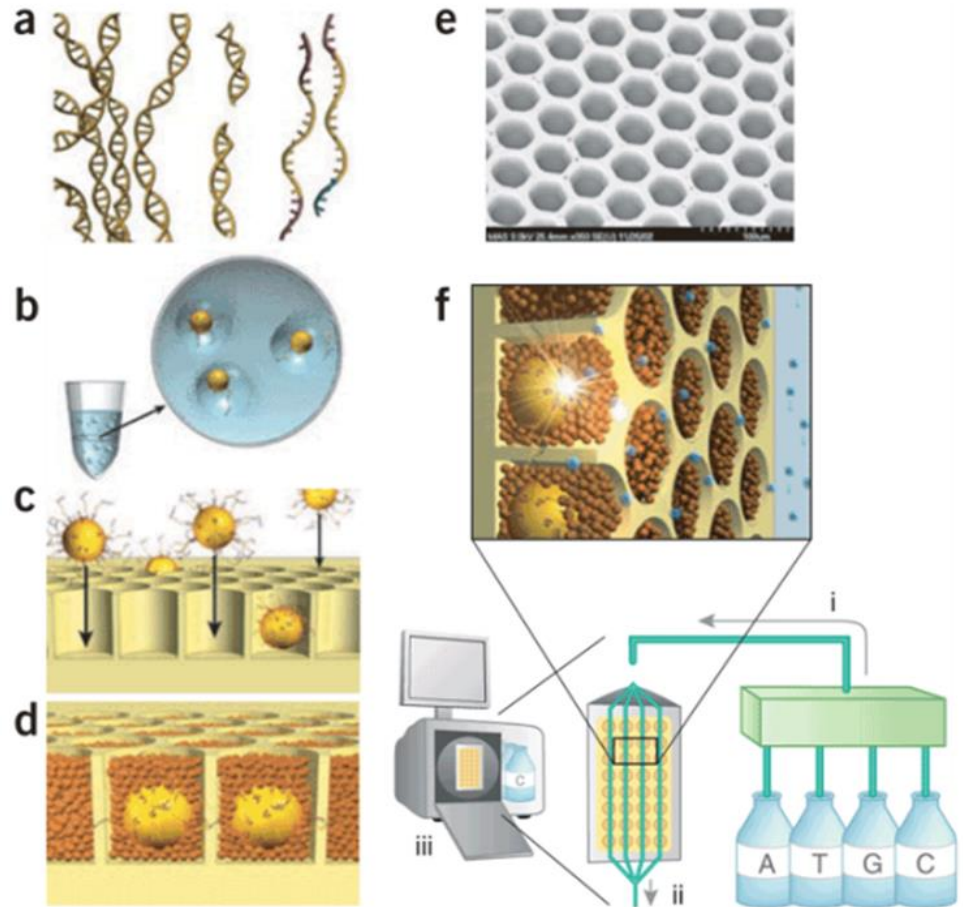
3. Illumina

- Quite short reads: 2x80-150bp (HiSeq), 2x300bp (MiSeq)

- 3 billion reads per flow cell (8 lanes; HiSeq), 20 million reads (MiSeq)

- Much cheaper than 454

- More error compared both with Sanger and 454
  - Error distribution different to 454, single base pairs more frequent, homopolymer errors rare
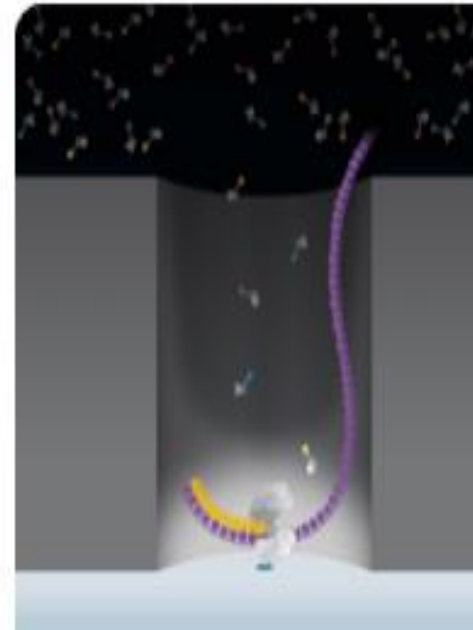

From: https://www.illumina.com/

H3ABioNet
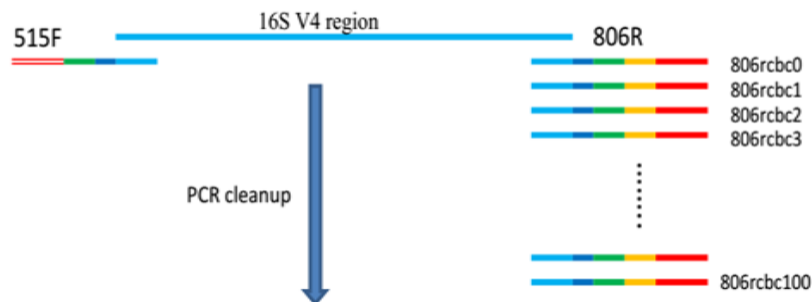Pan African Bioinformatics Network for H3Africa

4. PacBio SMRT

- Single Molecule Real-Time Sequencing nanotechnology:
  - Observe a single nucleotide being added to polymer (fluorescence)

- Long reads: 3 kb average

- 90 Mb per run

- Often combined with Illumina

From: http://www.pacificbiosciences.com

SUMMARY of PCR and Sequencing Steps
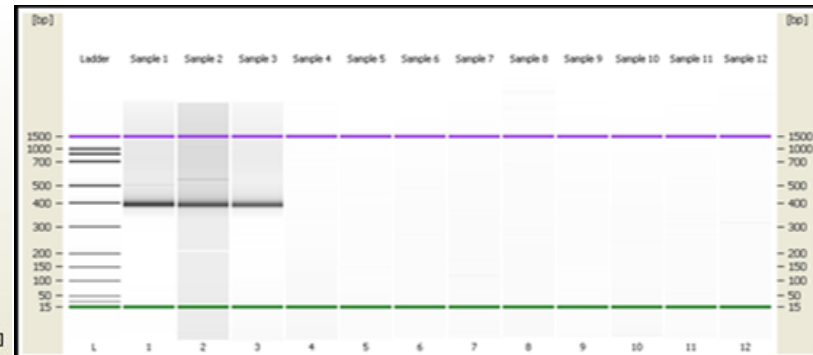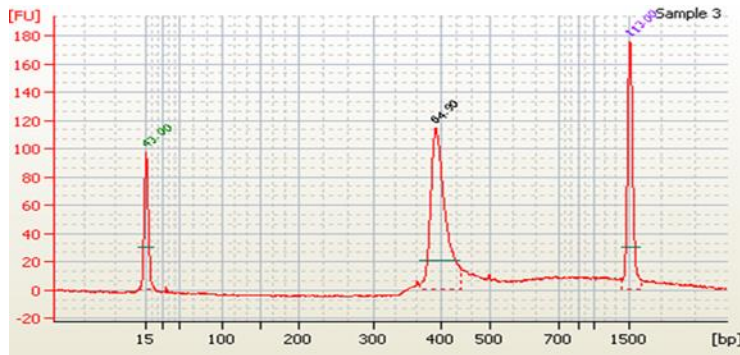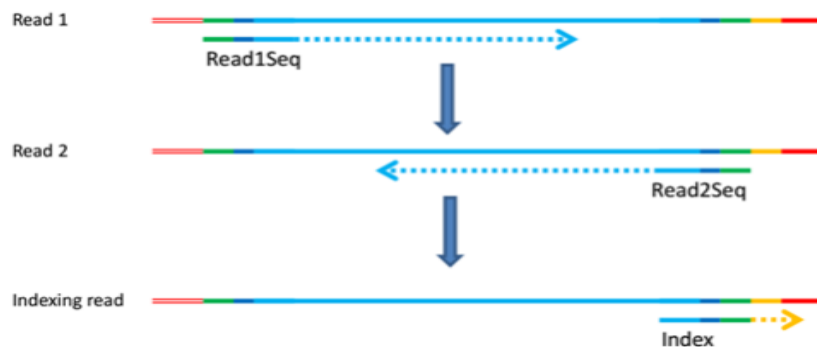
Output file formats

- **SFF** (standard flowgram format) - 454

- **Fastq - Illumina**

- **BAM** (binary of sequence alignment map) – Ion Torrent

- **Metadata** in tabular format that can be used for downstream analysis

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Module 4 : Part III

Quality Assessment of Raw Sequences

"Garbage in garbage out"

It takes a good lab practice to produce reliable data for the downstream processing

If we mess-up in Wetlab it can not be corrected in Dry lab

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Assessment of the Quality of Raw Reads

BM-AO-KITdnam-extractioncontrolspike-1-kitcomparison-P3-B07_S211_L001_R1_001.fastq

BM-AO-KITdnam-extractioncontrolspike-1-kitcomparison-P3-B07_S211_L001_R2_001.fastq

BM-AO-KITdnam-sequencing-control-1-wrrepeat-P3-C06_S222_L001_R1_001.fastq

BM-AO-KITdnam-sequencing-control-1-wrrepeat-P3-C06_S222_L001_R2_001.fastq

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).

