Functional microbiome research – bioinformatics section

Day 6 – Introduction to metabarcoding

We are on a new script today. Grab it with wget:

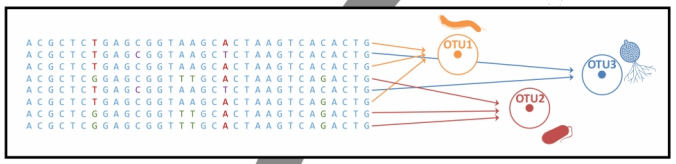
wget https://raw.githubusercontent.com/danchurch/FunctionalMicrobiomePractical2022/main/funmic2023/funMetabarcodeScript.txt

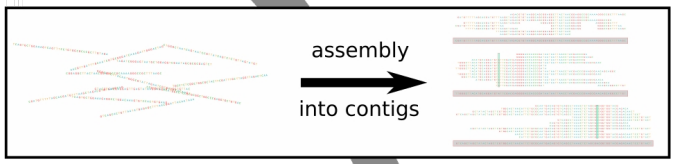
PCR amplification with barcode primers (ITS, 16S, ...)

Fragmenting and sequencing of all eDNA

Sequencing and identification of organisms using barcode sequence information

eDNA extraction



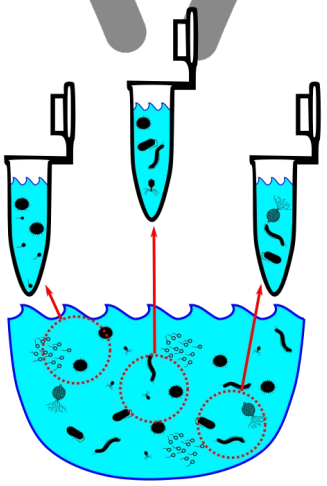


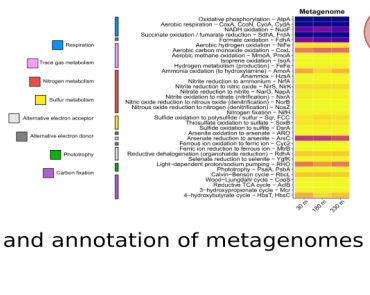
Metabarcoding

Beta diversity

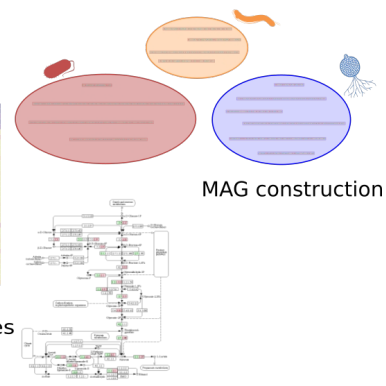
	Sam1	Sam2	Sam3
Otu1	0.34	0.32	0.29
Otu2	0.12	0.17	0.10
Otu3	0.07	0.03	0.11
Otu4	0.06	0.02	0.09

Taxonomy





Prediction and annotation of metagenomes



MAG construction

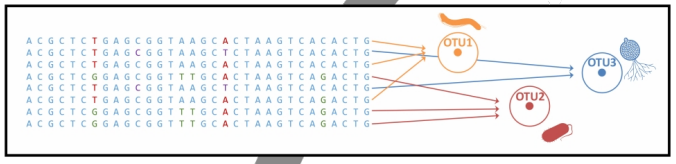
So far we have discussed the methods of getting sequence data out of an environmental sample, and processing that data to answer two questions:

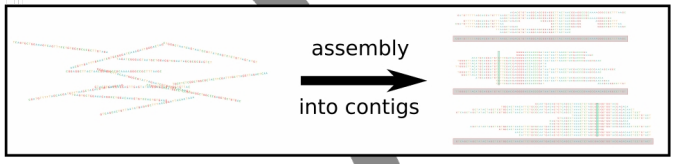
PCR amplification with barcode primers (ITS, 16S, ...)

Fragmenting and sequencing of all eDNA

Sequencing and identification of organisms using barcode sequence information

eDNA extraction





Metabarcoding

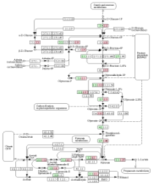
Beta diversity

	Sam1	Sam2	Sam3
Otu1	0.34	0.32	0.29
Otu2	0.12	0.17	0.10
Otu3	0.07	0.03	0.11
Otu4	0.06	0.02	0.09

