Lecture 12 and 13

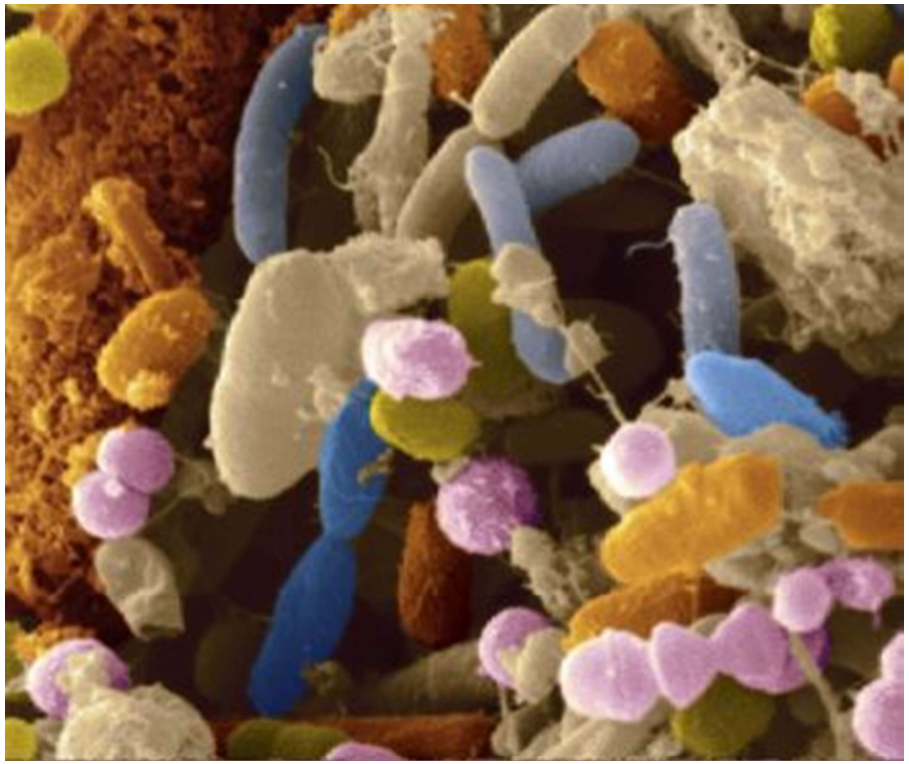
Introduction to bioinformatics (MVE510)

Autumn, 2020

Additional reading: The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Escobar-Zepeda A, de León AVP, and Sanchez-Flores A. Frontiers in Genetics, 6 2015.

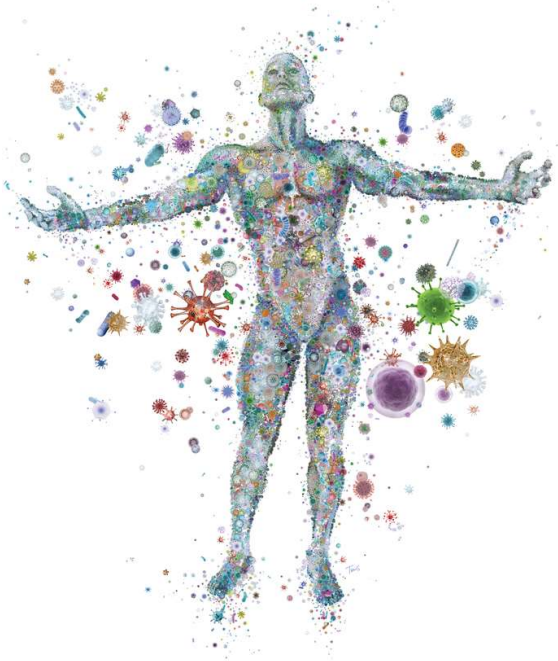
Microorganisms

- Present in every habitat and are integral members in almost every ecosystem on Earth
- The number of bacterial species is estimated to 10^7-10^9. Only a small proportion of these has been isolated and named



Microorganisms and their diversity

Human



1,000 bacterial species

Soil

10,000 bacterial species

Ocean

10,000 bacterial species

Metagenomics

- Microorganisms are often organized in complex communities
- Metagenomics is the study of the metagenome, which is the collective genome in a microbial community

- In metagenomics, DNA is randomly sampled from the metagenome which is used to derive information about the present microorganisms and their biological functions.

ACTCCTACGGGAGGCAGCAGTGGGGAATTTTGGACAATGGACGC
AAGTCTGATCCAGCCATTCCGTGTGCAGGACGAAGGCCTTCGGA

GTTGTAAACTGCTTTTGTACAGAACGAAAAGGTCTCTATTAATA
CTAGGGGCTCATGACGGTACTGTAAGAATAAGCACCGGCTAACG

ACTCCTACGGGAGGCAGCAGTGGGGAATTTTGGACAATGGACGC
AAGTCTGATCCAGCCATTCCGTGTGCAGGACGAAGGCCTTCGGA

CTCAGATCGTCGCTGTCTCTGCCAGTTAATCGCCATCTCTGCCA
GTTAATCGCCATCTCTGCCAGTTAATCGCTATCTCTGCCAGTTT

AATCGCCATCTCTGCCAGTTAATCGCCATCTCTGCCAGTTAATC
GCCATCTCTGCCAGTTAATCGCCATCTCTGACGAAATCCACCGC

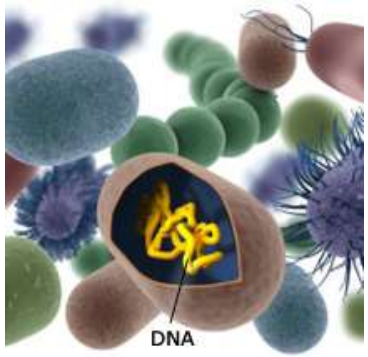
CTCAGATCGTCGCTGTCTCTGCCAGTTAATCGCCATCTCTGCCA
GTTAATCGCCATCTCTGCCAGTTAATCGCTATCTCTGCCAGTTT

CTCAGATCGTCGCTGTCTCTGCCAGTTAATCGCCATCTCTGCCA
GTTAATCGCCATCTCTGCCAGTTAATCGCTATCTCTGCCAGTTT

Metagenomics: Data volumes

One bacterial cell

- 1-5 megabases of DNA (10^6 bp)
- 1 000-5 000 genes



1 gram of soil

- 100 million cells
- 10 000 species
- 100 terabases of DNA (10^{14} bp)

Metagenomics – two approaches

Who are there?

- Analysis of the present microorganisms and their taxonomic affiliation (species, genus, order, etc)
- Estimation of abundance and diversity – how many species are there, how common are they and do they differ between samples?

What are they doing? ('gene-centric analysis')

- Analysis of the present genes and pathways
- Estimation of gene abundances and functional analysis of their biochemical role.