- Line 2 is the raw sequence letters.

- Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence

Phred score

Phred score (Q) describes the quality of the sequences and is presented by an integer value. The larger the value the more confidence there can be in the output.

The probability a base is called incorrectly is given by

$$(P) = \boxed{10^{(-Q/10)}}$$

Thus, $Q = -10 \log_{10}(P)$

- For example, if the probability of an error (P) = 0.01, then Q will be will be = 20. What if P = 0.0004? What will be the Q score?

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger								
Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A

H3ABioNet
Pan African Bioinformatics Network for H3Africa

It is RECOMMENDED to check the quality of raw reads before further processing of the raw Sequences.

FASTQC tool by Andrew, 2010 ; is a very powerful open source quality control tool for assessing high throughput sequence data

Details found here:

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Example of good Illumina data:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html
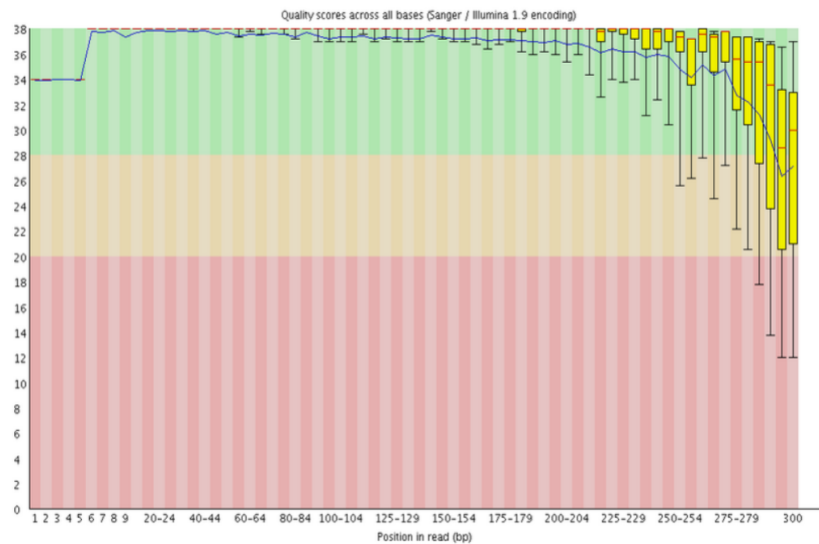
Example of bad Illumina data

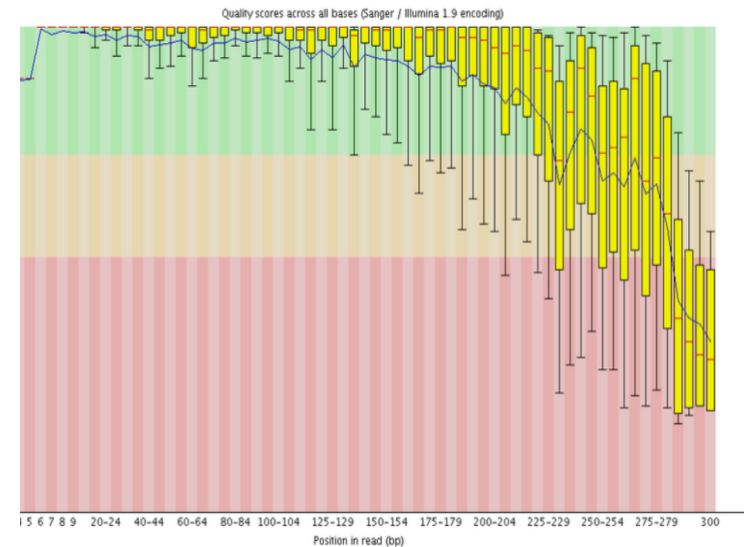https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

H3ABioNet

Pan African Bioinformatics Network for H3Africa

Sample of FASTQC output (1)

Sample FASQC output (2)

R1 (Forward read)

R2 (Reverse read)

What about this sample?   What about this?
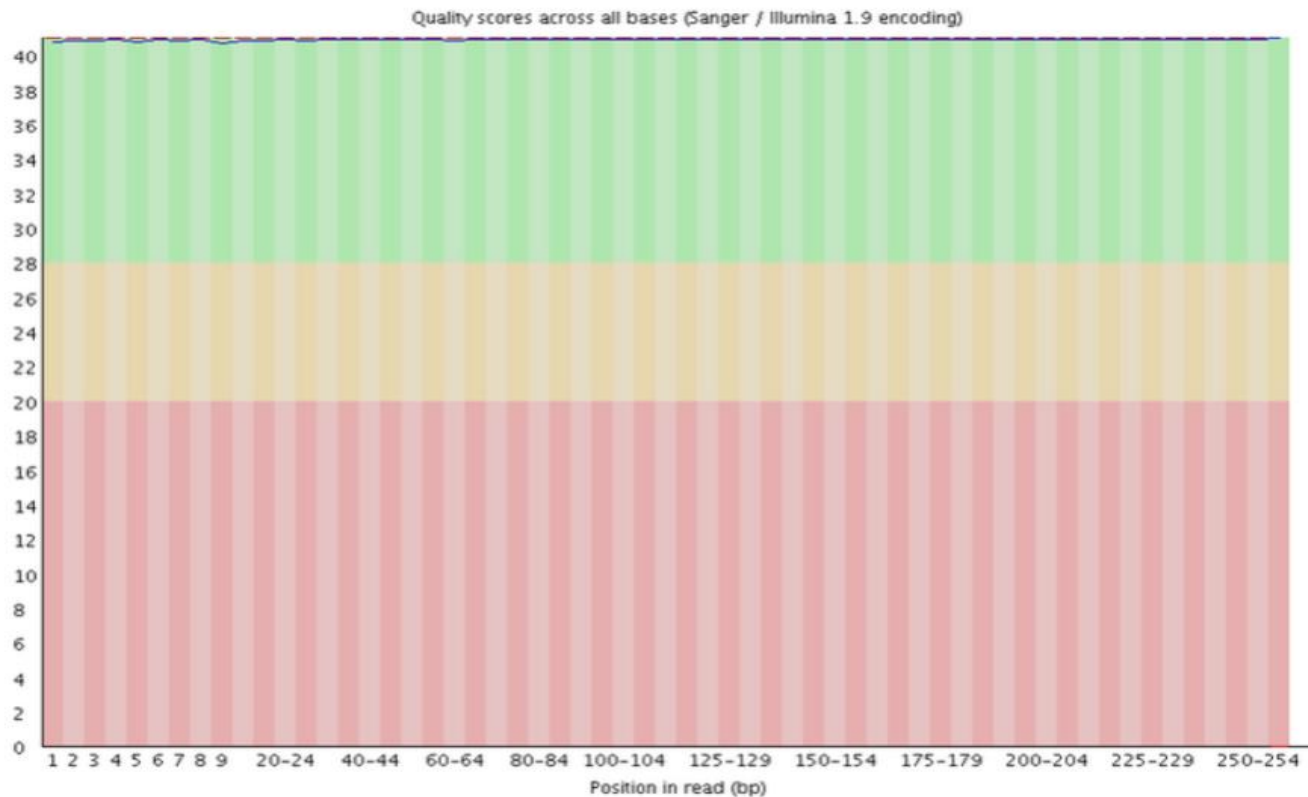
Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Position in read (bp)

Module 4: Part IV (a)

