H3ABioNet
Pan African Bioinformatics Network for H3Africa

# Data analysis of 16S rRNA amplicons

## Computational Metagenomics Workshop University of Mauritius

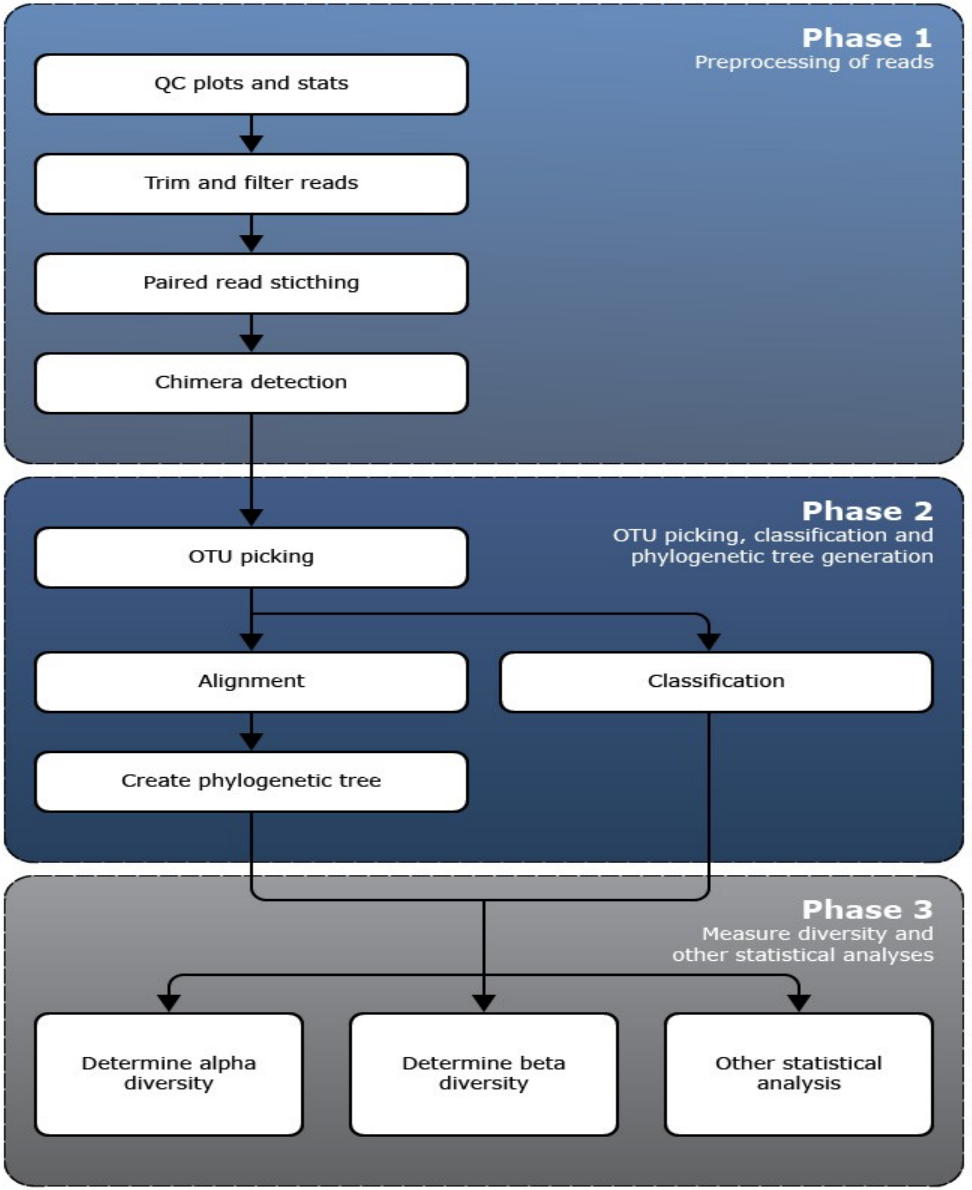
Gerrit Botha

December 2014

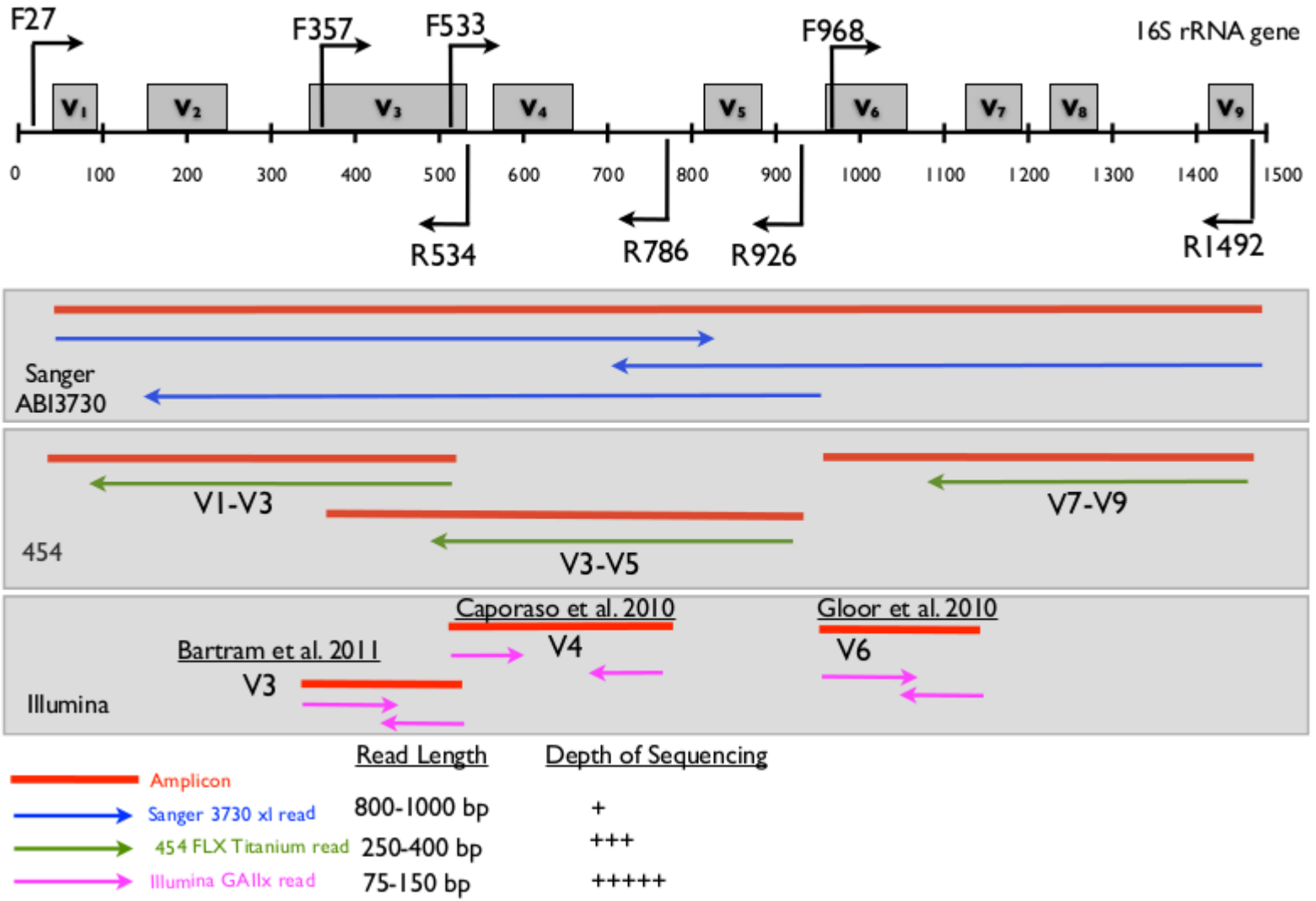
# Introduction

- The genes encoding the RNA component of the small subunit of ribosomes, commonly known as the 16S rRNA in bacteria and archaea, are among the most conserved 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 from an environmental sample and subsequent sequencing of a sufficiently large