Metagenomics

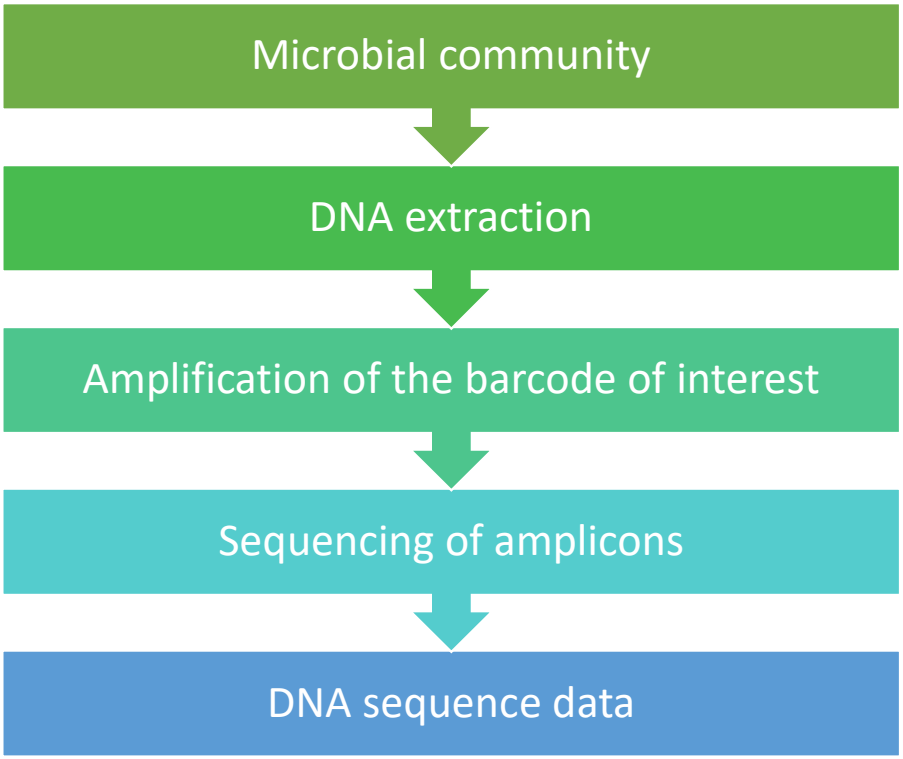
Amplicon sequencing (’metabarcoding’)

- Sequencing of a specific genomic region of interest (‘barcode’)
- General idea: Focus on only on a specific genetic marker that provides information about the organism. The marker is sometimes referred to as a barcode
- The marker is randomly sequenced from all microorganisms in the sample.

Shotgun metagenomic sequencing

- Total DNA from a sample is sequenced
- General idea: Analyze the biological function by analyzing what genes that are present.
- Random fragment from the entire metagenome is sequenced, including all of the genes in the present microorganisms

Amplicon sequencing



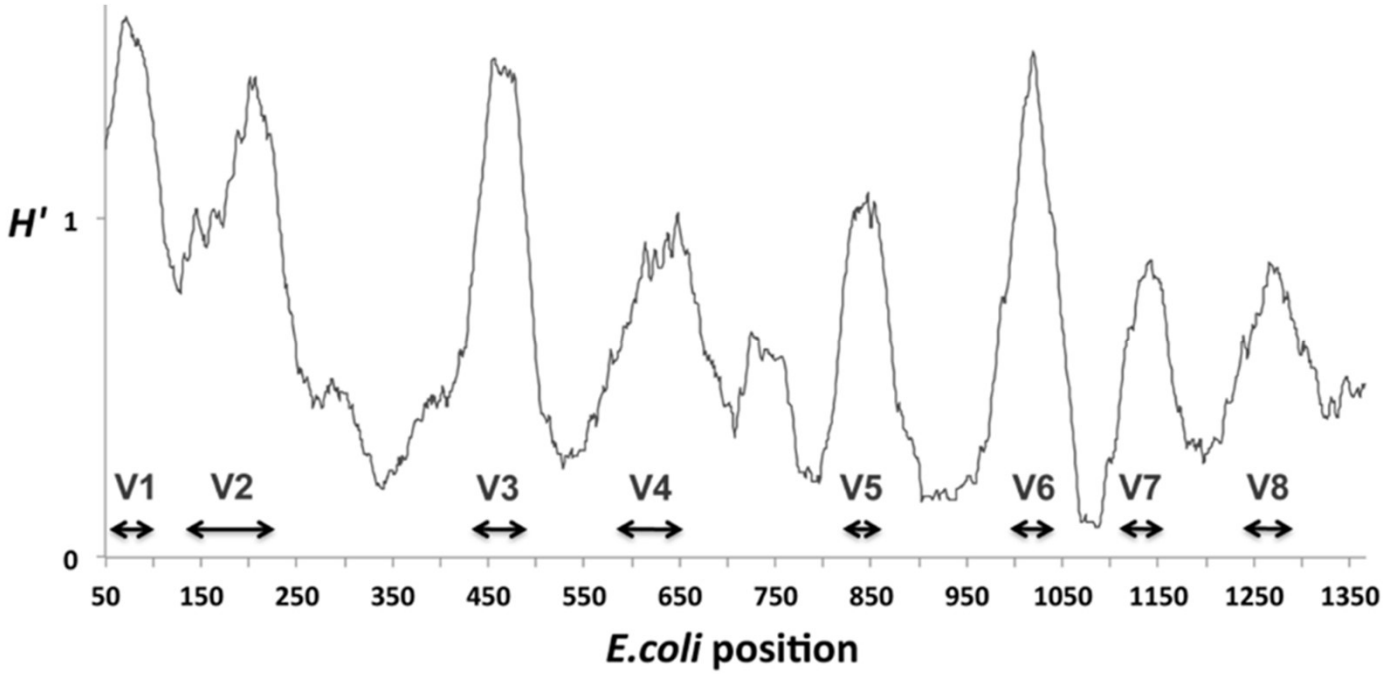
- Analysis starts from the total DNA present in a microbial community.
- The output is a large number of reads from the same marker (i.e. the same region) but from different microbial cells
- Amplification and sequencing of a specific barcode reduces the generated data volumes