16S rRNA Bioinformatics Processing Pipelines

H3ABioNet
Pan African Bioinformatics Network for H3Africa

16S rRNA diversity analysis workflow
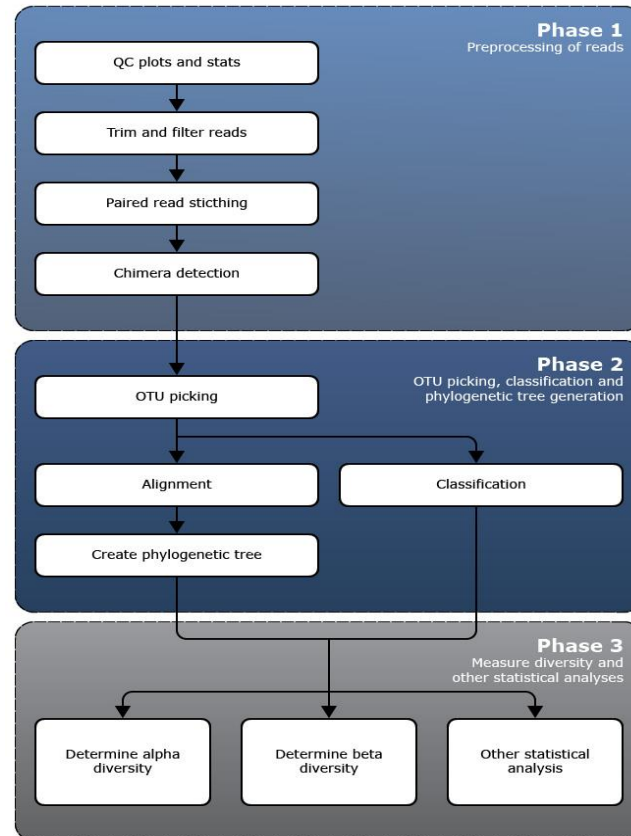
Pre-processing of input reads

- Demutliplexing

- Check for short read lengths

- Remove low quality bases

- Remove adapters and primers

- Remove chimeric sequences

- Merging of reads

Phase 1
Preprocessing of reads
- QC plots and stats
- Trim and filter reads
- Paired read sticthing
- Chimera detection

Phase 2
OTU picking, classification and phylogenetic tree generation
- OTU picking
- Alignment
- Classification
- Create phylogenetic tree

Phase 3
Measure diversity and other statistical analyses
- Determine alpha diversity
- Determine beta diversity
- Other statistical analysis

OTU picking

- An operational taxonomic unit is an operational definition of a species or group of species often used when only DNA sequence data is available.

- Clusters are formed based on sequence identity.

- 3 different approaches
  - De novo OTU picking
  - Closed reference OTU picking
  - Open reference picking

- A representative sequence is selected for downstream analysis.
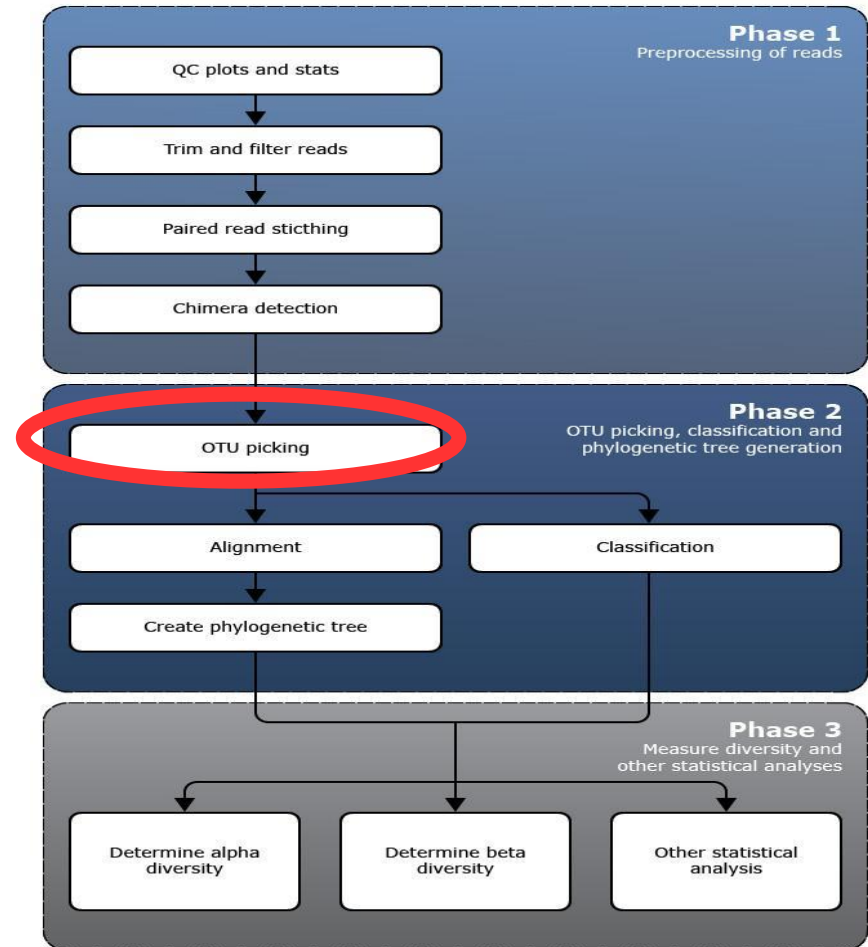
Classification

- A taxonomic identity is assigned to each representative sequence.

- Classification are done against three main reference databases with aligned, validated and annotated 16S rRNA genes: GreenGenes, Ribosomal Database Project (RDP) and Silva.

H3ABioNet
Pan African Bioinformatics Network for H3Africa
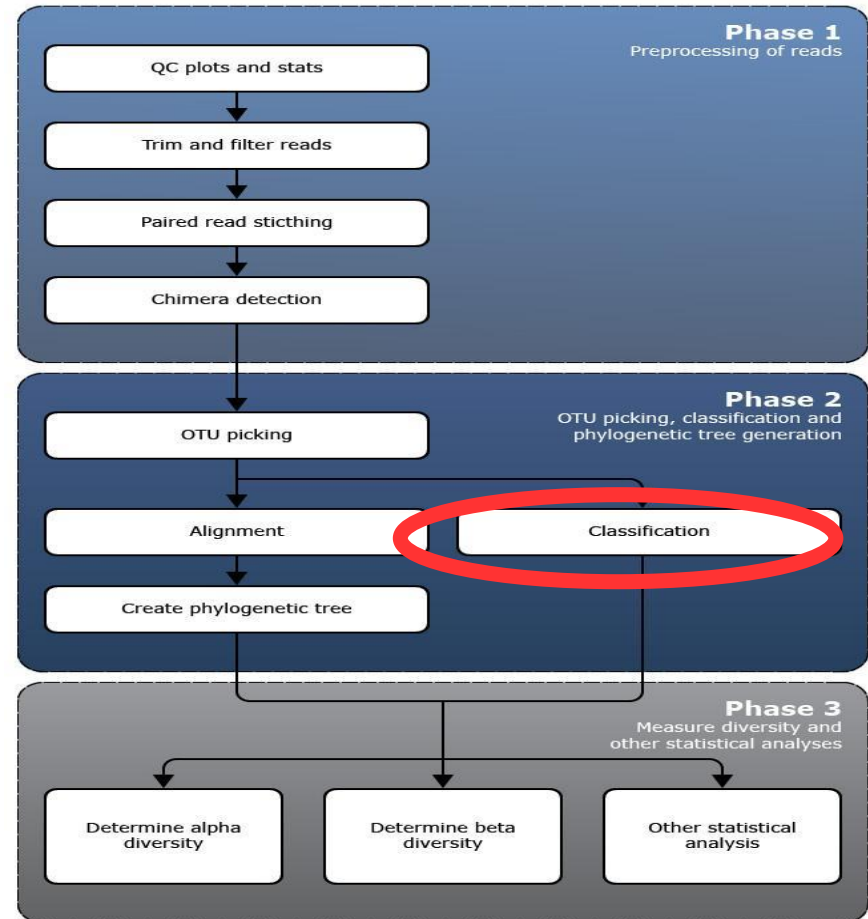
Alignment

- Alignment is the first step in generating a phylogenetic tree to understand the evolutionary relationship between samples.

- The aligners of choice are those that does alignments to a template (secondary structure) of the 16S sequence.