Taxonomy

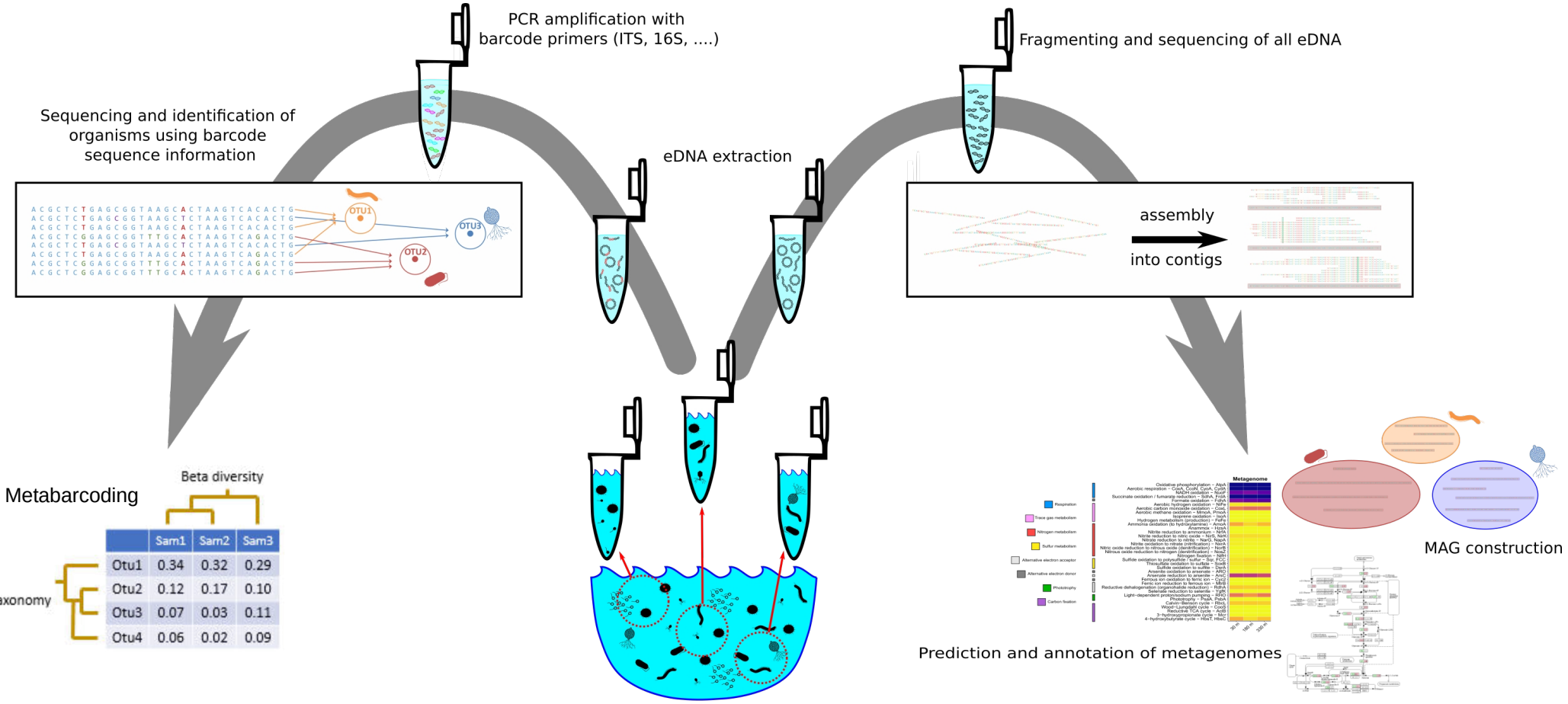
Prediction and annotation of metagenomes

MAG construction



So far we have discussed the methods of getting sequence data out of an environmental sample, and processing that data to answer two questions:

What can they do?



Who is there?

So far we have discussed the methods of getting sequence data out of an environmental sample, and processing that data to answer two questions:

What can they do?