Amplicon sequencing - considerations

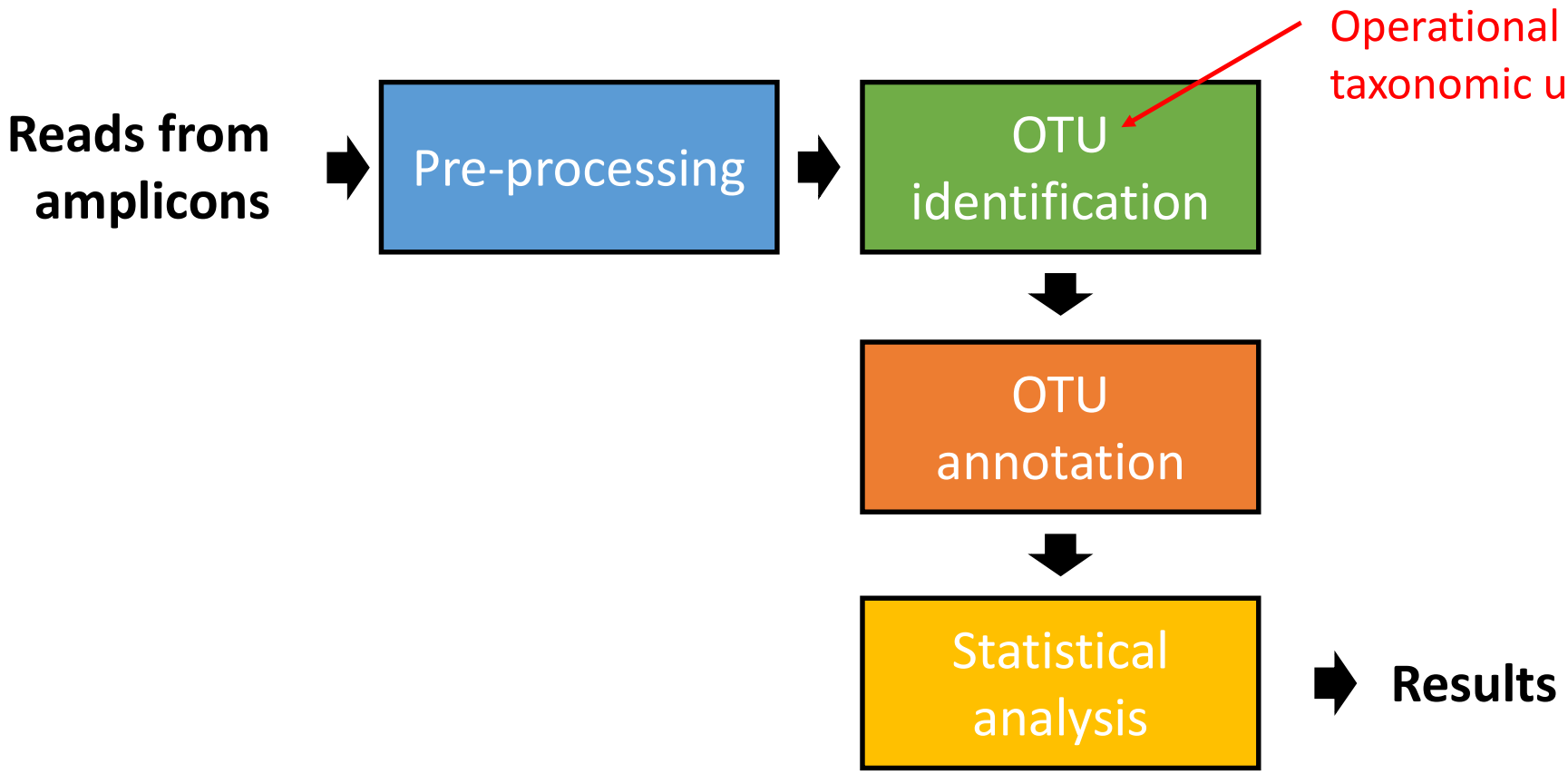
- The marker should be selected so that it
  - Varies between genomes from different organisms
  - Are flanked by conserved regions so that it can be easily amplified in as many species as possible
  - The length matches the sequencing technique
- Commonly used barcodes
  - The 16s ribosomal genes for bacteria
  - The 18s ribosomal genes for eukaryotes
  - The internal transcribed spaces (ITS) for fungi
  - Other functional genes (e.g. nif-genes)

Amplicon sequencing: the 16S rRNA gene

- Part of the small ribosomal subunit (30S)
- Contains both highly conservative and 'hypervariable' regions.