H3ABioNet
Pan African Bioinformatics Network for H3Africa
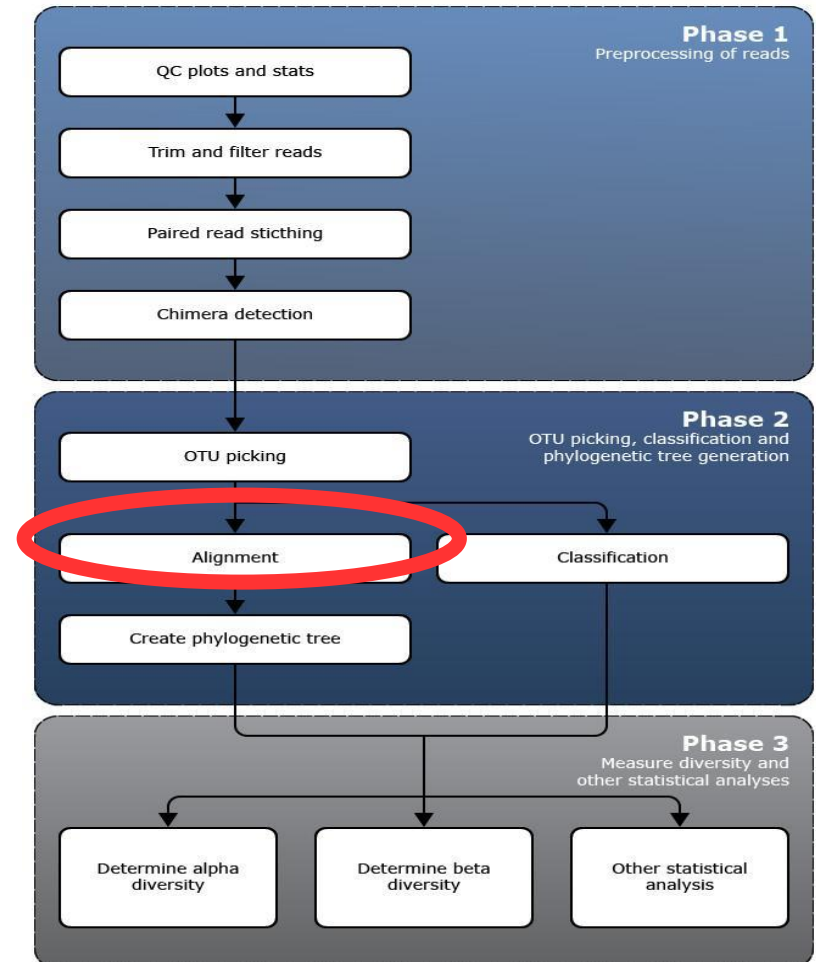
Create a phylogenetic tree

- The phylogenetic tree represents the relationship between the sequences in terms of the evolutionary distance from a common ancestor.

- In downstream analysis this tree is used for example in calculating the UniFrac distances and some alpha diversity measures.

H3ABioNet
Pan African Bioinformatics Network for H3Africa
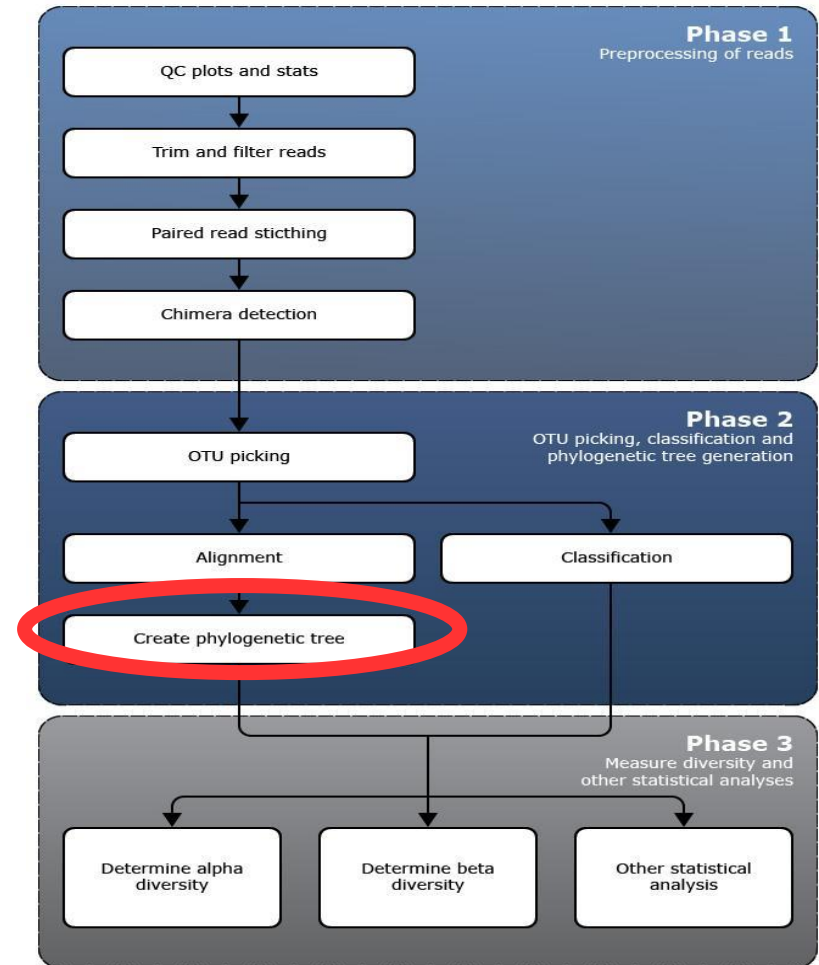
Other 16S rRNA processing pipelines

4.3.1. Mothur Microbiome pipeline

Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities▽

Patrick D. Schloss,[1,2]* Sarah L. Westcott,[1,2] Thomas Ryabin,[1] Justine R. Hall,[3] Martin Hartmann,[4] Emily B. Hollister,[5] Ryan A. Lesniewski,[6] Brian B. Oakley,[7] Donovan H. Parks,[8] Courtney J. Robinson,[2] Jason W. Sahl,[9] Blaz Stres,[10] Gerhard G. Thallinger,[11] David J. Van Horn,[2] and Carolyn F. Weber[12]

Department of Microbiology, University of Massachusetts, Amherst, Massachusetts[1]; Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan[2]; Department of Biology, University of New Mexico, Albuquerque, New Mexico[3]; Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada[4]; Department of Soil and Crop Sciences, Texas A&M University, College Station, Texas[5]; Department of Soil, Water, and Climate, University of Minnesota, St. Paul, Minnesota[6]; Department of Biological Sciences, University of Warwick, Coventry, United Kingdom[7]; Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada[8]; Environmental Science and Engineering, Colorado School of Mines, Golden, Colorado[9]; Department of Animal Science, University of Ljubljana, Ljubljana, Slovenia[10]; Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria[11]; and Department of Biological Sciences, Louisiana State University, Baton Rouge, Lousiana[12]

mothur aims to be a comprehensive software package that allows users to use a single piece of software to analyze community sequence data. It builds upon previous tools to provide a flexible and powerful software package for analyzing sequencing data. As a case study, we used mothur to trim, screen, and align sequences; calculate distances; assign sequences to operational taxonomic units; and describe the α and β diversity of