number of individual amplicons enables the analysis of the diversity of clades in the sample and a rough estimate of their relative abundance.
- Follow the SOP here: http://h3abionet.org/tools-and-resources/sops/16s-rrna-diversity-analysis

# 16S rRNA diversity analysis workflow

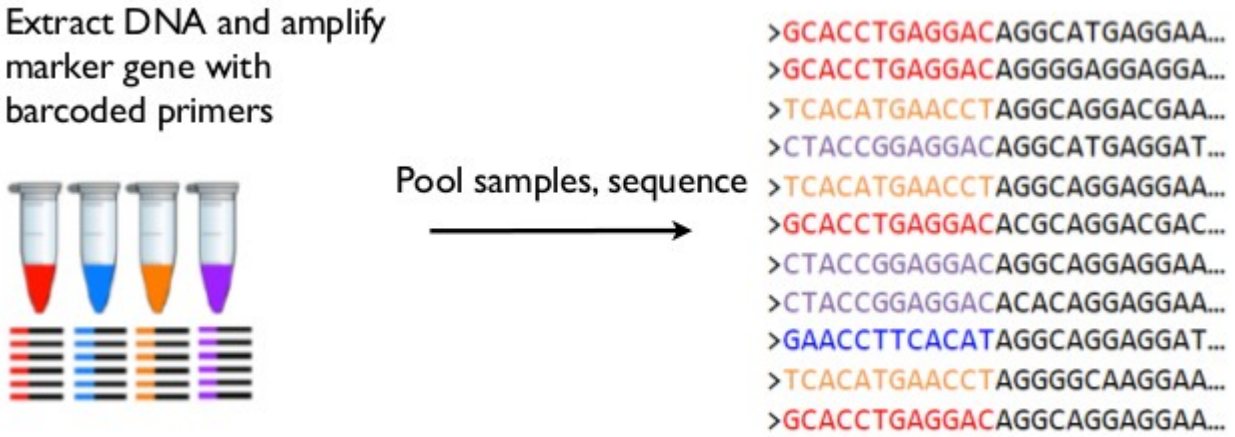


# Sample 16S rRNA gene region



# Sample pooling

- Twelve base error-correcting barcodes allow hundreds of samples per run




- *Micah Hamady, et al., Nature Methods, 2008. Error-correcting barcodes for pyrosequencing hundreds of samples in multiplex*