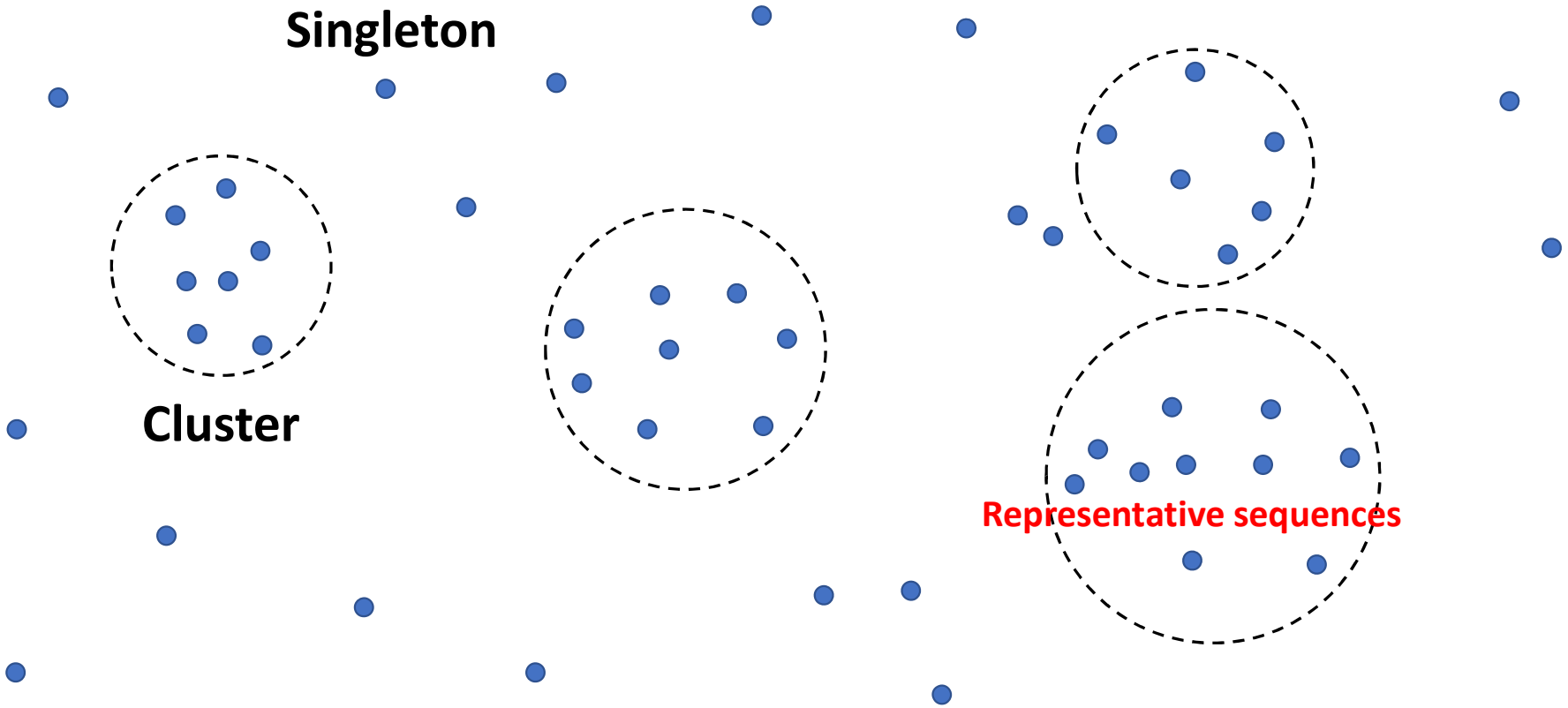
Amplicon sequencing – data analysis

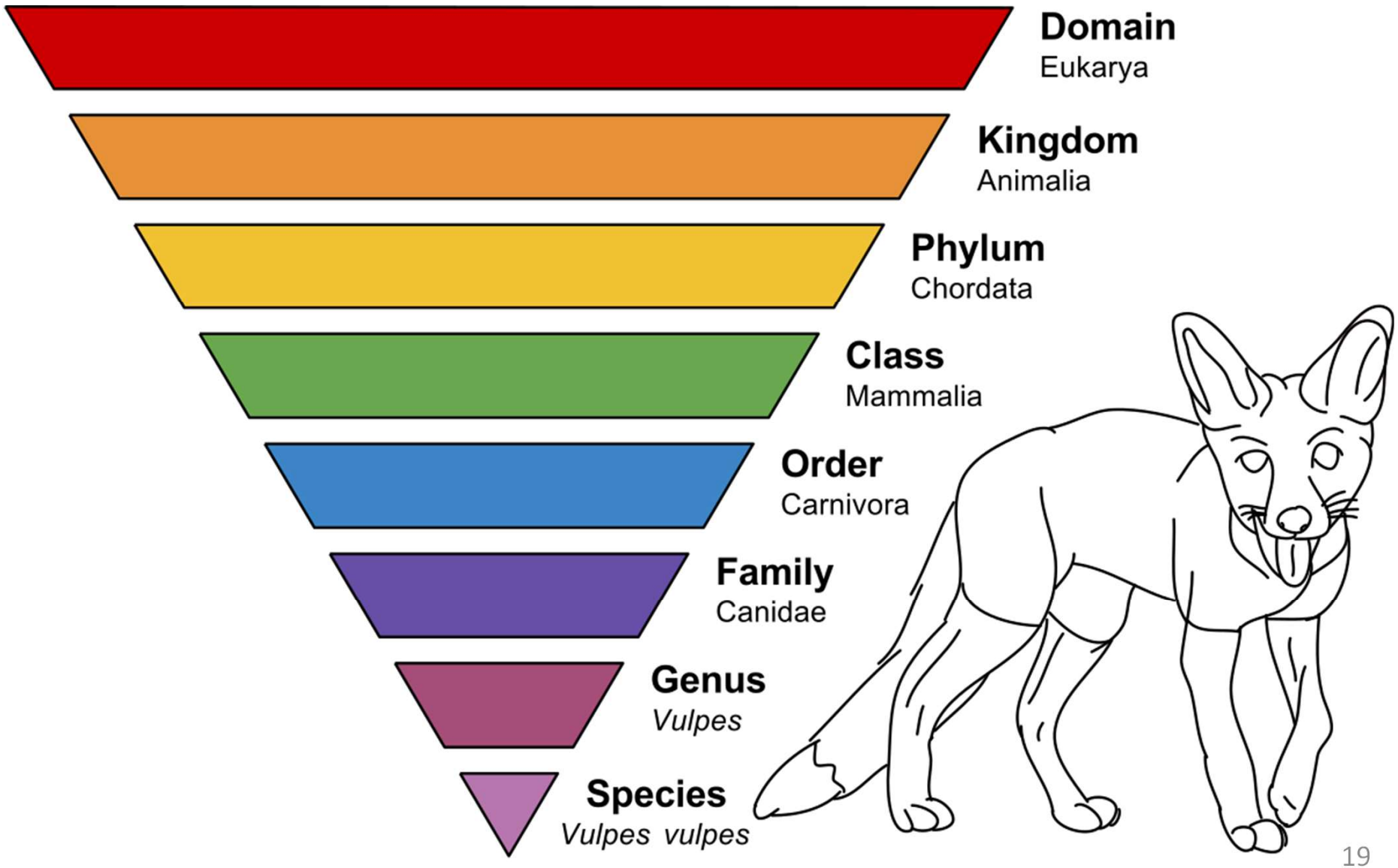


Amplicon sequencing – Identification of OTUs

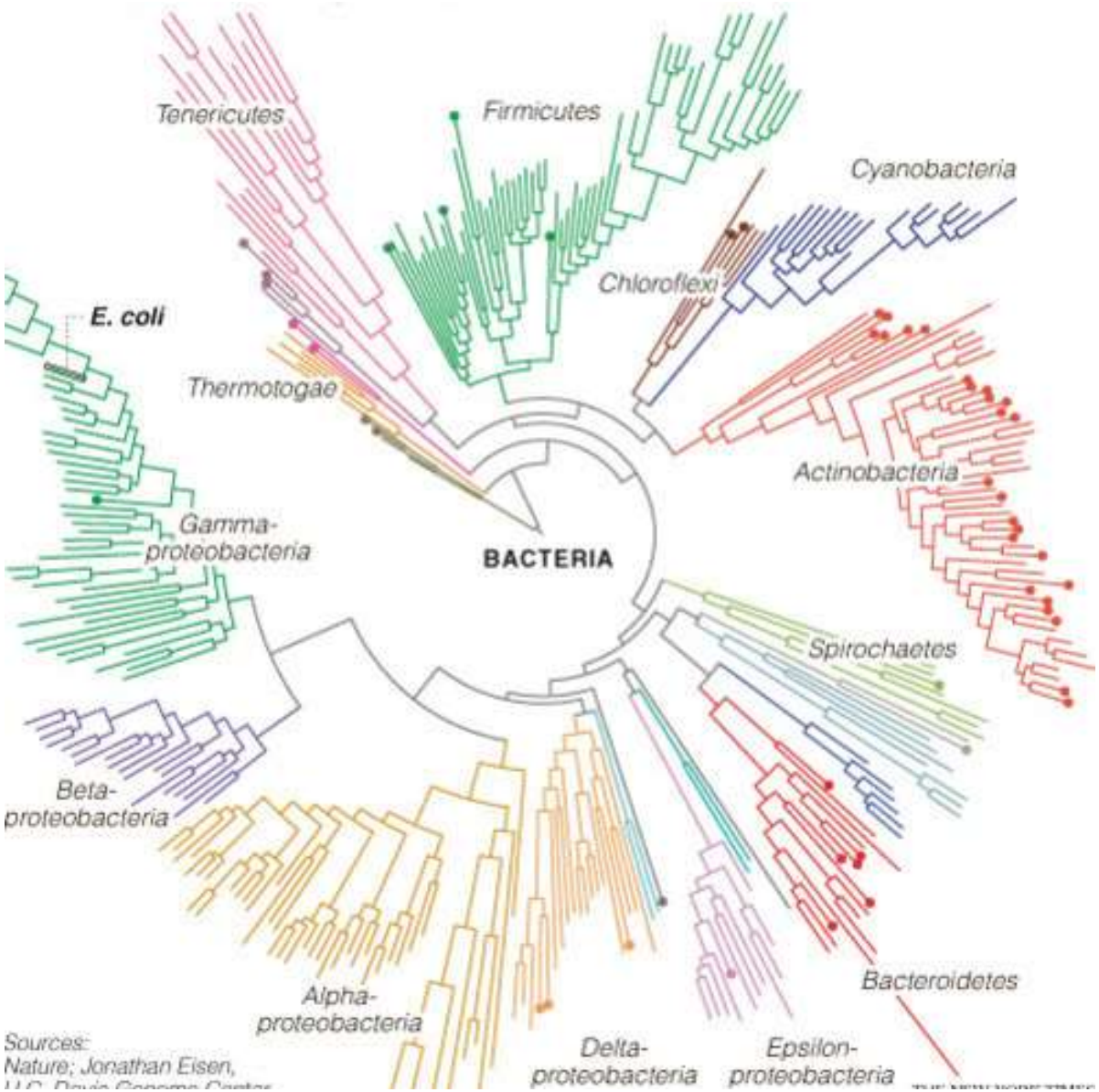
- An operational taxonomic unit (OTU) is a putative species formed by clustering sequences from amplicons.
- Sequences that are sufficiently similar are clustered together and assumed that they come from the same type of organism.
- Sequences that does not cluster with any other sequence are called singletons. These sequences are OTUs but are, in many cases, discarded since they are only observed once.
- A commonly used similarity threshold used for 16S is 97%. Those sequences that have a similarity >97% are clustered together and are thus assumed to be from the same OTU.