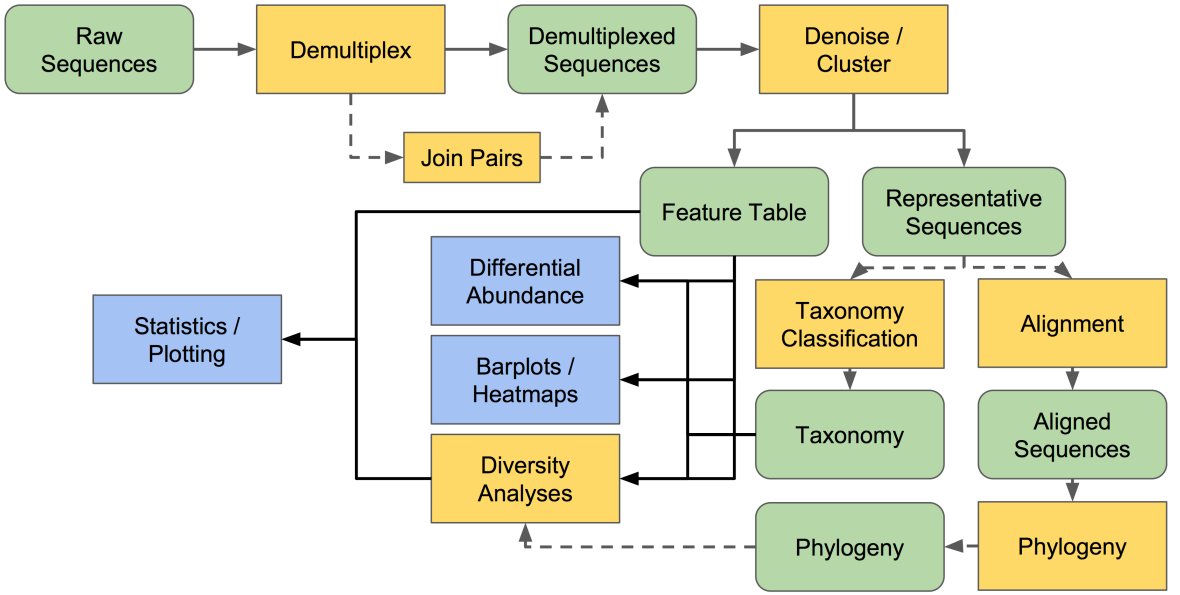
Metabarcoding

Regardless of barcode chosen, preliminary bioinformatics pipelines usually have the following basic steps:

1) Quality control and filtering
2) Denoising and clustering
3) Taxonomic assignment

Metabarcoding

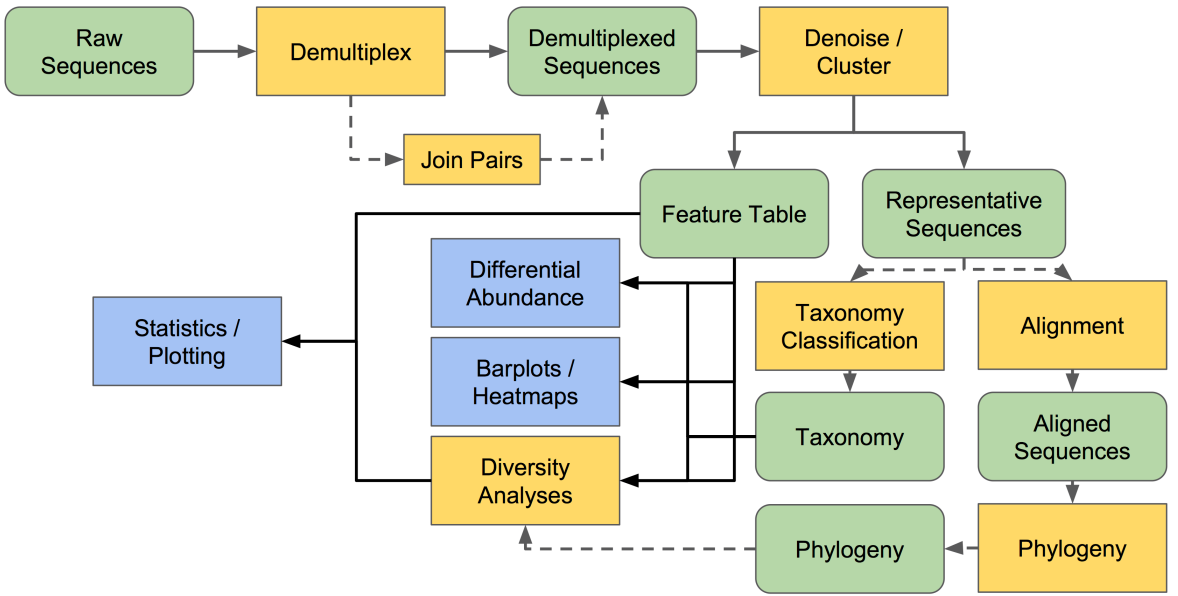
We will be doing most of these steps inside the Qiime2 environment.



