**16S rRNA gene diversity and multiplicity**  
Bachelor project – 20 ECTS

Author: Lasse Schnell Danielsen

Supervisor: Mikael Lenz Strube

Et billede, der indeholder plante

Automatisk genereret beskrivelse

# Introduction (background information)

# Methods and materials

## Analysis of WGS data

### Collecting the sequences

### Running ribdif on the sequences

### Running abricate on the sequences

### Selecting a measurement of the diversity in the sequences

* PCA
* Entropy vs Information content vs nt subs

## Getting information from databases

### NCBI

### Bacdive + Growthrate data

## Joining the datasets

|  |
| --- |
|  |
| **Figure X. Flow diagram of workflow for gathering the data** |

# Results and analysis

## Data exploration

### Data quality and missing Data

* Unequal data distribution
  + What
    - As can be seen from figure X, there is an uneven distribution of phylums in the dataset.
    - This is a problem since this will introduce a qunativtative bias toward heavly sequenced species
  + Why?
    - Some species are sampled more
      * Since they are more relevant to the scientific community
    - Bias towards culturable species
      * Culture -> WGS (?)
  + How to fix?
    - We can decrease the effect by joining at a species level. But we lose information.
    - The effect is present no matter how high we join phylogenetically
    - We begin losing a ton of information by going to the genus level
      * As can be seen from the change in variance at this level

|  |
| --- |
|  |
| **Figure X. A: Distribution of the entropy across the sequences. B: 16S gene copy number for different phylum:** |
|  |
|  |
|  |

|  |
| --- |
|  |
| **Figure X. Overview of selected attributes and their missing values:** |

* The join
  + The data was joined
    - We lost X amount of data
* Missing data
  + Intro
    - Since the dataset was joined based on the WGS data
    - We have 16S gene data on all
    - But not from bacdive
  + In the plot we have
    - Missing values of selected attributes from bacdive
  + We can observe
    - There is missing data for a lot of attributes.
  + Some of the attributes
* This means
  + It will be hard to make an analysis which involves all attributes explaining the environment.
  + Therefore, I will instead analyze some of them separately and them together when it makes eg. now using PCA, with the ones without too many missing vcalues

### PCA

|  |
| --- |
|  |
| **Figure X. A: PCA of selected attributes for different phylia, colored by order:** |

* The first two principal components explain little of the variance
  + Many of the attributes are probably independent
  + This aligns well with the article describing that there is little evidence for direct relationship between ecological information.
* A few of the principal vectors(?) overlap
  + This means that they might be able to explain some of the same ecological information.
    - Many of them make sense
      * Genes\_nc and n16
      * Total\_Seq\_length – pseudogenes – coding genes
* We can distinguish somewhat between some phylums
  + For example Actinomycetota and Bacillota
* For the different phylum we can generally distinguish between genera.
  + This is especially clear in Actinomycetota and Bacillota
    - Here the distinction is attributed to some being sporeforming
    - Lastly it is interesting that we based on these two phylum and these attributes, probably could build a simple ML model, which given a random bacteria could predict whether it was sporeforming or not
    - Based on the principal components
      * Sporeforming seems to have higher n16 + nc
      * Longer genes +
      * Optimal temps in comparatively lower temperatures
  + This I relevant as the different bacterial genera often share similar ecological lifestyles.
    - Therefore we can use this collection of data to get an idea about the different bacteria’s lifestyle
    - And see if they correlate with the number of 16S gene copies and div
  + As an example can be seen the plot below, how some lifestyles are isolated in the PCA
  + Lastly the same information is practically present in the plot of the 16S genes against sequence length. Here we can observe, as described by PC1, that small genomes are present with high temperatures. Which can also be seen from the data. This makes sense ecological, as there I less need for variation here.
  + Et billede, der indeholder diagram

    Automatisk genereret beskrivelse
  + Furthermore we can see that large genomes tend to have larger GC%. And interestly enough that orders with large n16 tend to have lower GC%.
    - Ofc there is effect of high temp living have lower n16 content
    - But it is also general
    - Based on this corr we can nearly estimate n16 based on seq length and GC%
    - Et billede, der indeholder diagram

      Automatisk genereret beskrivelse

|  |
| --- |
|  |
| **Figure X. A: PCA of selected attributes for different phylia, colored by order:** |

# Discussion

# Perspectivation

## Other notes

If we see trends across pylum it is not random due to tax

* Try all div methods, and use PCA to see which effect they describe