Amplicon sequencing – data analysis





Red fox (*Vulpes vulpes*)


Tenericutes
Firmicutes
Cyanobacteria
Chloroflexi
E. coli
Thermotogae
Actinobacteria
Gamma-
proteobacteria
BACTERIA
Spirochaetes
Beta-
proteobacteria
Bacteroidetes
Alpha-
proteobacteria
Delta-
proteobacteria
Epsilon-
proteobacteria
Sources:
Nature; Jonathan Eisen,

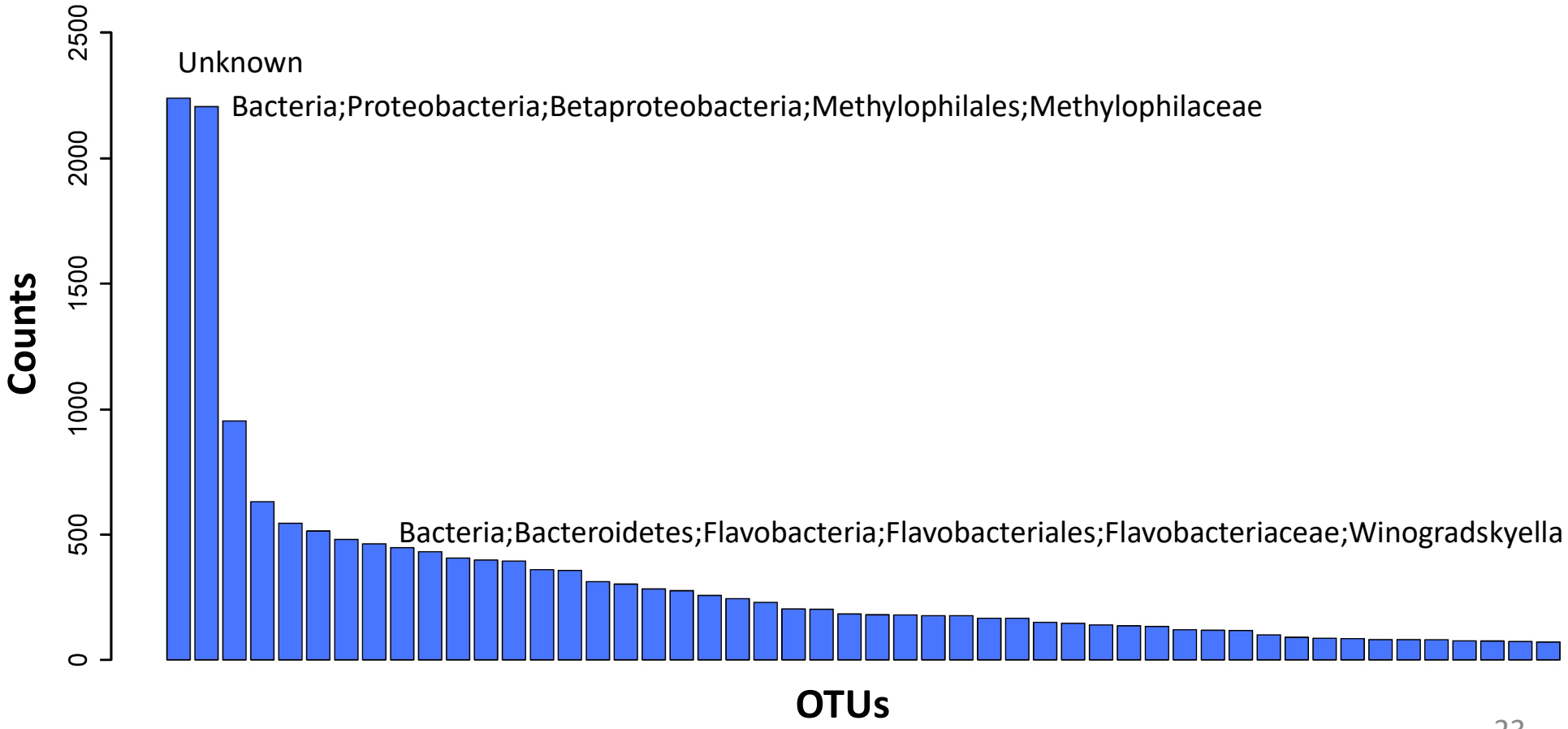
Amplicon sequencing – Annotation of OTUs