https://docs.qiime2.org/2021.8/tutorials/overview/

Metabarcoding

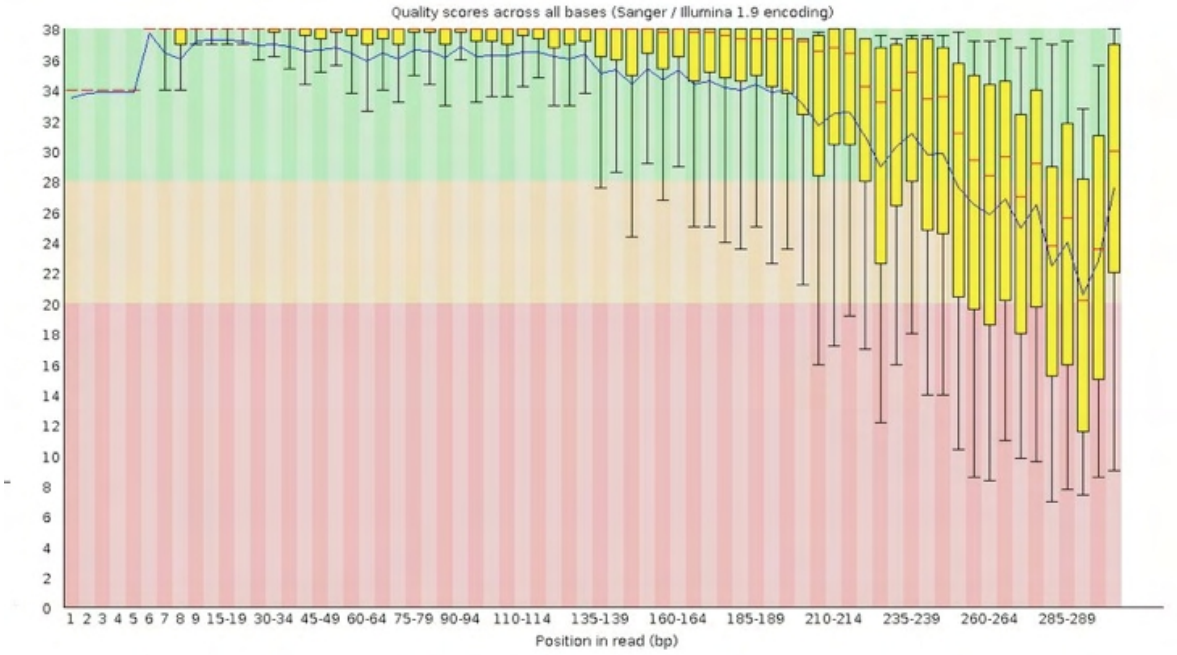
We will be doing most of these steps inside the Qiime2 environment.



https://docs.qiime2.org/2021.8/tutorials/overview/

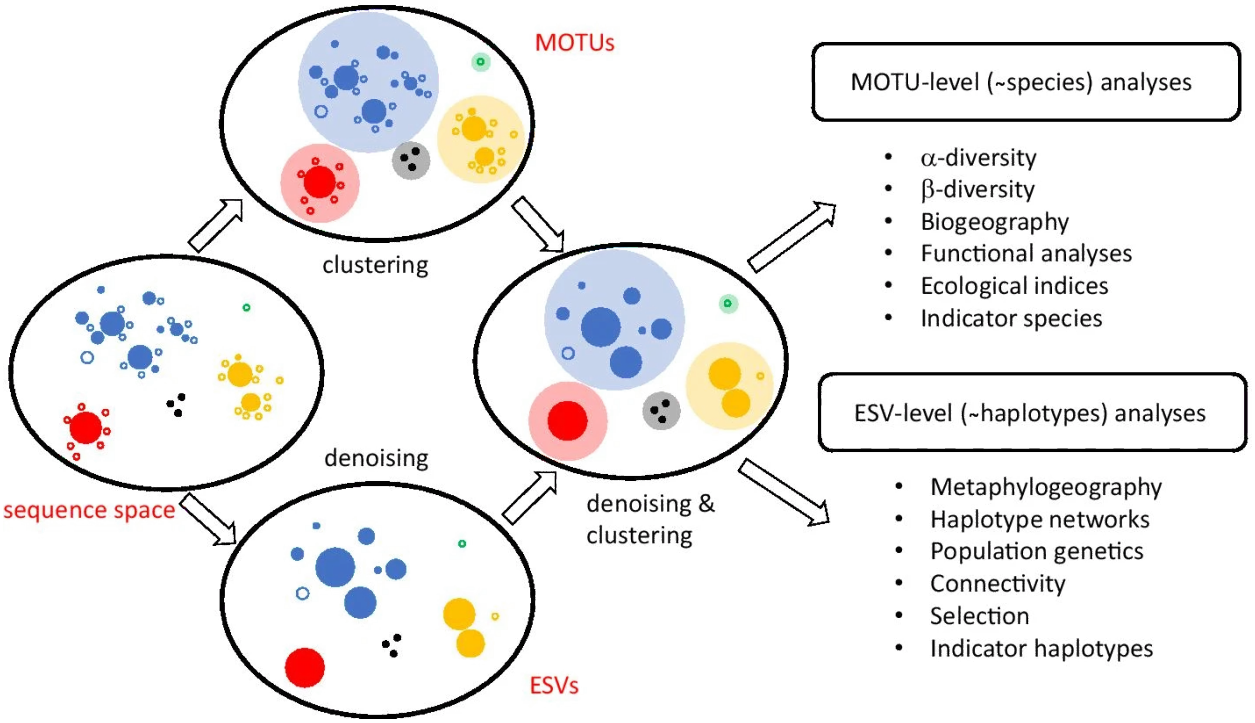
Qiime was one of the first comprehensive pipelines for handling metabarcode data, and remains one of the most popular.