https://www.mothur.org/wiki/MiSeq_SOP

H3ABioNet
Pan African Bioinformatics Network for H3Africa

MOTHUR - Analyses Workflow
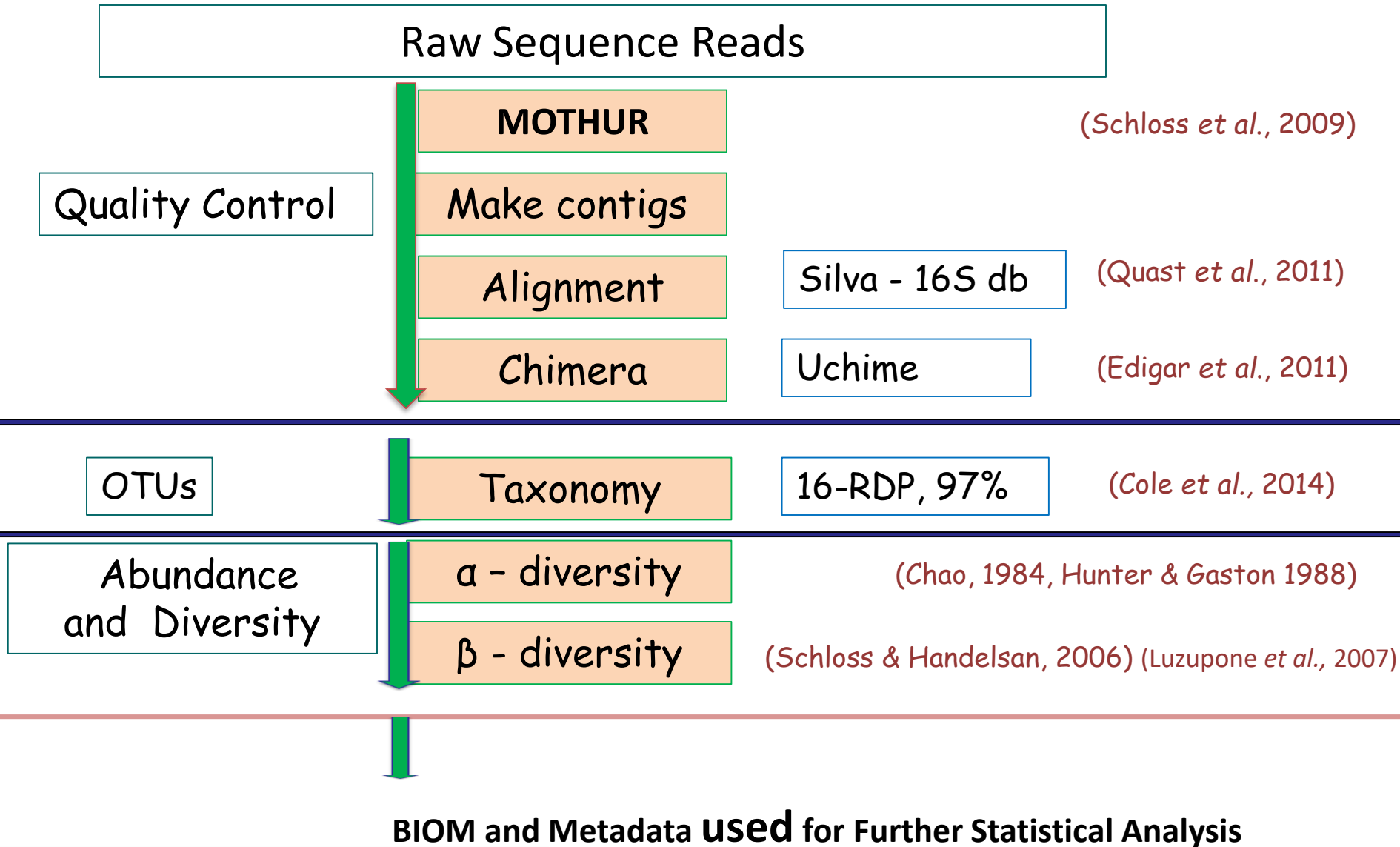
Raw Sequence Reads

MOTHUR — (Schloss *et al.*, 2009)

Quality Control — Make contigs

Alignment — Silva - 16S db — (Quast *et al.*, 2011)

Chimera — Uchime — (Edigar *et al.*, 2011)

OTUs — Taxonomy — 16-RDP, 97% — (Cole *et al.*, 2014)

Abundance and Diversity — α – diversity — (Chao, 1984, Hunter & Gaston 1988)

β - diversity — (Schloss & Handelsan, 2006) (Luzupone *et al.*, 2007)

BIOM and Metadata used for Further Statistical Analysis

H3ABioNet
Pan African Bioinformatics Network for H3Africa

QIIME PIPELINE

QIIME allows analysis of high-throughput community sequencing data

J Gregory Caporaso[1,12], Justin Kuczynski[2,12], Jesse Stombaugh[1,12], Kyle Bittinger[3], Frederic D Bushman[3], Elizabeth K Costello[1], Noah Fierer[4], Antonio Gonzalez Peña[5], Julia K Goodrich[5], Jeffrey I Gordon[6], Gavin A Huttley[7], Scott T Kelley[8], Dan Knights[5], Jeremy E Koenig[9], Ruth E Ley[9], Catherine A Lozupone[1], Daniel McDonald[1], Brian D Muegge[6], Meg Pirrung[1], Jens Reeder[1], Joel R Sevinsky[10], Peter J Turnbaugh[6], William A Walters[2], Jeremy Widmann[1], Tanya Yatsunenko[6], Jesse Zaneveld[2], and Rob Knight[1,11]

Rob Knight: rob.knight@colorado.edu

Developed by Caporaso et al., 2010; https://QIIME2.org/

What is QIIME?

Quantitative Insight Into Microbial Ecology

- It is a Next-generation Microbiome Bioinformatics Platform that is **Extensible, Free, Open source and Community Developed**.

- It is a powerful suite of microbiome analysis packages where researchers can start an analysis with **raw DNA sequence data and finish with publication-quality figures and statistical results**

- **QIIME2** is a complete **redesign** of the QIIME1 while retaining the features that makes it a powerful and widely-used tool

H3ABioNet
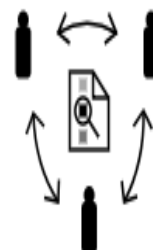Pan African Bioinformatics Network for H3Africa

QIIME2 Key Features

Tracking and Exploration

Easy sharing, Suite of tools
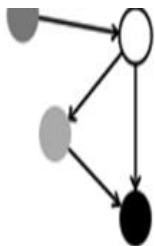
Automatically track your analyses with decentralized data provenance — no more guesswork on what commands were run!

Interactively explore your data with beautiful visualizations that provide new perspectives.

Easily share results with your team, even those members without QIIME 2 installed.

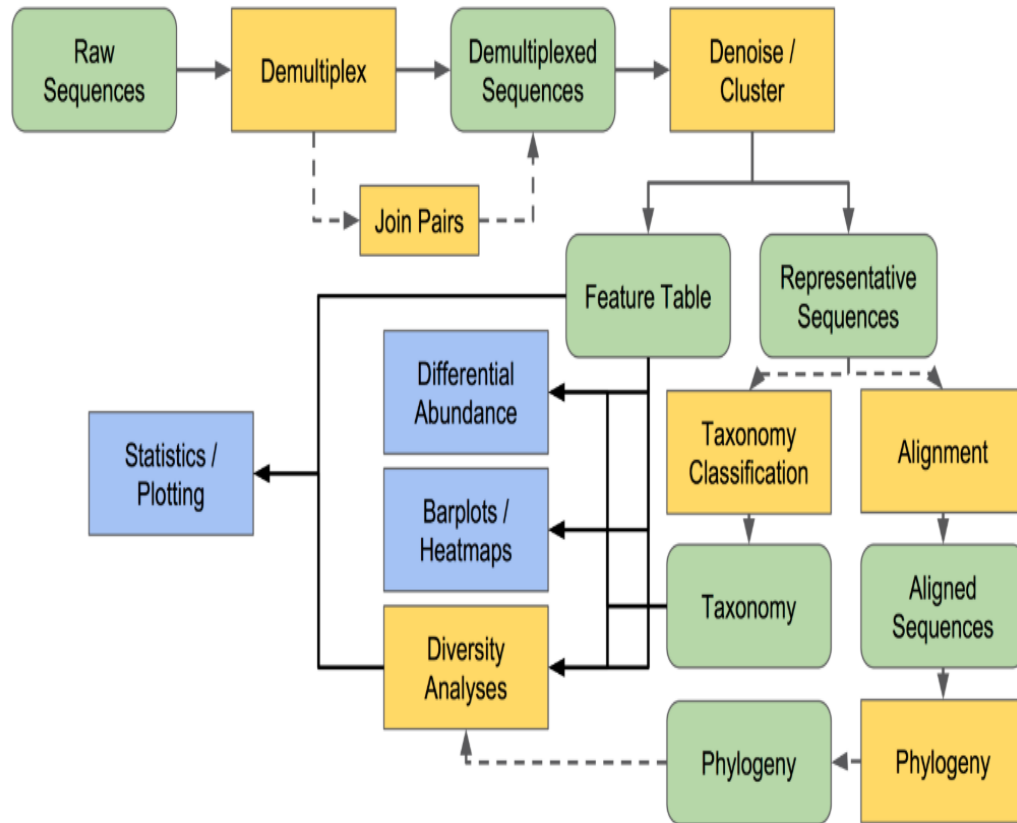Plugin-based system — your favorite microbiome methods all in one place.

