Vocabulary

Amplicon

A piece of RNA or DNA that is the result of an amplification or replication process. In sequencing, amplicons are generally the products of polymerase chain reaction (PCR) and are usually phylogenetically informative.

Examples: 16S, 18S and 23S.

Barcode

In multiplex sequencing, where multiple samples are pooled for sequencing, a barcode is a short DNA sequence added to a DNA fragment prior to sequencing to indicate its sample of origin.

Demultiplexing

The sorting of sequences into samples of origin based on barcode sequences

OTU (**O**perational **t**axonomic **u**nit)

Cluster of reads that are above a threshold of similarity to one another. In 16S sequencing, reads are generally clustered at 97% similarity to produce “species”-level OTUs.

ASV (**A**mplicon **s**equence **v**ariant)

ASVs are meant to represent sequences truly found in nature and may differ by as little as a single nucleotide. The process of defining ASVs includes the removal of sequences that are thought to have arisen due to sequencing error.

Metadata

A set of data that gives information about other data. Barcode sequences are examples of metadata, as they describe something about the sequencing data (sample of origin).

Zero

Believe it or not, zeros don't always mean zero. Zeros can actually indicate three different things:

- **True zeros:** A feature (OTU/ASV) is missing from a sample.
- **Rounding (Sampling) zeros:** A feature exists below the detection limit.
- **Count zeros:** A feature exists in a sample, but counting was not thorough enough to observe it.

In sequencing data, zeros are considered count zeros, meaning we assume that any sequence we see ≥1 time could be seen in another sample if sequenced with enough depth.

Rarefaction

A statistical method used to estimate species richness based on the number of individuals sampled. Rarefaction is not recommended for sequencing data.

Alpha diversity

The diversity held within a single sample (local diversity).

Beta diversity

The spatial variation of diversity. Cannot be defined outside of the concept of diversity across space.

Gamma diversity

The species diversity across all samples in a sequencing study (regional diversity).

Phylogenetic diversity

A measure of diversity that incorporates the phylogenetic distance among species in a sample or unit of interest.

Community

The collection of all populations occupying a geographic area at the same time.

Population

A group of organisms of the same species that live in the same area and interbreed.

Compositional data

Data that only carry relative information and often are constrained to a specific sum (e.g. 100%). Sequencing data are compositional because the sequencer imposes an artificial constraint on counts, as sequencers are not capable of sequencing every DNA strand in a sample. This also means that sequencing a particular strand of DNA prevents another strand of DNA from being sequenced, meaning that sequencing counts are not true counts.

One of the better known examples of compositional data is soil composition. All soils are made up of some percentage of sand, clay, and loam which must add up to 100%. If one soil has proportionately more clay, that automatically means that it must contain less sand or loam.

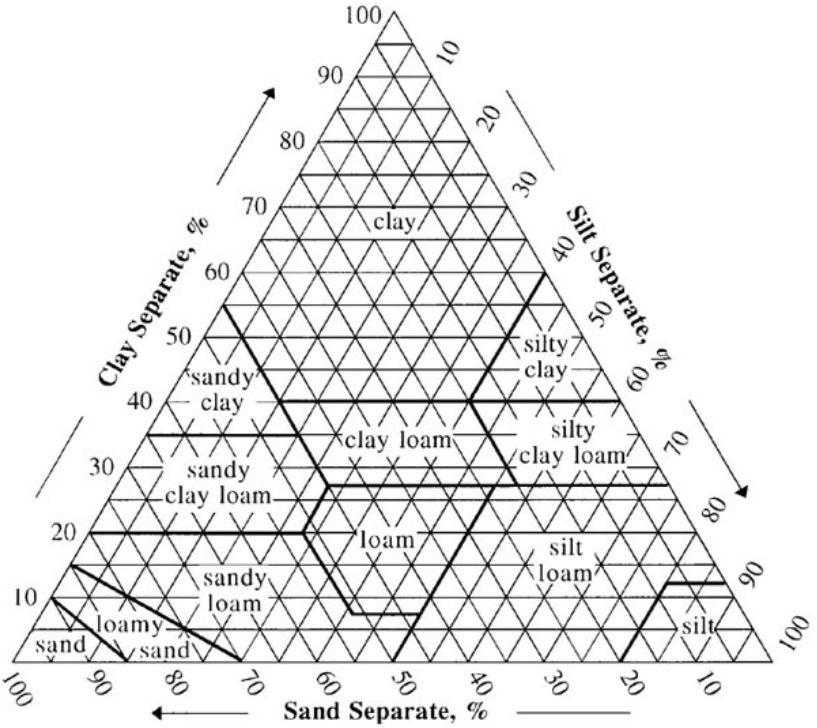

Figure courtesy of the USDA [Natural Resources Conservation Service]