Reference database – 16s

>GY194060.4884.6412 Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Unidentified
AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTAGAACGCTGAGGTTTGGTGTTTACACTAGACTGATGAGTTGCGAACGGGTGAGTAACGCGTAGGTAACC
TGCCTCATAGCGGGGGATAACTATTGGAAACGATAGCTAATACCGCATAAGAGTAATTAACACATGTTAGTTATTTAAAAGGAGCAATTGCTTCACTGTGAGATGGACCTGCGTTGTATTAGCT
AGTTGGTGAGGTAAAGGCTCACCAAGGCGACGATACATAGCCGACCTGAGAGGGTGATCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTAGGGAATCTTCGGCA
ATGGACGGAAGTCTGACCGAGCAACGCCGCGTGAGTGAAGAAGGTTTTCGGATCGTAAAGCTCTGTTGTTAGAGAAGAACGTTGGTAGGAGTGGAAAATCTACCAAGTGACGGTAACTAACCAG
AAAGGGACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGGCGGTTCTTTAAGTCTGAAGTTAAAGGCAGTGGCTT
AACCATTGTACGCTTTGGAAACTGGAGGACTTGAGTGCAGAAGGGGAGAGTGGAATTCCATGTGTAGCGGTGAAATGCGTAGATATATGGAGGAACACCGGTGGCGAAAGCGGCTCTCTGGTCT
GTAACTGACGCTGAGGCTCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGTGCTAGGTGTTAGGCCCTTTCCGGGGCTTAGTGCCGCAGCTAACGC
ATTAAGCACTCCGCCTGGGGAGTACGACCGCAAGGTTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACCTTACCAGGTCTT
GACATCCTTCTGACCGGCCTAGAGATAGGCTTTCTCTTCGGAGCAGAAGTGACAGGTGGTGCATGGTTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCT
ATTGTTAGTTGCCATCATTAAGTTGGGCACTCTAGCGAGACTGCCGGTAATAAACCGGAGGAAGGTGGGGATGACGTCAAATCATCATGCCCCTTATGACCTGGGCTACACACGTGCTACAATG
GTTGGTACAACGAGTCGCAAGCCGGTGACGGCAAGCTAATCTCTTAAAGCCAATCTCAGTTCGGATTGTAGGCTGCAACTCGCCTACATGAAGTCGGAATCGCTAGTAATCGCGGATCAGCACG
CCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCACGAGAGTTTGTAACACCCGAAGTCGGTGAGGTAACCTTTTAGGAGCCAGCCGCCTAAGGTGGGATAGATGATTGGG
GTGAAGTCGTAACAAGGTAGCCGTATCGGAAGGTGCGGCTG

>AC201869.46386.47908 Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Candidatus
Regiella;Candidatus Regiella insecticola
AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGCAGCGGGGAGTAGCTTGCTACTCTGCCGGCGAGCGGCGGACGGGTGAGTAAAGTCTGGGGAT
CTGGCTTATGGAGGGGGGTAACCATTGGAAACGGTGGCTAATACCGCATGATGTCGCAAGACCAAAGTGGGGGACTTTAGGGCCTCACGCCATAAGATGAACCCAGATGAGATTAGCTAGTAGG
TGCGGTAATGGCGTACCTAGGCGACGATCTCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGAAACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGC
GGAAGCCTGATGCAGCCATGCCGCGTGTGTGAAGAAGGCCTTAGGGTTGTAAAGCACTTTCAGCGGGGAGGAAGGCGGTAAGAGTAATATGCTTATCGATTGACGTTACCCGCAAAAGAAGCAC
CGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCGAGCGTTAATCGGAATAACTGGGCGTAAAGGGCATGTAGGCGGTAAGTTAAGTCAGATGTGAAATCCCCGAGCTTAACTTGG
GAACGGCATTTGAGACTGGCGGGCTAGAGTTTTGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACAAGGACT
GACGCTGAGGTGCGAAAGCGTGGGTAGCAAACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGATGTCGATTTGGAGGTTGTGGCCTAGTGTTATGGCGTCCGAAGCTAACGCGATAAAT
CGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCTACTCTTGACATCC
ACGGAATTGAGCAGAGATGCTTAAGTGCCTTCGGGAGCCGTGAGACAGGTGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGTCCTTT
GTTGCCAGCGCGTGATGGCGGGAACTCAAAGGAGACTGCCGGTGATAAACCGGAGGAAGGTGGGGATGACGTCAAGTCATCATGGCCCTTACGAGTAGGGCTACACACGTGCTACAATGGCGCA
TACAAAGAGCGGCGAGCGTGCGAACGTAAGCGAACCTCATAAAGTGCGTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATGAAGTCGGAATCGCTAGTAATCGTGGATCAGAATGCTACG
GTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTGGGCGGCAAAAGAAGTAGGTAGCTTAACCTTCGGGGGGGCGCTTACCACTTTGTGGTTCATGACTGGGGTGAAG
TCGTAACAAGGTAACCGTAGGGGAACCTGCGGTTG

Amplicon sequencing – OTU distribution



Amplicon sequencing – statistical analysis

The OTU distribution are typically used to

- Compare differences between samples ('comparative metagenomics')
  - What species are present in the gut of sick compared to healthy individuals?
  - What species are affected in polluted environments?
  - This analysis is very similar to the analysis of RNA-seq data.
- Estimate the biodiversity in a sample
  - How many species are present?
  - How are the species distributed?

Amplicon sequencing – diversity

Alpha diversity

The diversity on the local level (in a habitat). This include, for example, diversity at a specific geographical site or in a specific individual.

Beta diversity

Diversity between habitats. This include, for example, the diversity between geographical sites or between individuals.

Alpha diversity

Let p_i be the proportion of OTU $i = 1, \dots, n$.

Richness

Unique number of OTUs. Can easily be calculated by counting the number of OTUs where $p_i > 0$.

Evenness

Shannon diversity index. Estimates the evenness of a distribution by calculating the entropy,

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

A higher value indicates a more uniform distribution.

Alpha diversity

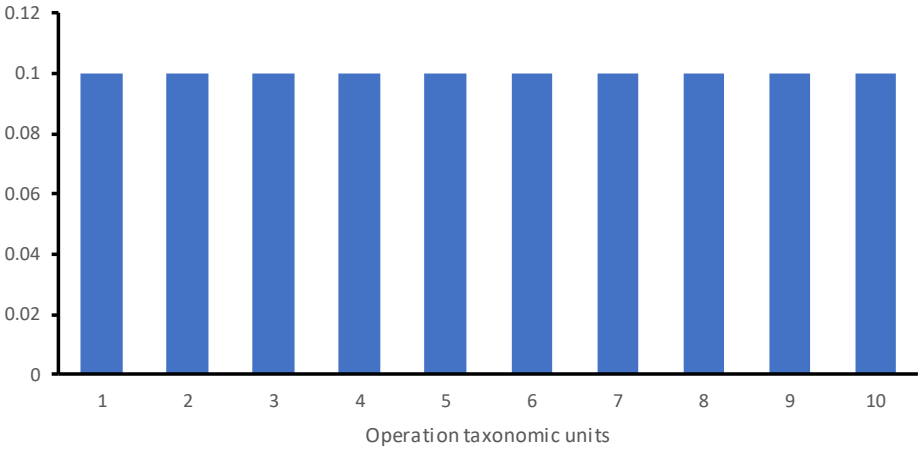
Evenness

Simpson's diversity index. Calculates evenness by estimating how likely that two randomly picked OTUs are from the same species.

