# Plan van Aanpak

## Naturalis

Naturalis Biodiversity Center is een onderzoekinstelling en natuurhistorisch museum in Leiden. Naturalis is leider van een grootschalig infrastructuurproject met het doel om alle meercellige organisme in Nederland te kunnen identificeren aan de hand van een van de drie modaliteiten: beeld, geluid, of DNA/ In het kader van dit ARISE-project zal ook op grote schaal metabarcoding gaan plaatsvinden, bijvoorbeeld van bodemschimmels. Dit is waar mijn afstudeerstage over zal gaan de komende maanden. In eerste instantie zal dit gaan om het herkennen van soorten, maar met deze samples kunnen ook gegevens worden verzameld over de biodiversiteit op gemeenschapsniveau. Bijvoorbeeld: hoe fylogenetisch/taxonomisch divers is een sample ten opzichte van een andere (genormaliseerd voor PCR biases etc.), m.a.w. de alfa-diversiteit, die met een aantal statistieken – zoals fylogenetische dispersie - uitgedrukt kan worden. Of, wat is de fylogenetische turnover tussen samples, dus beta-diversiteit, bijvoorbeeld uit te drukken in UNIFRAC-afstand. In deze opdracht gaat de student een workflow prototyperen om deze indices uit te rekenen. De pijplijn loopt vanaf ruwe data (FASTQ), die dus met command-line bioinformatics opgewerkt moet gaan worden (shell/python), tot de berekeningen van de indices (R).

Korte beschrijving van een stappenplan:

* Pre-processing van NGS metabarcoding data (trimmen, filteren, read merging)
* Mappen van reads tegen referentiebackbone (UNITE e.a.)
* Uitrekenen van de alfadiversiteitsindices in R, zodat voor elke samples een of meer metrics beschikbaar zijn vwb de diversiteit binnen de sample
* Uitrekenen van de betadiversiteitsindices in R, zodat de turnover (of juist overeenkomst) tussen de samples in pairwise distance uitgedrukt kan worden, bijvoorbeeld om zo samples te classificeren

## Eerste week

* Schrijven van een plan van aanpak
* Wat voor data ga ik werken?
* Galaxy porter onderzoeken

Stappenplan uitwerken

* Pre-processing van NGS metabarcoding data (trimmen, filteren, read merging)

Metabarcoding is de barcode van DNA/RNA op een manier die de gelijktijdige identificatie van veel taxa binnen hetzelfde monster mogelijk maakt. -> Hoe ziet dit soort data er precies uit? Volgende stap is om te bepalen hoeverre we het willen trimmen en filteren; wat voor soort outliers oid

Metabarcoding is paired end reads -> merge van overlapped reads?

First step is quality control.

Next trimming methods.

I’m now looking at the rougly made plan and work out each of the steps more detailed. Also looking for some scientific articles.

Now a’m looking at the step for pre-processing the NGS metabarcoding data, so the quality control, trimming methods ect. I want to work that out this week or beginning next week beaceaus

Alfa-beta diversity

Galaxy workflow system

Ruwe reads met its dna, misschien multiplexen oid trimmen <- misschien moet dat gebeuren maar niet de scope van mn project.

Specifieke wat doe je met metabarcoding data. Clusteren? (op een soort soorten)

Of AS feas (alle unieke reads)

Hoe kan je de diversiteit uitrekenen <- inlezen na woensdag

Splitsing of allemaal in r doen <- browser / of misschien in galaxy

Pubmed te lezen <- dag 4 week 1

<https://github.com/naturalis/mebioda/tree/master/doc/week1>

<https://github.com/naturalis/mebioda/tree/master/doc/week1#day-4---metabarcoding>

ZENODO.org <- data met elkaar delen

## Pre-processing NGS metabarcoding data

The first step in pre-processing the NGS metabarcoding data is to look at the data. What is the quality of the data, striking features of the data

Filtering low quality reads, overlapping short reads, trimming low-quality ends -> using HTS tools

Workflow

* Clip any synthetic oligonucleotides (adaptors/ primers)
* Trim loq quality bases, filter short reads
* Novo / or mapping assembly
* Further annotation
* Variant calling
* Consensus sequence computation

Onderzoeken:

UNITE

Rarefaction -> hoe werkt dit

Unifrac distance

Bray-curtis distance

Hoe werkt mantel test

Species acummulation curve

UNITE is een database en sequentiebeheeromgeving. Er is een referentiedataset waar je zoekopdrachten aan kan geven ; alleen bodemschimmels is optie.