# Inputs - File formats

- SFF (standard flowgram format) - 454
- Fastq - Illumina
- BAM (binary of sequence alignment map) – Ion Torrent PGM
- Metadata that can be used for downstream analysis

# Fastq format

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

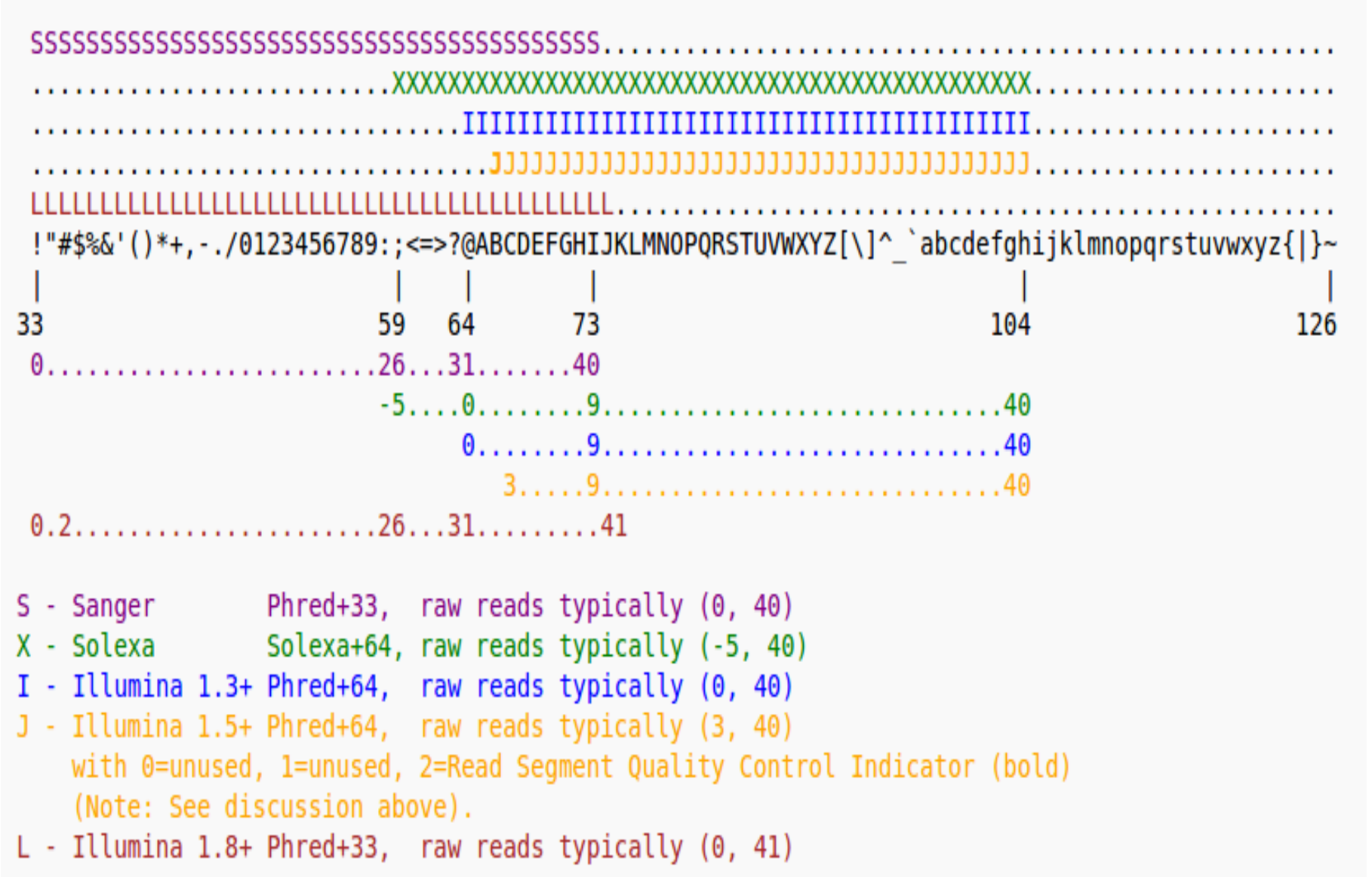
- Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence

From: http://en.wikipedia.org/wiki/FASTQ_format

# Pred scores (1)

- Used to describe the quality of the output and present this by an integer value. **The larger the value the more confidence there can be in the output.**
- The probability a base is called incorrectly = $10^{(-Q/10)}$
- Q = 10, The probability that the call is wrong = 0.1
- Q = 20, The probability that the call is wrong = 0.01
- Q = 30, The probability that the call is wrong = 0.001
- For Illumina a > 30Q base call quality value is good