Metabarcoding



1. Real life barcode data are messy, so raw sequence files have to be examined. Low quality reads are dropped. Non-target end sites (primers, adapters, etc) are trimmed, as are low quality regions on either end (phasing!)

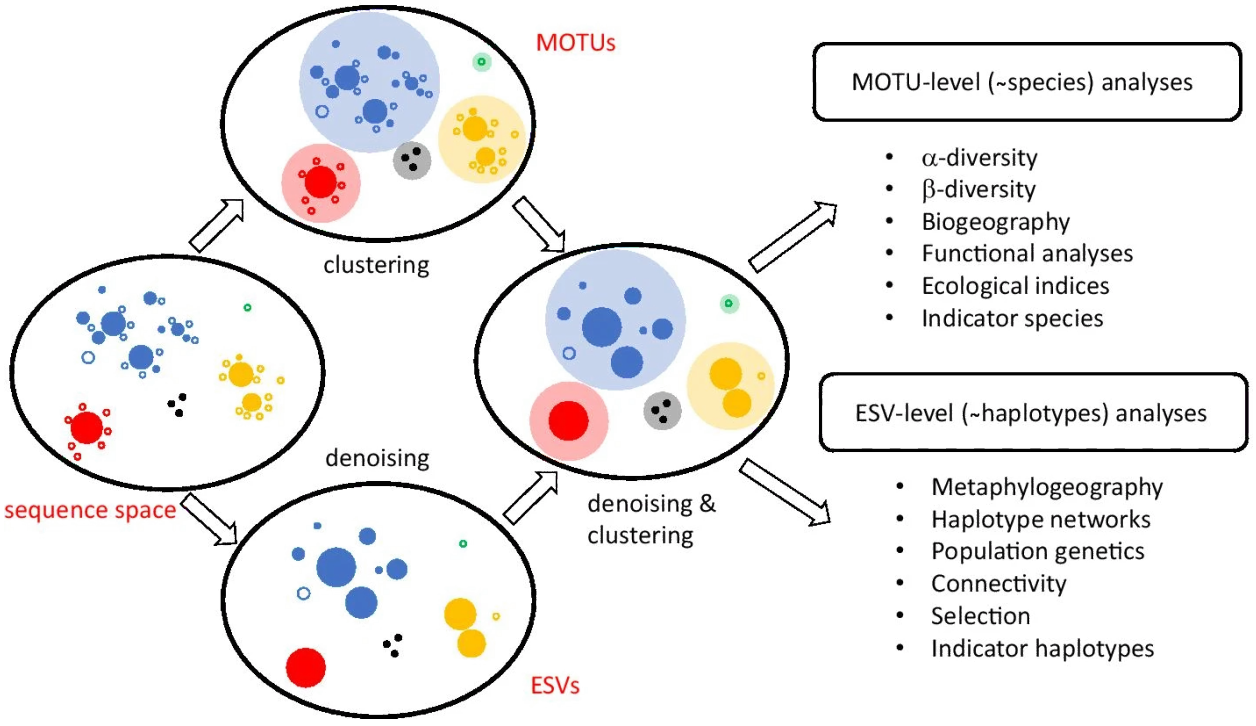
Metabarcoding



(Antich et al. 2021)

2. Next reads are grouped together. Clustering of highly similar reads is done because variability in barcodes occurs from at least two sources: sequencer error (technical) and natural variation within species (biological).