Module 4: Part IV.. cont…… Bioinformatics Pipeline

QIIME2 CONCEPTUAL WORKFLOW

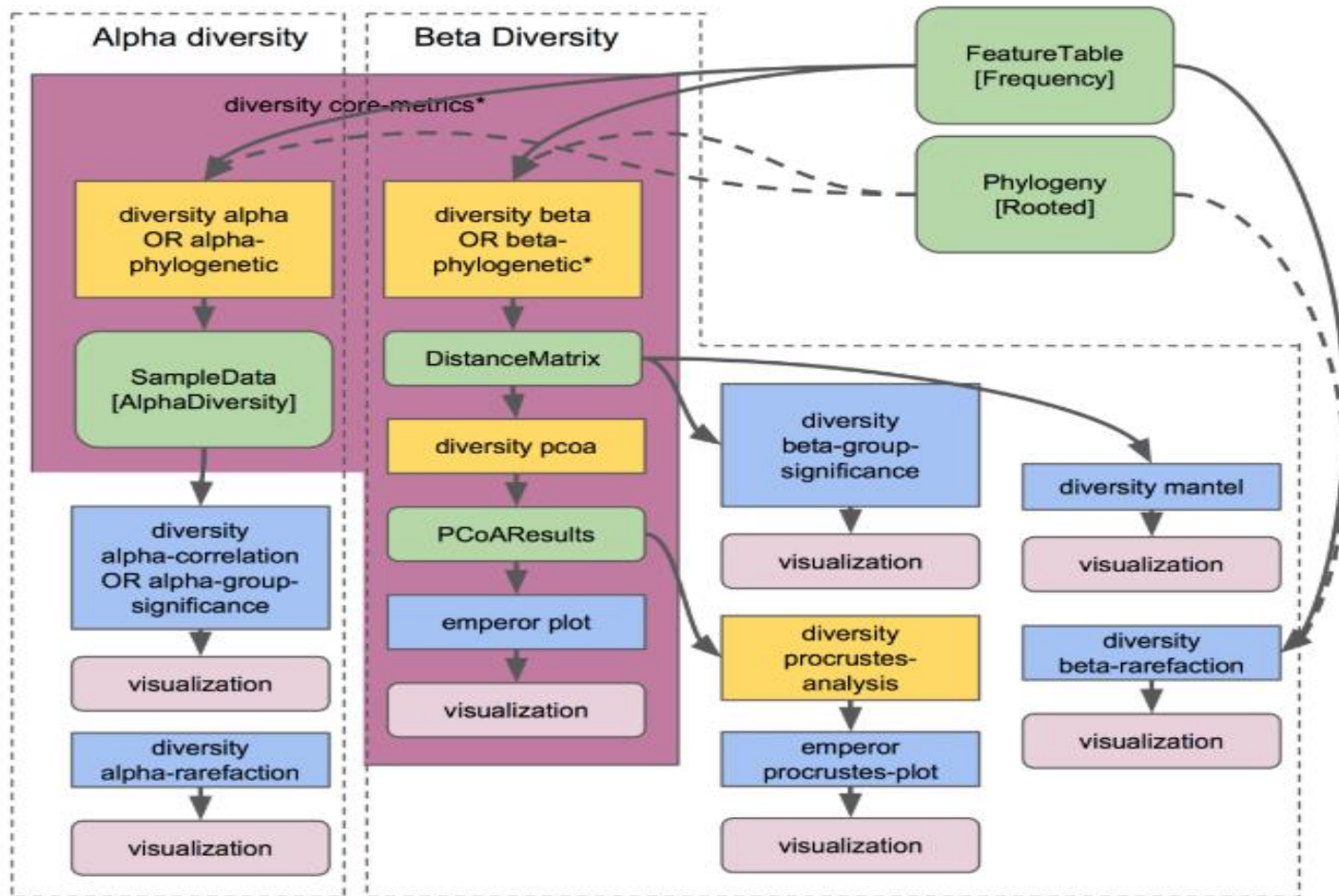Data are imported as a QIIME2 artifact to be used by a QIIME2 action (except the metadata).

Users may enter the workflow at different stages. Most will have raw sequence (e.g., FASTQ or FASTA) data, which should be imported.

Other users may start with demultiplexed sequence data or even a feature table

Alpha and Beta Diversity Analysis

q2-diversity plugin…..

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Example of Selected q2-plugins

Analyse longitudinal microbiome data:

q2-longitudinal plugins: Performs statistical analyses of longitudinal studies,

◊ i.e., where samples are collected from patients/subjects/sites repeatedly over time.

Predict the future (or the past)

q2-sample-classifier plugin: predict sample metadata as a function of feature data
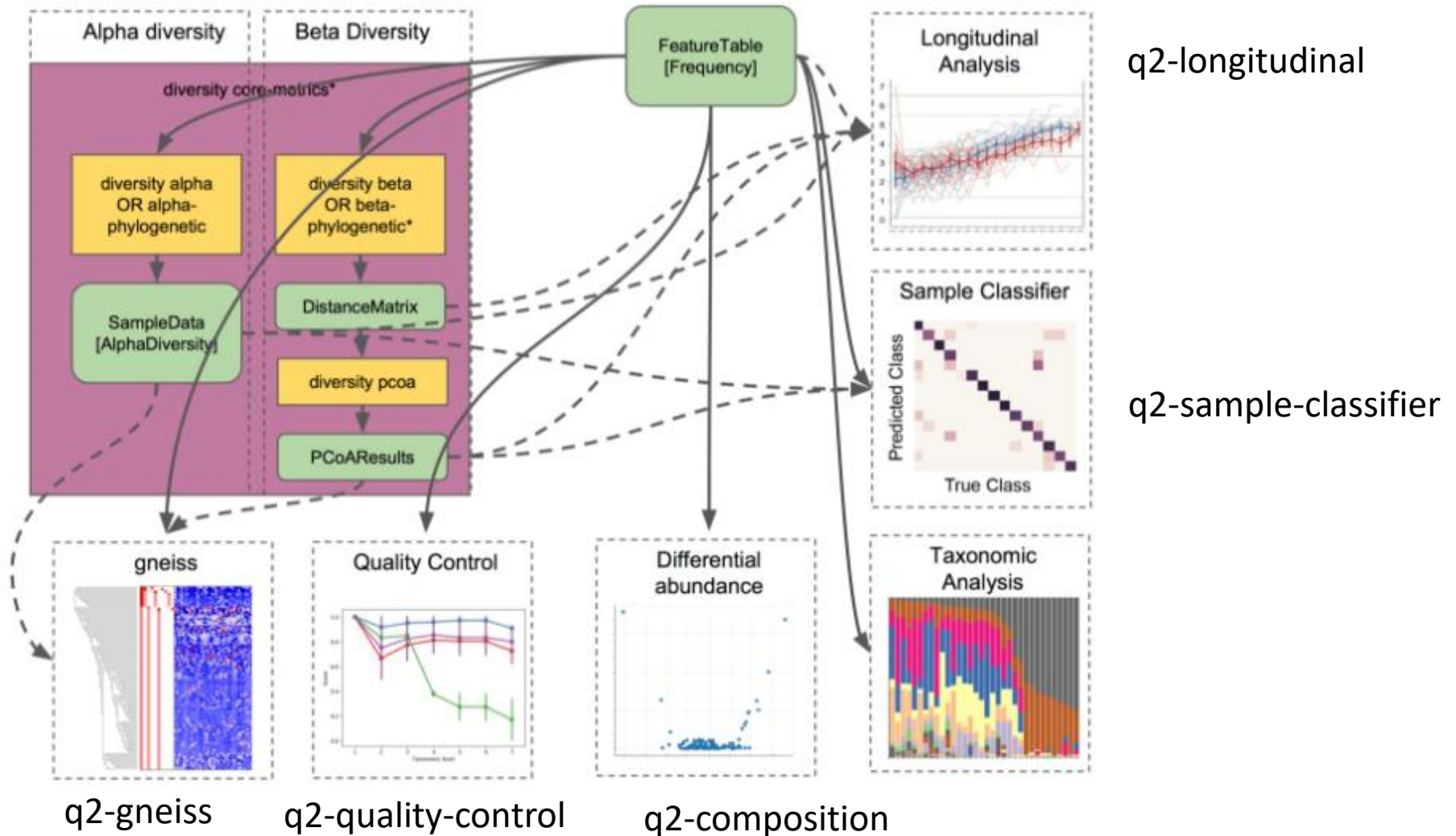
◊ i.e. can we use a fecal sample to predict cancer susceptibility? Or predict wine quality based on the microbial composition of grapes before fermentation?