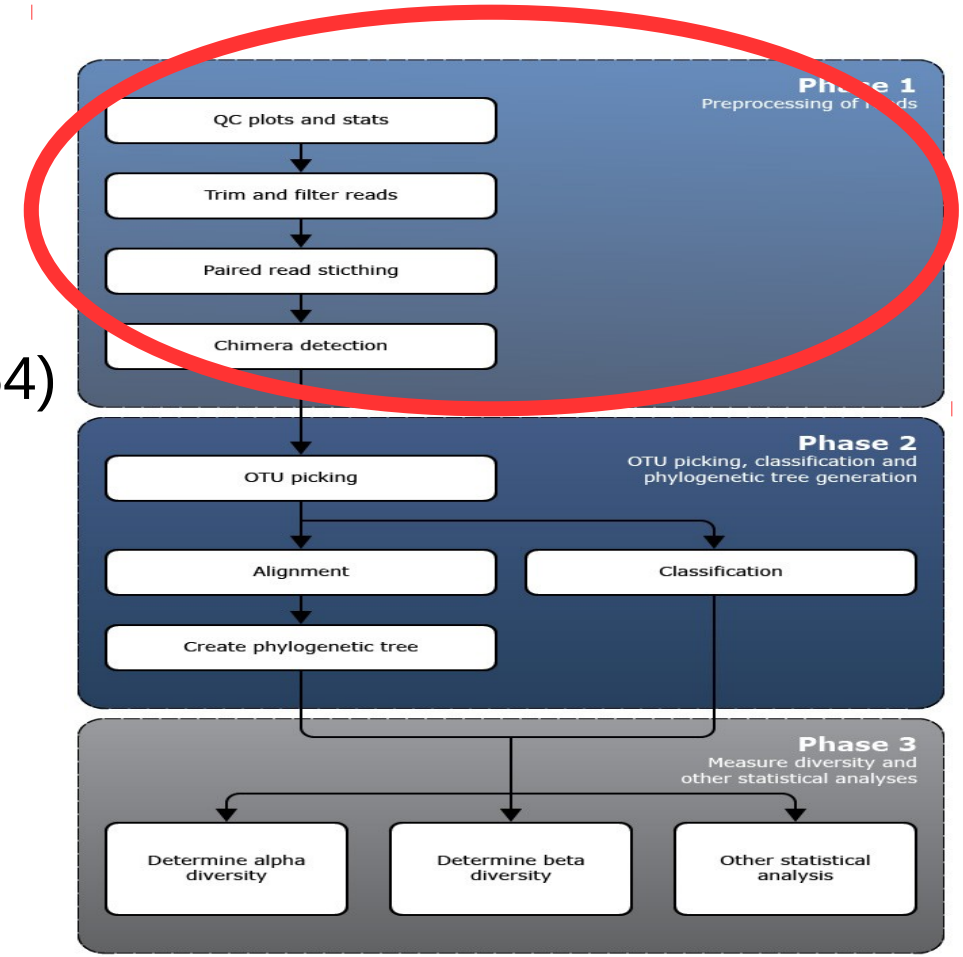
# Phred scores (2)




