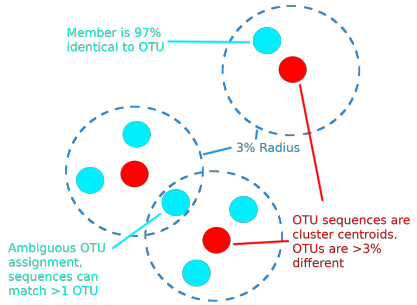
- Sanger Sequencing 1977
- Barcoding and taxonomic reference libraries
- Next Generation Sequencing
- Costs go down by orders of magnitude

Genomics; A new hope?

- The invention in 1977 of Sanger-based DNA sequencing3 which revolu- tionised all branches of the biological sciences, could not be used for environmental bulk samples because they contained potentially thousands of species, and separating them for sequencing was prohibitively difficult.
- It enabled the development of Barcoding => identifiying species by small segment of slow varying dna => Build taxonomic libraries based on barcodes
- NGS method of high-throughput multi species identification using degraded DNA found in the environment (
- Costs went down by orders of magnitude and made DNA sequencing available to ecological studies

└Operational Taxonomic Units



Sequncing => OTUs; genetic proxies for species. Sequences clustered with 97% similarity and are given a signle otu name and taxonomy. Number of sequences in an OTU read => read counts

Machine learning for ecosystem assessment in the Amazon, using environmental DNA data

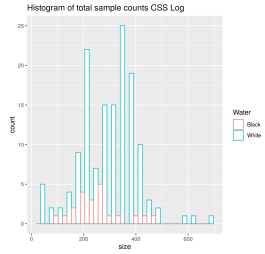
Data

Sampling

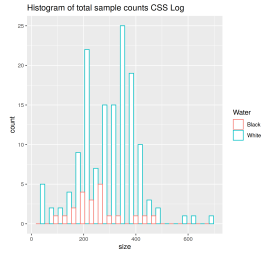
- Collected 164 samples from Northern Peruvian rivers
- Sequenced using Next generation sequencing
- 675 OTUs matched with taxonomies from reference libraries
- Discarded the rest
- Recorded colour of water and location

- Collected 164 samples from northern peruvian rivers
- The rivers can be seen in this photo highlighted by the yellow colour
- Most abundand sequences in each otu were matched with a taxonomy. OTUs without one where discarded
-
- We also collected metadata about each sample. This included water colour and location of the sample

- This is an example of Black and white water in the amazon. Essentially coffee with and without milk. Different concentration of minerals will inevitably lead to different species inhabitting the waters. It is our response variable; Can we predict water colour from species composition?
- Unbalanced data set;143 white 21 black, most black water samples in the east



- Ended up with 164 samples and a totla of 675 OTUs. Sparse matrix mostly zeroes. Thought appropriate to use dimensionality reduction
- used oridnation methods => common in ecology for uncovering patterns in data
- Principal components analysis finds axis with max variance and then orthogonal with second most
- PCoA is generalisation of PCA. We can calculate dissimilarities between samples using distance metrics other than euclidean and then perform eigenanalysis on the distance matrix of the samples rather than covariance matrix. Used bray curtis commonly used in ecological studies.



- nmds preserves order relations in dissimilarities between samples in the projection of data. EXAMPLE the pairs of closest samples have to also be represented by the smallest interpoint distance in the ordination projection.
- high correlation
- high variance in total read counts per sample. Used css normalisation (common method) to reduce variance. Log reduces further

- Location of train and test sets will affect performance of classifiers.
- Splitting data simulates different learning conditions .
- Maximally similar
- Maximally dissimilar
- Random with equal distributions of black water in all sets

- TEsting the classifiers will be affected by the location of the samples in the training and test sets.
- Splitting the data set in different ways is akin to testing the classifiers under different consitions
- e.g. In maximally similar setting we simulate testing the models in geographical regions they have already seen
- In maximaly dissimilar, we do the exact opposite by isolating the test from train
- we also tried random sampling