Differential abundance

q2-composition plugin (based on ANCOM) and q2-gneiss plugin

◊ Determines which features are significantly more/less in different groups of samples

H3ABioNet
Pan African Bioinformatics Network for H3Africa

QIIME2 Statistical Analysis Types

q2-longitudinal

q2-sample-classifier

q2-gneiss q2-quality-control q2-composition

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Example (1) Plots- Microbial Profile
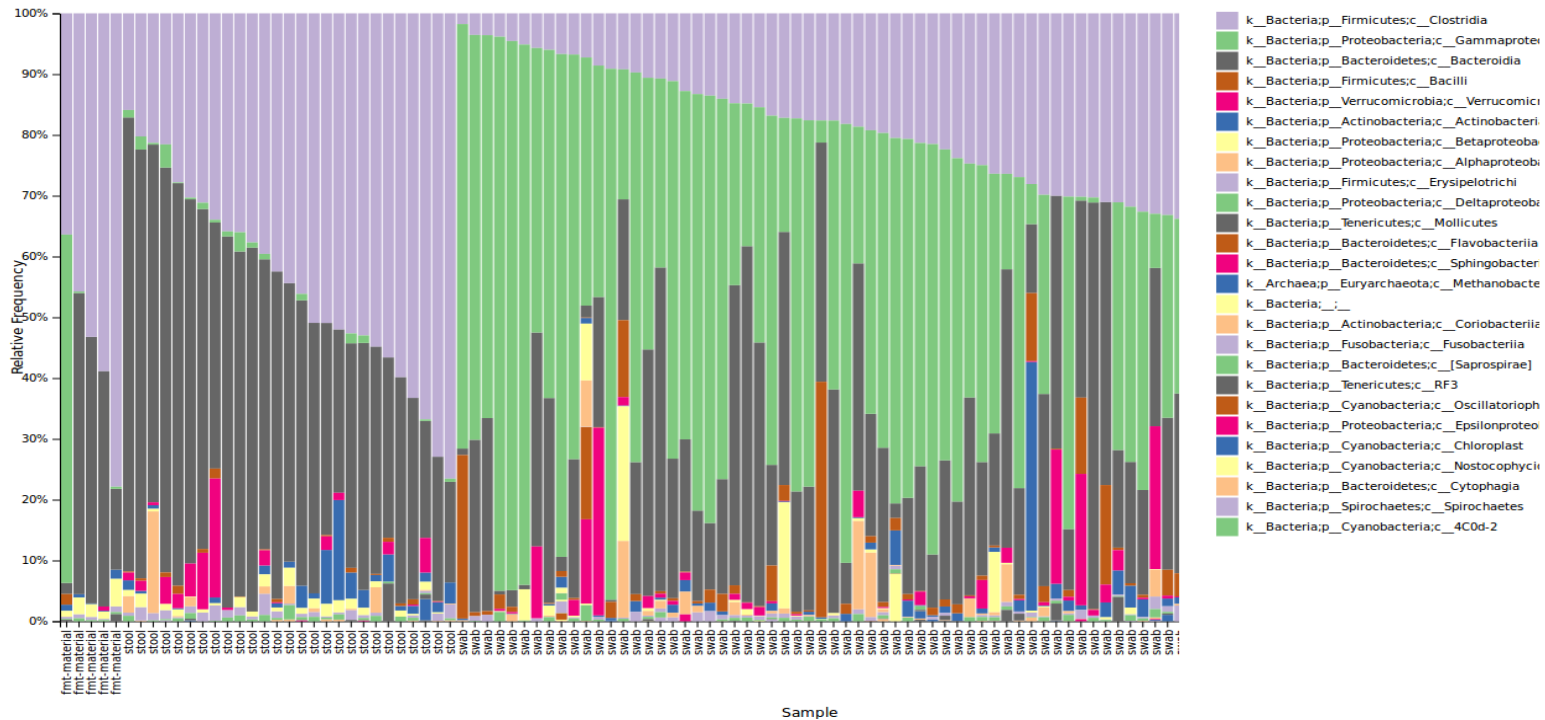
Example (2) Plots- beta diversity - PCoA plots
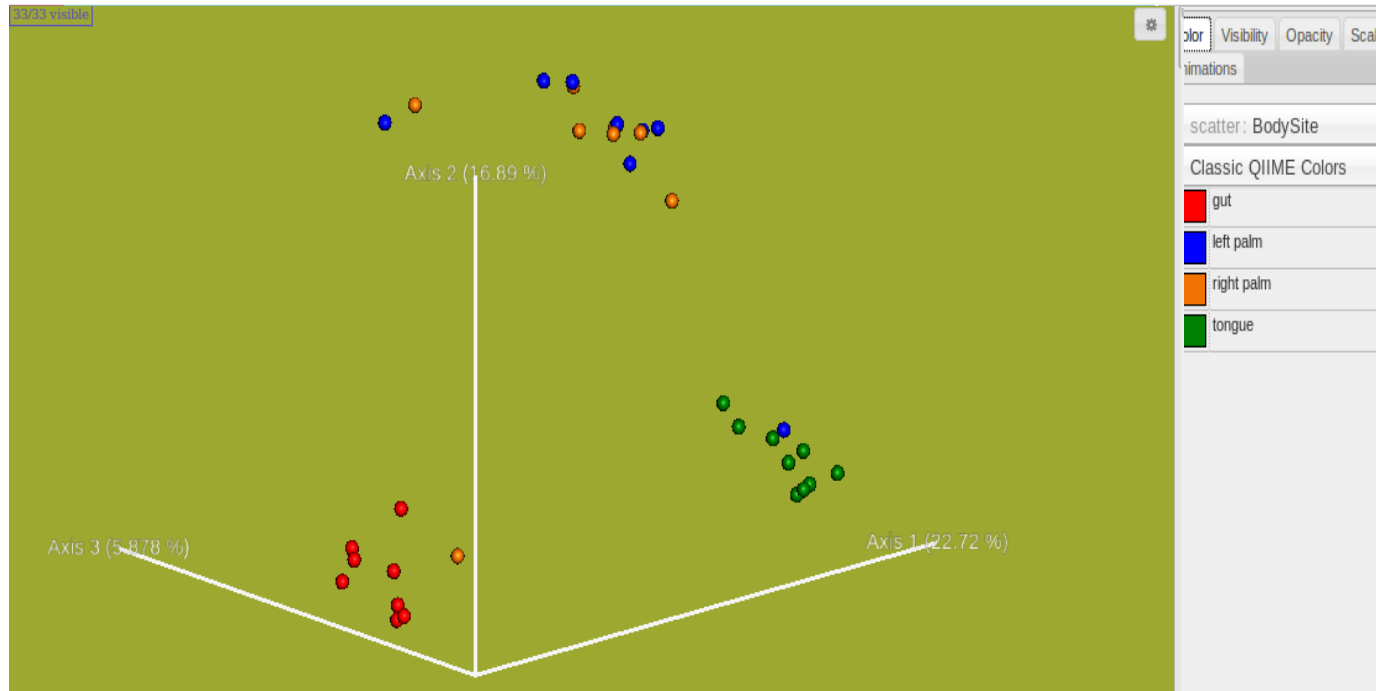
QIIME2 cont…

Further readings
 https://qiime2.org/

Installation of qiime2 here ….
https://docs.qiime2.org/2019.7/install/
https://docs.qiime2.org/2019.7/