From: http://en.wikipedia.org/wiki/FASTQ_format

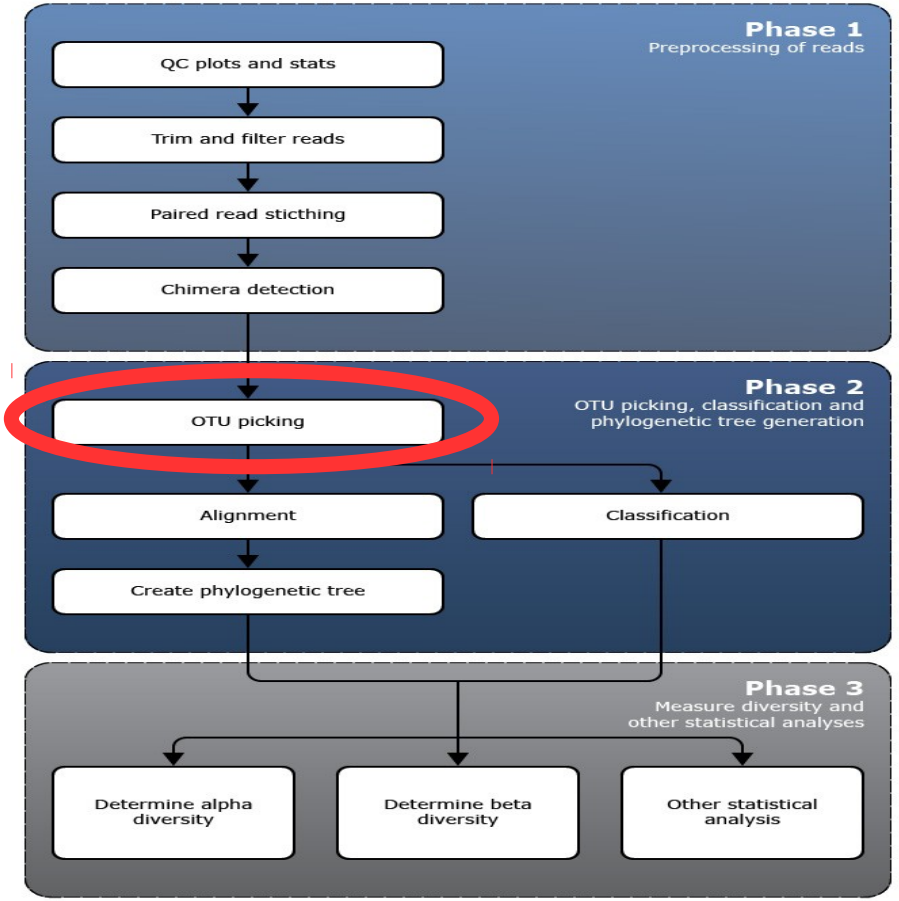
# Preprocessing of input reads

- Quality control
  - Check for barcode mismatches
  - Check for primer mismatches
  - Check for homopolymer runs (454)
  - Check for ambiguous bases
  - Check for short read lengths
  - Remove low quality bases
  - Remove adapters
  - Remove chimeric sequences
- Demutliplexing
- Read merging (Illumina forward and reverse reads)



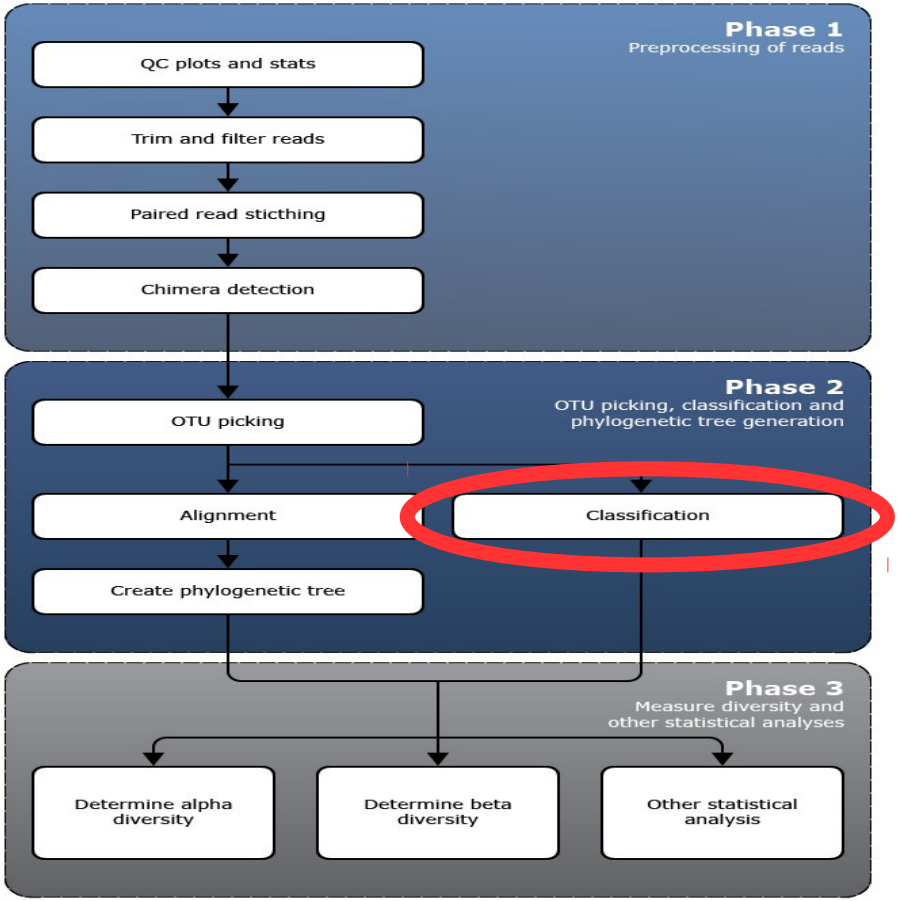
# OTU picking

- 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.