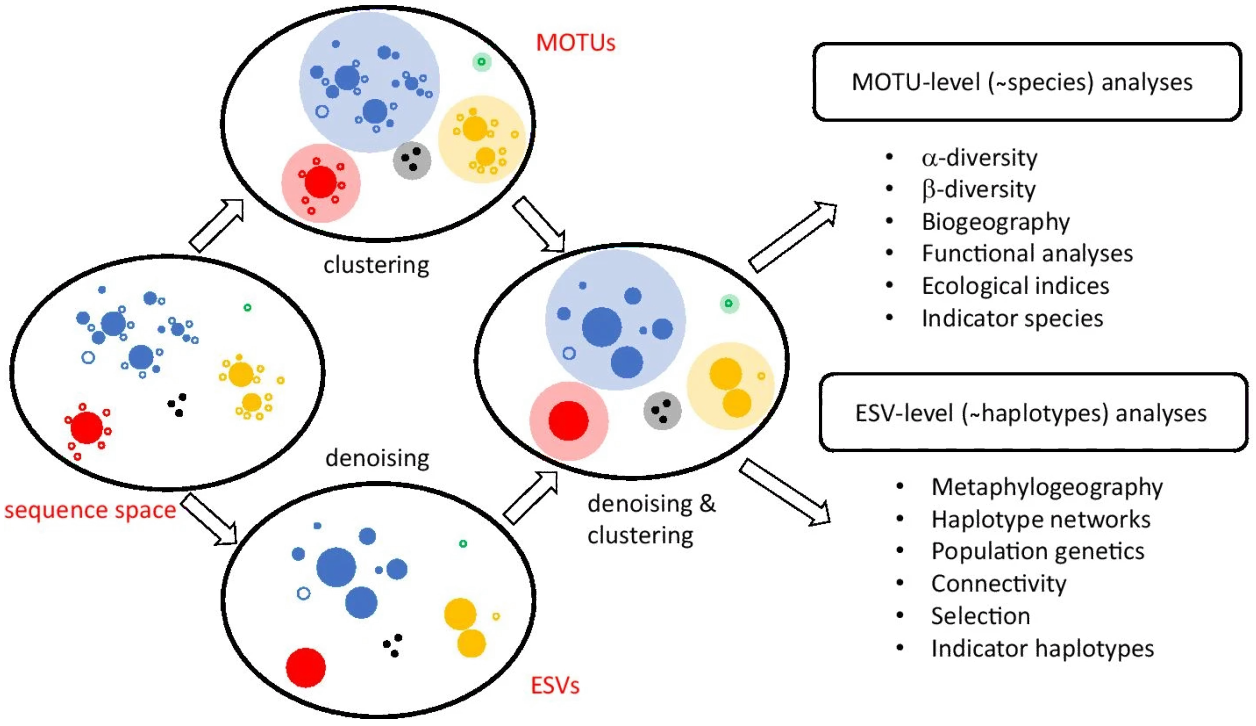
Metabarcoding



(Antich et al. 2021)

Sequencer error (technical variation) can sometimes be reduced through a strict zero-radius clustering process called de-noising.

Metabarcoding



(Antich et al. 2021)

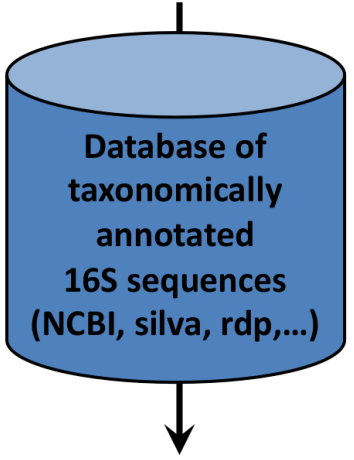
Denoising is usually done by machine-learning algorithms that characterize the types of errors in an individual run. The results are called “ASV’s”, “ESV’s”, “ZOTU’s”, etc. etc. These methods are best developed for Illumina data.

Metabarcoding

The end result of clustering (and denoising) are OTUs (operational taxonomic units). These are usually supposed to represent species- or genus-level groups. A representative consensus sequence for each OTU is generated.

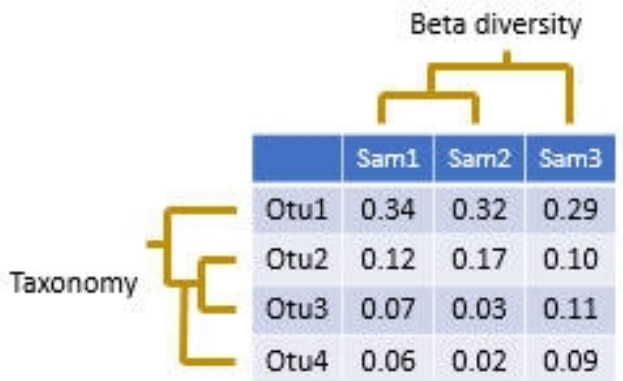
Metabarcoding

A C G C T C T G A G C G G T A A G C A C T A A G T C A C A C T G
A C G C T C T G A G C G G T A A G C T C T A A G T C A C A C T G
A C G C T C G G A G G G G T A A G C A C T A A G T C A G A C T G



3. This consensus sequence from each OTU can then be aligned against reads from well-studied organisms to find likely taxonomy for the OTU – who is it?