Variety of tutorials found here…
https://docs.qiime2.org/2019.7/tutorials/

Subscribe to qiime2 forum for assistance here..
https://forum.qiime2.org/

DADA2 16S rRNA Bioimformatic Pipeline

16SrRNA Intermediate Bioinformatics Online Course: Int_BT_2019

Samson KM

H3ABioNet

Pan African Bioinformatics Network for H3Africa

DADA2: High resolution sample inference from Illumina amplicon data

Benjamin J Callahan, Paul J McMurdie, [...], and Susan P Holmes

Additional article information

Associated Data

► Supplementary Materials

Abstract

We present DADA2, a software package that models and corrects Illumina-sequenced amplicon errors. DADA2 infers sample sequences exactly, without coarse-graining into OTUs, and resolves differences of as little as one nucleotide. In
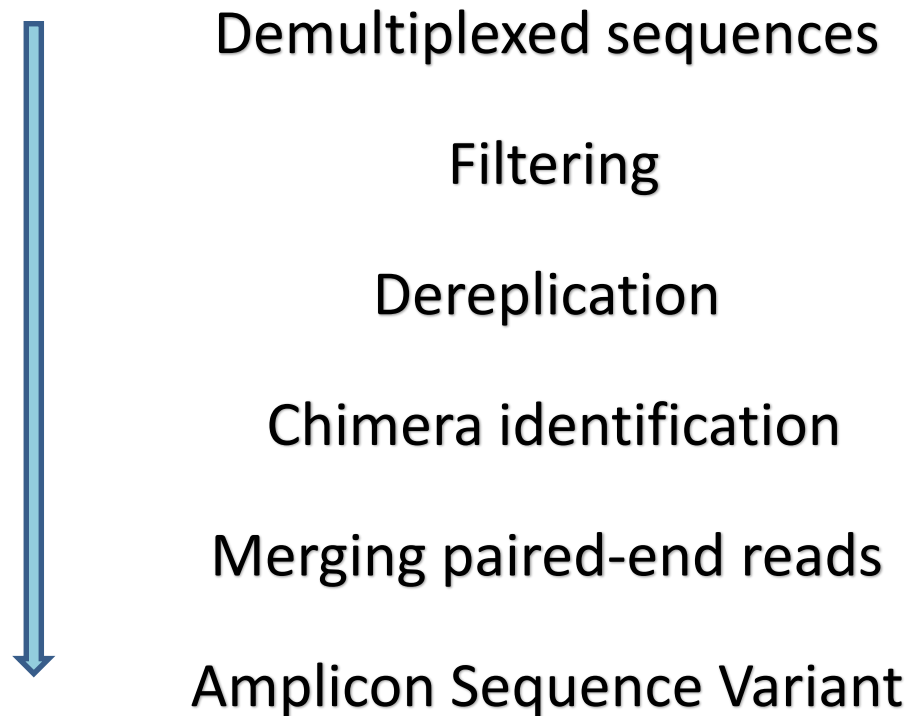
Developed by Callahan et al., 2016

DADA2 cont…

- **DADA**2 stands for - **D**ivisive **A**mplicon **D**enoising **A**lgorithm

- It is an **Open source** R-based package which corrects amplicons errors directly without constructing OTUs

- It infers sample sequences exactly and **resolves differences of as little as one nucleotide**

- The output are the Sequence Amplicon Variant (ASV) also named Exact Sequence Variant (ESV), as opposed to the OTUs

- In vie of the potential of this pipeline, the **H3ABioNet have developed DADA2 Nextflow Pipeline** for the Microbiome data processing. You will use this pipeline in practical session.

H3ABioNet
Pan African Bioinformatics Network for H3Africa

In view of that, The H3ABioNet have a developed a DADA2 Nextflow pipeline

H3ABioNet
Pan African Bioinformatics Network for H3Africa

The DADA2 workflow

Demultiplexed sequences

Filtering

Dereplication

Chimera identification

Merging paired-end reads

Amplicon Sequence Variant

Import into R Phyloseq package for downstream Analysis

DADA2 cont...

Further reading on DADA2

Installation of DADA2

https://bioconductor.org/packages/devel/bioc/vignettes/dada2/inst/doc/dada2-intro.html

Tutorials available here:

https://benjjneb.github.io/dada2/tutorial.html

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Module 4: 16S Microbiome Pipeline Summary
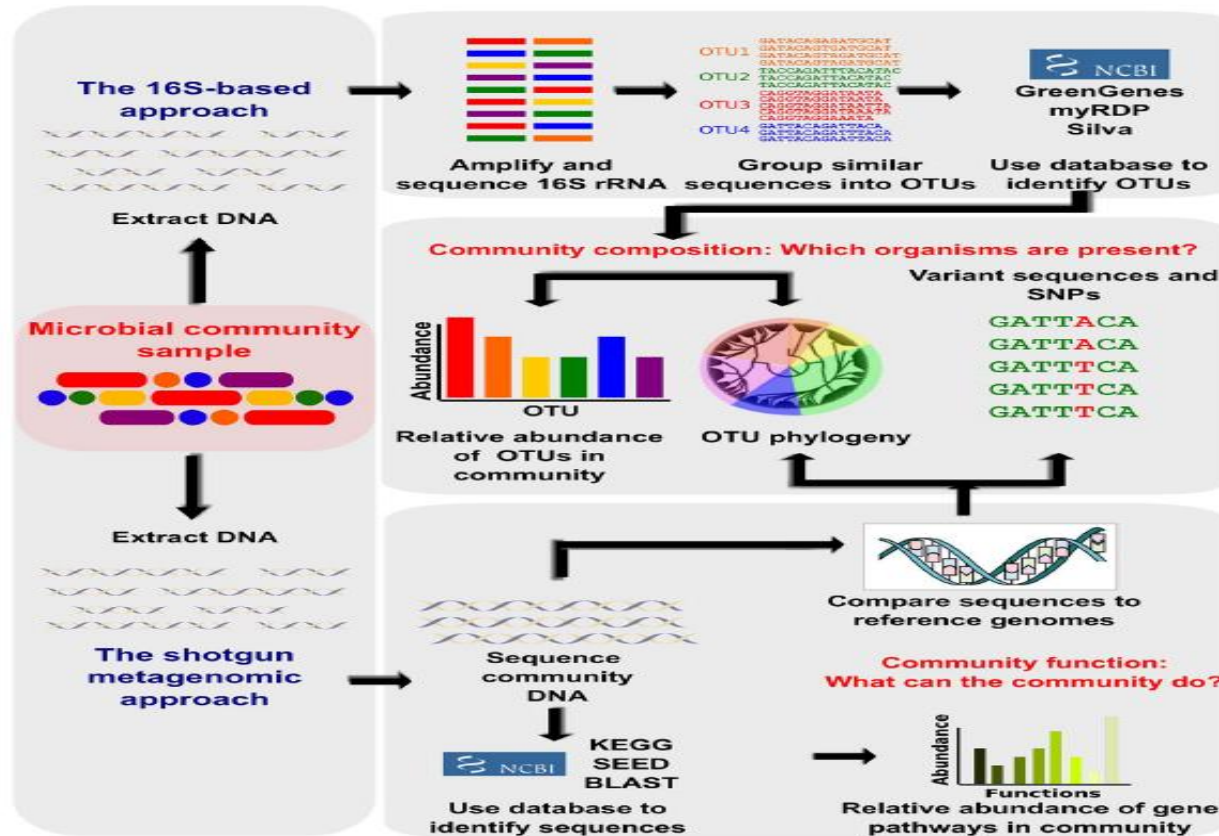
16S rRNA diversity vs shotgun analysis

Figure 1. Bioinformatic methods for functional metagenomics. Studies that aim to define the composition and function of uncultured microbial communities are often referred to collectively as "metagenomic," although this refers more specifically to particular sequencing-based assays. First, community DNA is extracted from a sample, typically uncultured, containing multiple microbial members. The bacterial taxa present in

H3ABioNet
Pan African Bioinformatics Network for H3Africa

SUMMARY

Raw sequence data

16S rRNA PIPELINE
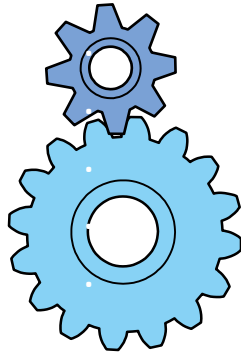
BIOM file
otus_repsetOUT.fa
Phylogenetic tree
Taxonomy file
otu_table

STATISTICIAN

Other available pipelines

- UPARSE:
  http://www.drive5.com/uparse

- IM Tornado: https://github.com/pjeraldo/imtornado2

- FROGS: https://github.com/geraldinepascal/FROGS

- VSEARCH: https://github.com/torognes/vsearch

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Acknowledgements

H3ABioNet for some contents of these slides

The CBIO at UCT for the inhouse microbiome-pipeline

Verena Ras and Gerrit Botha, for their devoted time on this training

Dar es Salaam Institute of Technology : My employer

Thank you

Asanteni