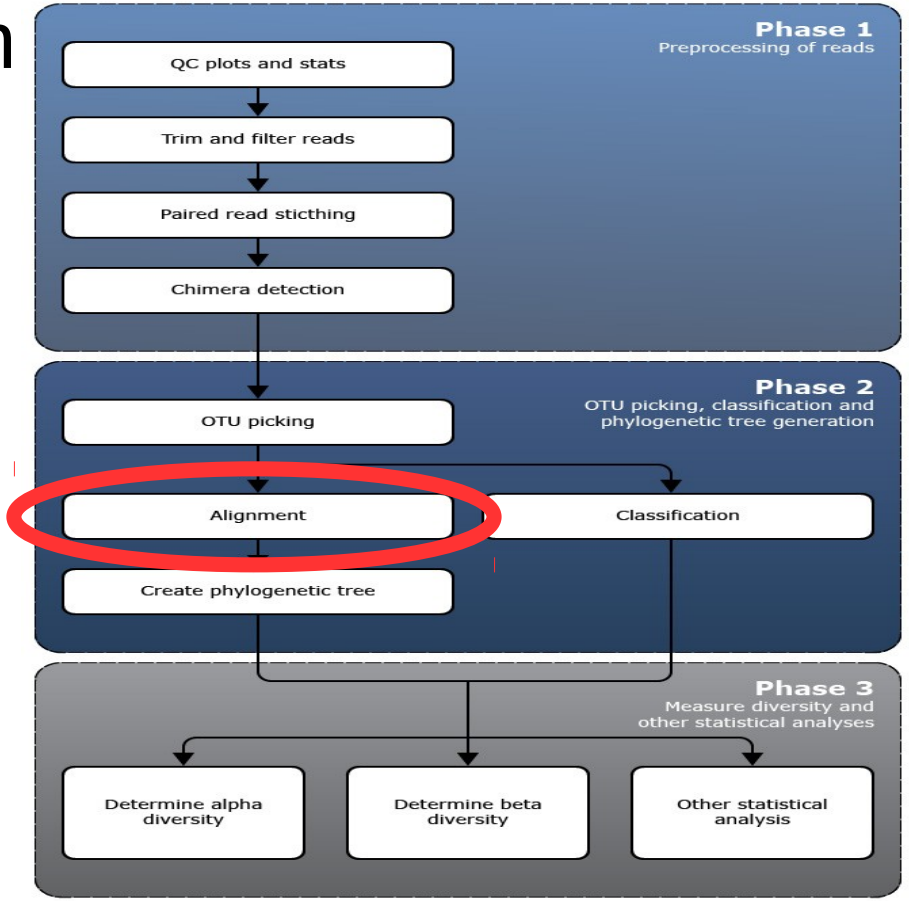
# Classification

- 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.



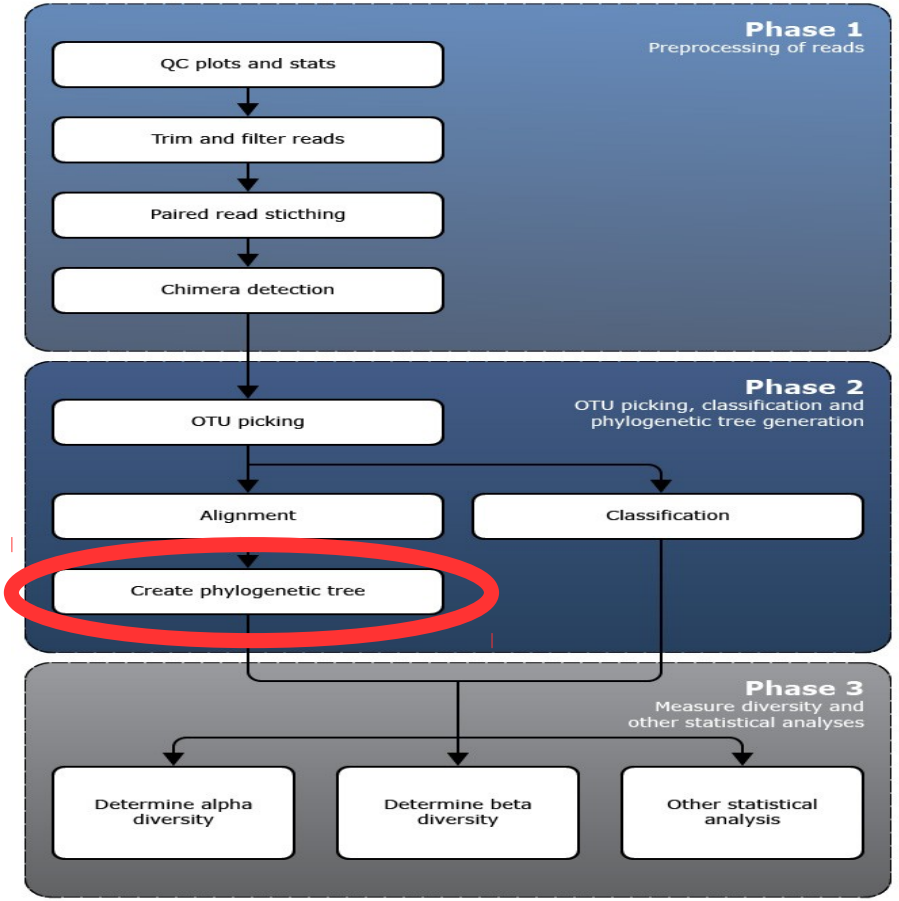
# Alignment

- 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.