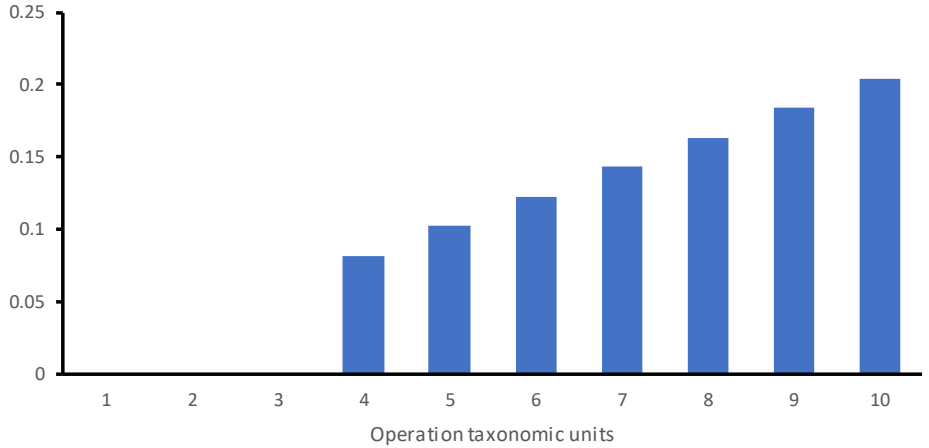
$$\lambda = \sum_{i=1}^{n} p_i^2$$

The Simpson index is a value between 0 and 1. A higher value indicates a less uniform distribution (max value achieved if $p_i = 1$ for some *i*).

Diversity - an easy example



Richness, unique OTUs: 10
Evenness, Shannon: 2.30
Evenness, Simpson: 0.10



Richness, unique OTUs: 7
Evenness, Shannon: 1.90
Evenness, Simpson: 0.15

Diversity - rarefaction

- The diversity indices are dependent on the sequencing depth.
- Examples
  - A higher sequencing depth means more detected OTUs and therefore a higher richness.
  - More detected OTUs means a higher Shannon index (maximum is *log(N)* where *N* is the number of OTUs)
- In order to make indices between samples comparable they need to be rarefied, i.e. subsampled to the same sequencing depth.
  1. Select a sequencing depth *n*.
  2. For each sample, randomly sample *n* fragments without replacement.
  3. Count new OTU abundances for the rarefied data.

Metagenomics

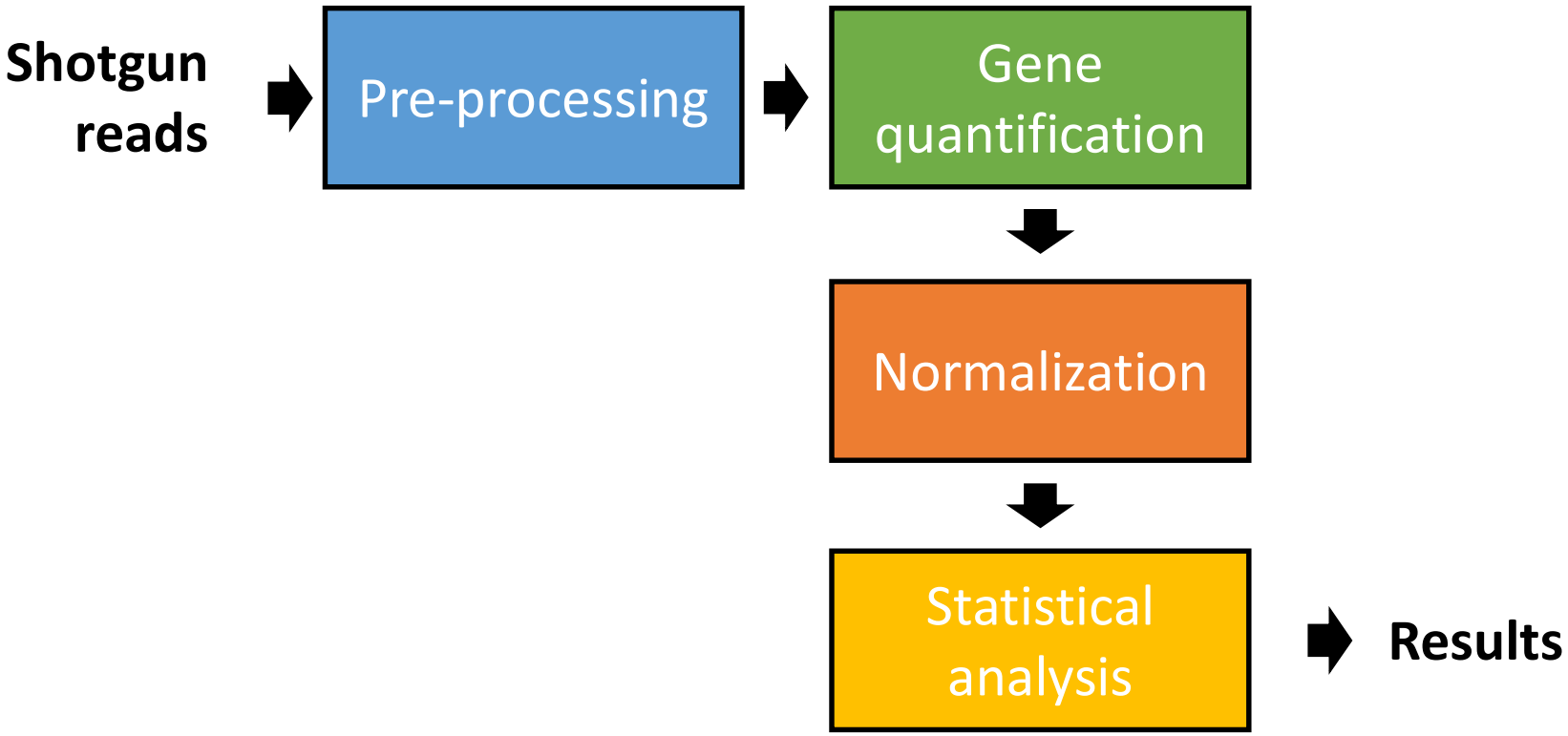
Amplicon sequencing ('metabarcoding')

- Sequencing of a specific genomic region of interest ('barcode')
- General idea: Focus on only on a specific genetic marker that provides information about the organism. The marker is sometimes referred to as a barcode
- The marker is randomly sequenced from all microorganisms in the sample.

Shotgun metagenomic sequencing

- Total DNA from a sample is sequenced
- General idea: Analyze the biological function by analyzing what genes that are present.
- Random fragment from the entire metagenome is sequenced, including all of the genes in the present microorganisms