UNITE is gebruik op de rauwe data -> misschien staat hierover nog iets in verslag van Sophie.

De resultaten hiervan bestuderen

https://www.cd-genomics.com/microbioseq/rarefaction-curve-a-measure-of-species-richness-and-diversity.html

Rarefaction kan worden gebruikt om re bepalen op een specifiek monstel voldoende is gesequenced om zijn identiteit te bepalen.

Als een bepaald sample super vaak is gesequeced en een ander niet dat klopt niet, dus dan ga je over in rarefaction. Je gaat de laagste (of de ene laagste als je de laagste laat vallen) hoeveelheid aantal sequences doen bij alle samples. Je verliest zo veel informatie helaas, en je riskeert dat je unieke feature verliest. Het moet helaas want anders kan je geen diversiteit berekenen. Leuk zijn om iets te bedenken waardoor dit niet nodig is.

Unifrac distance is een afstandsmetriek die wordt gebruikt voor het vergelijken van gemeenschappen (branch lenghts ect). <https://rdrr.io/cran/abdiv/man/unifrac.html>

Bray-Curtis distance: The Bray-Curtis dissimilarity is always a number between 0 and 1. If 0, the two sites share all the same species; if 1, they don’t share any species.

Mantel test: testen van de correlatie tussen twee matrics -> <https://jkzorz.github.io/2019/07/08/mantel-test.html>

Species accumulation curve -> hoeveel species aan de ene kant en hoeveel individuen aan de andere kant.

Ik heb Sophie haar scriptie gelezen met het oogpunt dat ik hier een vervolg onderzoek op ga uitvoeren. De data zijn verzameld en (minimaal) geanalyseerd, er zijn veel vervolg onderzoeken mogelijk op de data. Of dit mij 8 maanden bezig gaat houden weet ik niet super goed.

Als vervolg hierop zou ik meer onderzoeken kunnen uitvoeren op de biodiversiteit van de samples; een rarefaction curve, unifrac distande, bray-curtis distance, mantel test ect (dit zijn ook de zoekopdrachten die Rutger mij heeft gegeven).

Een pipeline die nog een minimale kwaliteitscontrole uitvoert door de unidentified ect eruit filtert, een bestand maakt om een rarefaction curve mee te kunnen maken (inkorten van veel samples -> verlies van data wil je dit ook voor de andere onderzoeken gebruiken). Diepere kijk in de alpha en beta diversiteit.

**Data-analyse**

De beschreven statistische toetsen die uitgeoefend worden op de data worden uitgevoerd in het programma R. De DADA2 ITS pipeline workflow die wordt toegepast is ook een R workflow. Voor het voorbereiden van de data voor de rarefaction curve zal er gebruik worden gemaakt van de bash commandlines. Voor het ontwikkelen van een rarefaction curve zal Rarefy worden gebruikt6. Rafefy is een R package. Met Rarefy is het mogelijk een rarefaction curve maken van de alfa en betadiversiteit, maar ook diverse andere diversiteitsmetriek of het voorspellen van fylogenetische waarden. Voor het ontwikkelen van de UniFrac distance wordt een R package gebruikt genaamd phyloseq7. Voor het uitvoeren van de Mantel test zijn meerdere R packages nodig. De vegan package en de geosphere package. De geosphere package geeft je de mogelijkheid om de afstand te bekijken tussen twee samples uit hetzelfde monster.