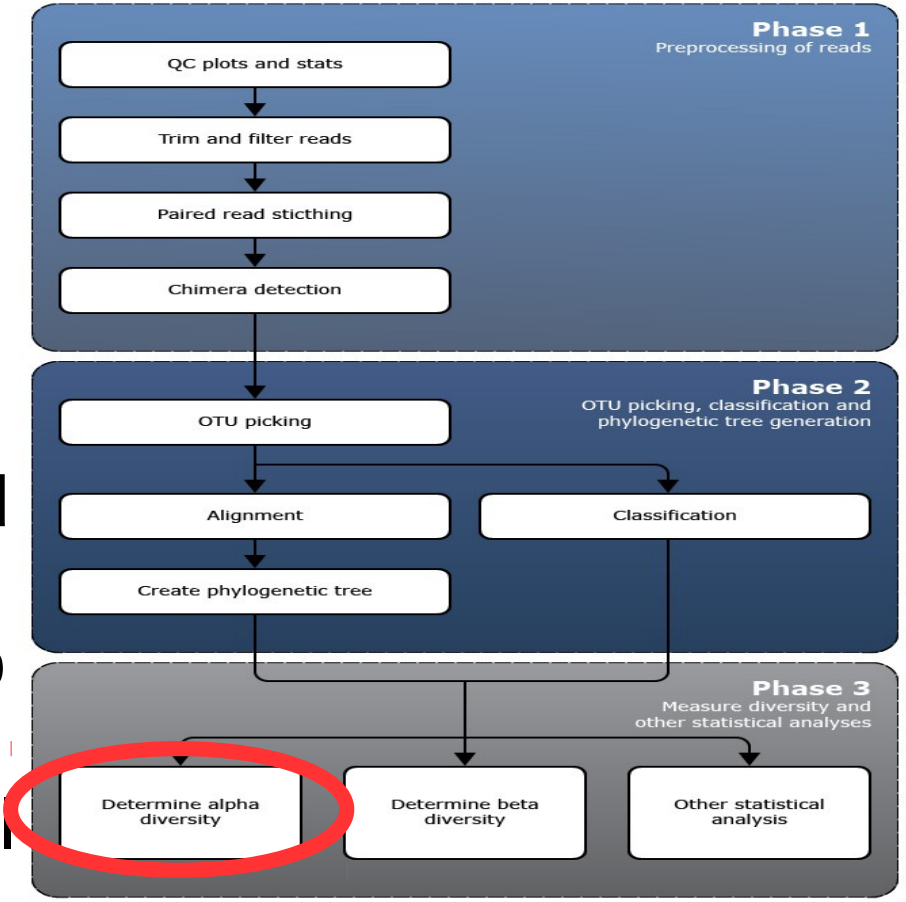
# Create a phylogenetic tree

- 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.



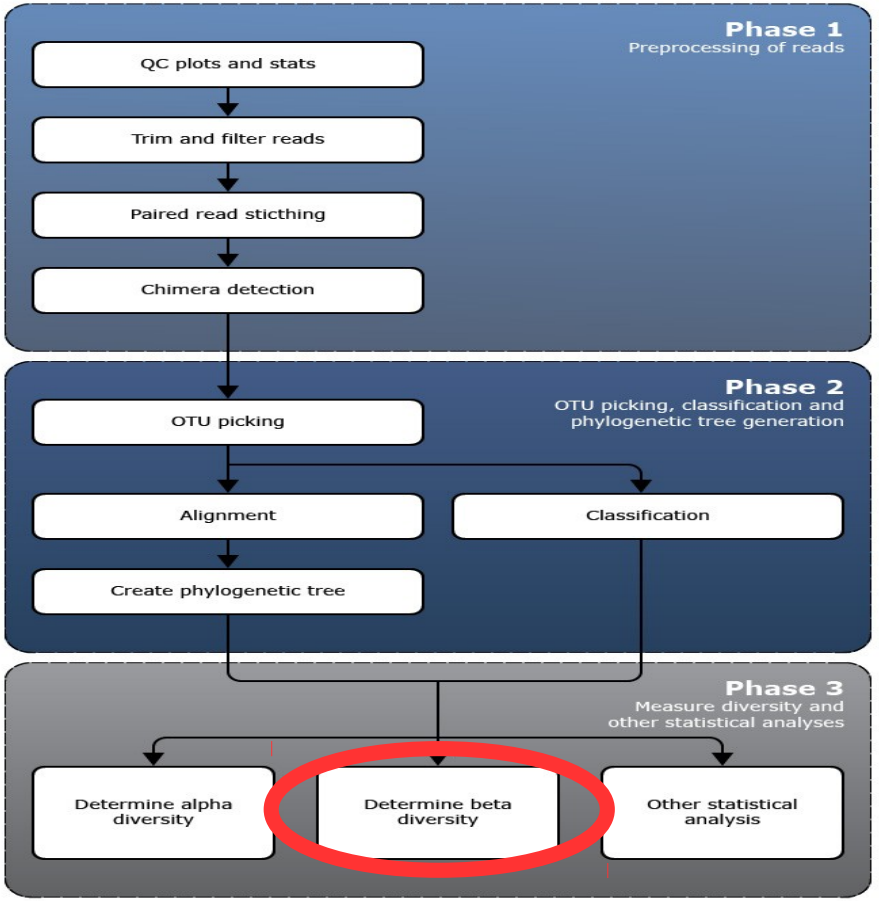
# Determine alpha diversity

- Alpha diversity is a measure of diversity within a sample.
- It gives an indication of richness and/or evenness of species present in a sample.
- Rarefaction analysis is required to understand the actual diversity within a sample and to determine if your sequencing effort is sufficient and if the total diversity within the sample has been captured.