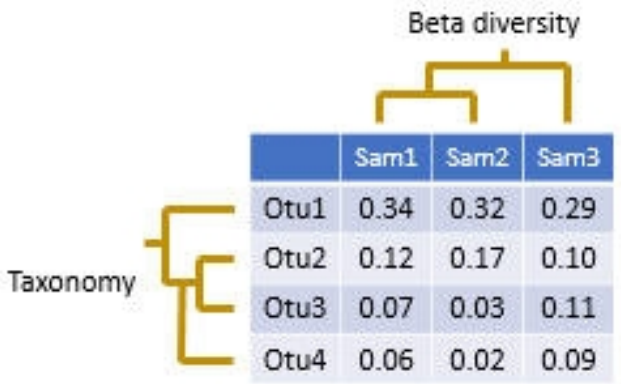
Metabarcoding



	Sam1	Sam2	Sam3
Otu1	0.34	0.32	0.29
Otu2	0.12	0.17	0.10
Otu3	0.07	0.03	0.11
Otu4	0.06	0.02	0.09

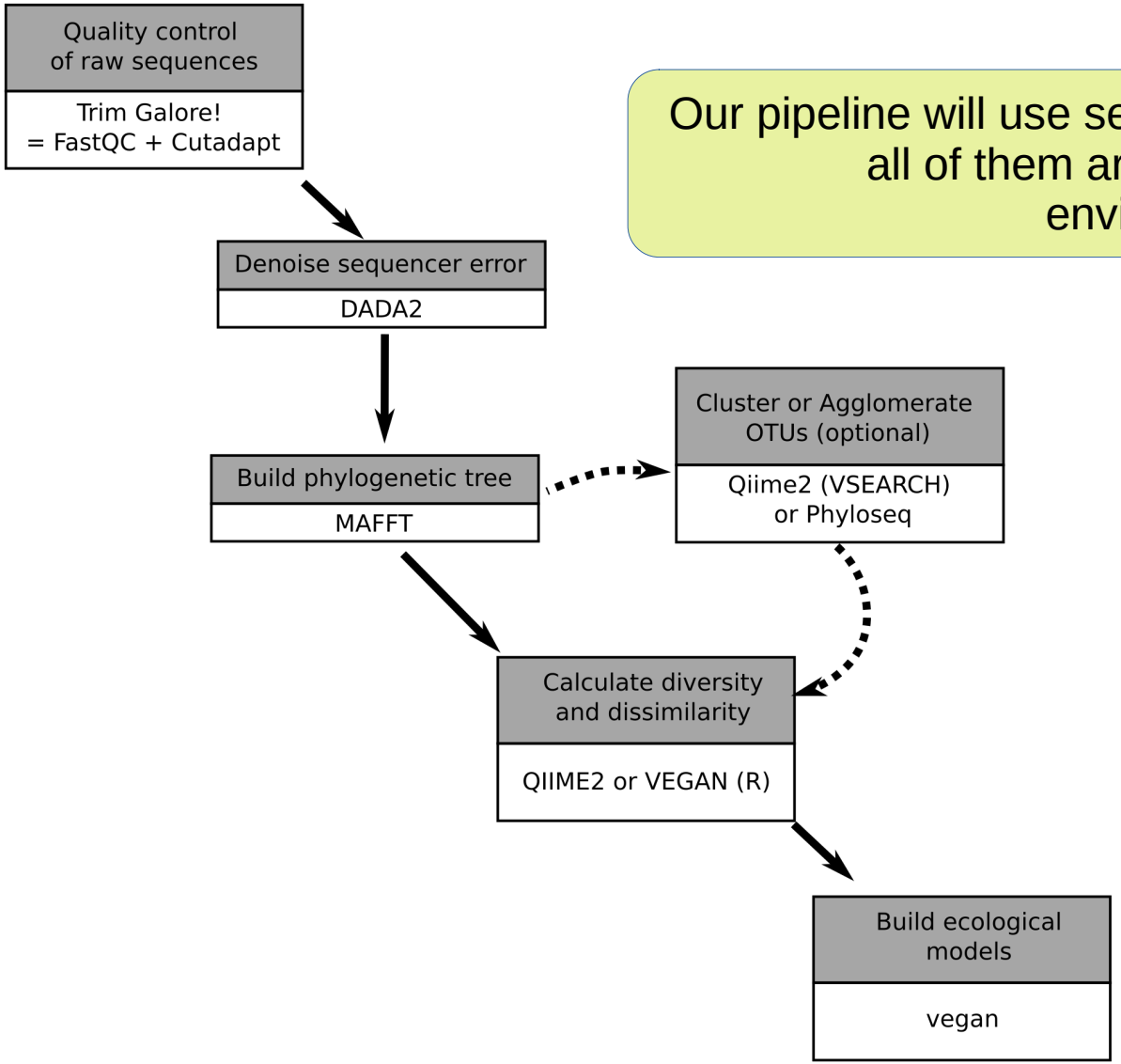
We also map the member reads back to this consensus senquence/taxonomy for abundances, and note which samples contain this OTU. The result is our **community matrix**, the basic unit of all multivariate community analysis.