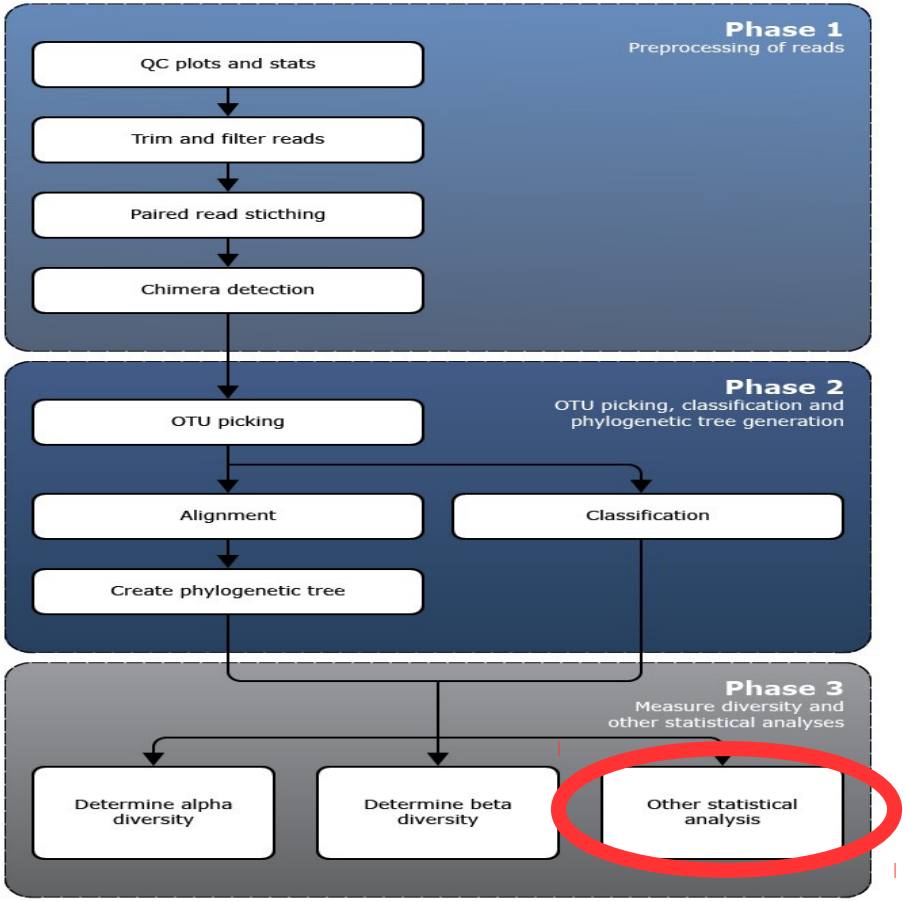
# Determine beta diversity

- Beta diversity is a measure of diversity between samples.
- One of the most commonly used metrics is the UniFrac distance that compares samples using phylogenetic information. An all-by all or pairwise matrix of the beta diversity metrics between all the samples in the study is generated and can be visualised in different ways such as a tree, graph, network, ordination methods.



# Other statistical analysis

- Additional statistical tests between samples or groups of samples can be done in QIIME, R, phyloseq and ade4 using
  - OTU tables
  - metadata
  - phylogenetic info



# Pipelines that can do parts of the workflow

- UPARSE: http://www.drive5.com/uparse/
- IM Tornado: http://sourceforge.net/projects/imtornado/
- QIIME: http://qiime.org/
- Mothur: http://www.mothur.org/

# Tool list

- FASTQC - http://www.bioinformatics.babraham.ac.uk/projects/fastqc
- PRINSEQ - http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi
- SolexaQA - http://www.biomedcentral.com/1471-2105/11/485
- PEAR - http://bioinformatics.oxfordjournals.org/content/early/2013/10/18/bioinformatics.btt593.full.pdf
- PANDASeq - http://www.biomedcentral.com/1471-2105/13/31
- FLASH - http://bioinformatics.oxfordjournals.org/content/early/2011/09/07/bioinformatics.btr507.full.pdf
- UCHIME - http://drive5.com/usearch/manual/uchime_algo.html
- ChimeraSlayer - http://nebc.nox.ac.uk/bioinformatics/docs/chimeraslayer.html
- Perseus - http://www.biomedcentral.com/1471-2105/12/38/
- UPARSE - http://www.drive5.com/uparse/
- UCLUST - http://www.drive5.com/uclust/downloads1_2_22q.html
- RDP classifier - http://sourceforge.net/projects/rdp-classifier/files/rdp-classifier/
- RTAX - https://github.com/davidsoergel/rtax
- PyNAST - http://www.ncbi.nlm.nih.gov/pubmed/19914921
- INFERNAL - http://infernal.janelia.org/
- FastTree - http://www.microbesonline.org/fasttree/
- IM-TORNADO - http://sourceforge.net/projects/imtornado/
- Mothur - http://www.mothur.org/
- QIIME - http://qiime.org/
- UNIFRAC - http://bmf.colorado.edu/unifrac/
- R packages
  - phyloseq - http://www.bioconductor.org/packages/release/bioc/html/phyloseq.html
  - ade4 - http://cran.r-project.org/web/packages/ade4/index.html