Metabarcoding

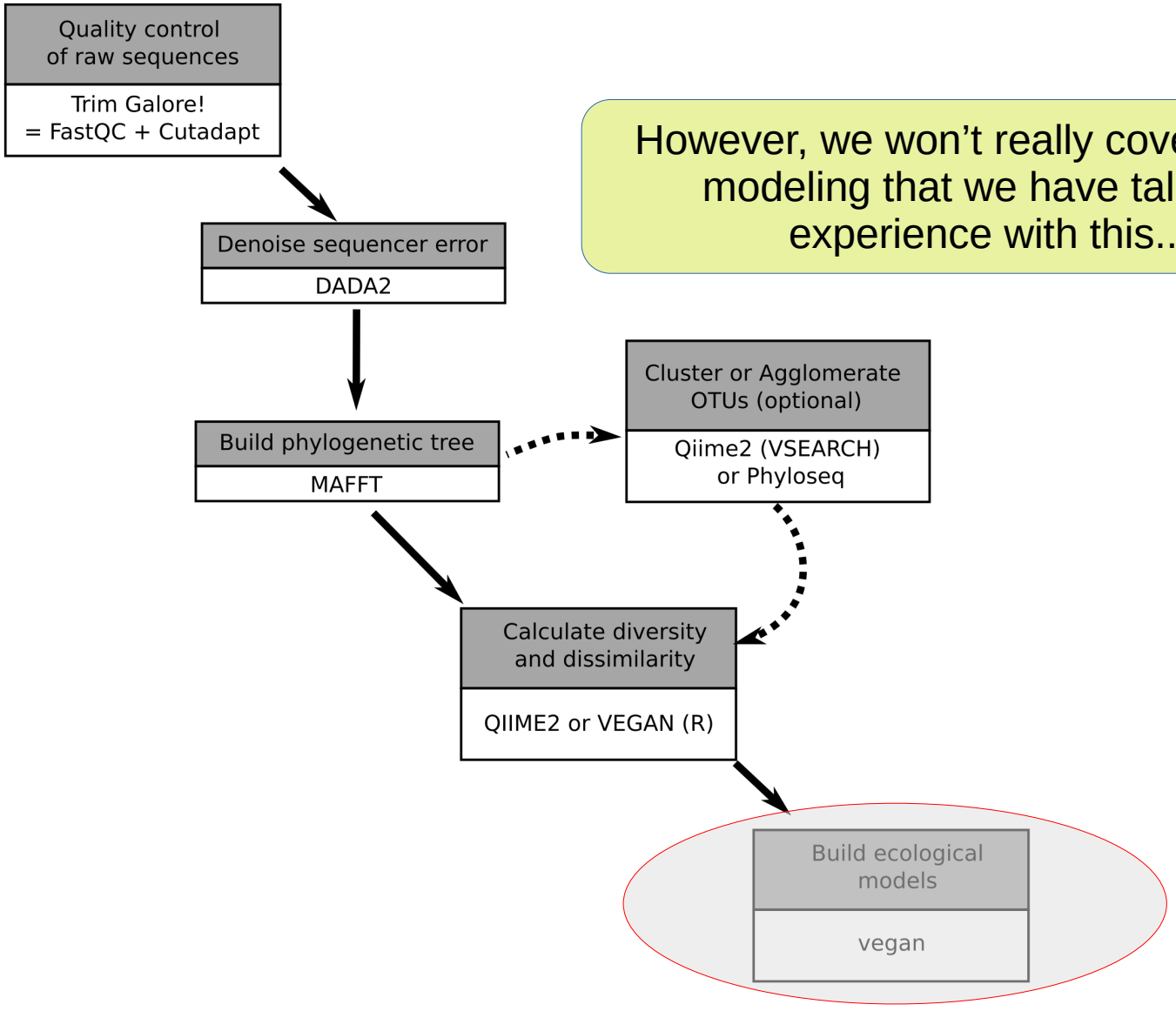


	Sam1	Sam2	Sam3
Otu1	0.34	0.32	0.29
Otu2	0.12	0.17	0.10
Otu3	0.07	0.03	0.11
Otu4	0.06	0.02	0.09

With our OTU table, and our taxonomic classification, and with phylogenetic information, we can begin to analyze alpha and beta diversity, make ordinations, and make models.



Our pipeline will use several software packages, but almost all of them are wrapped into the Qiime2 environment/pipeline.



However, we won't really cover in-depth the mathematical modeling that we have talked about lecture. To get experience with this...find a research lab!