Shotgun metagenomic sequencing – data analysis

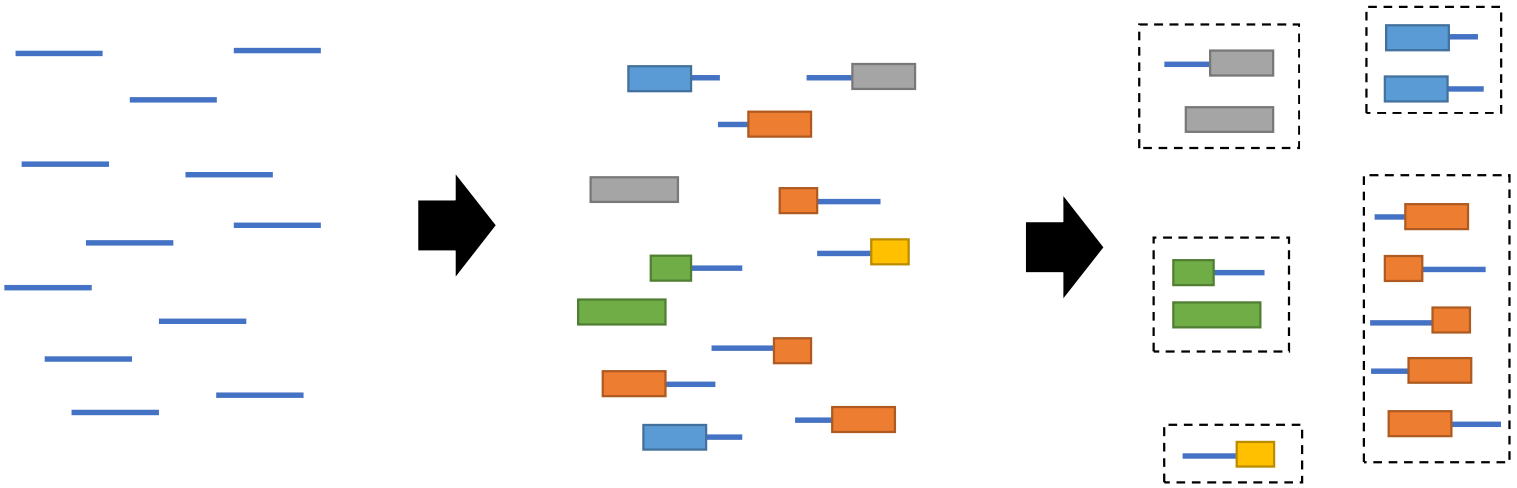


Shotgun metagenomics - gene quantification

Direct binning of metagenomic reads

- Search each metagenomic fragments for the presence of genes.
- A vast number of the microbial genes are not present in the databases. The search therefore requires sensitive aligners and approximate matches are often accepted.
- Requires relatively long reads (preferably >500bp) and not generally possible to do for short reads (genes that are well-conserved are easier to identify shorter reads).
- 'Bins' are finally formed by counting the number of reads for each type of gene.

Shotgun metagenomics - direct binning



Shotgun reads

Identification of genes

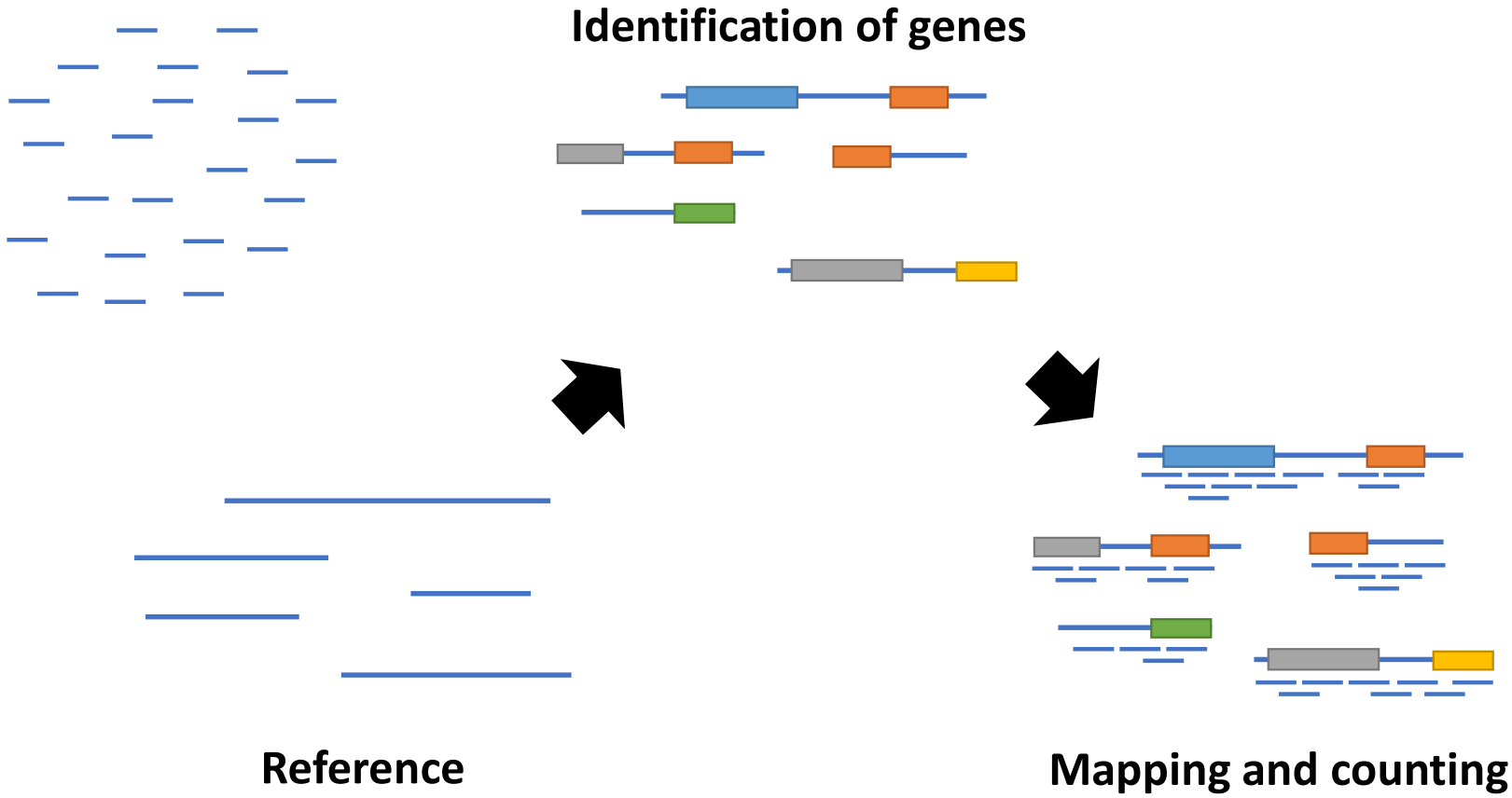
Sorting and counting

Shotgun metagenomics - gene quantification

Reference-guided binning

- Guided binning uses an annotated reference database that contains the genomes of the microorganisms present in the sample.
- Each metagenomic fragment is mapped against the reference database.
- 'Bins' are formed by counting the number of reads matching each type of gene present in the genomes.
- Typically done for data with short reads (<500bp).

Shotgun metagenomics - reference-guided binning



Shotgun metagenomics - gene quantification

Commonly used reference databases for guided binning

- Genomes
  - >10k completely sequenced and >400k partially sequenced bacterial genomes are available in the repositories. High bias towards pathogenic bacteria and many environmental bacteria are missing.
- Gene catalogues
  - Available for the human microbiome and contains today around 10 million genes.
- *De novo* assembly of the metagenome
  - Longer sequences are created directly from the metagenomic data. Often the only choice for environmental metagenomes.

Shotgun metagenomics - normalization

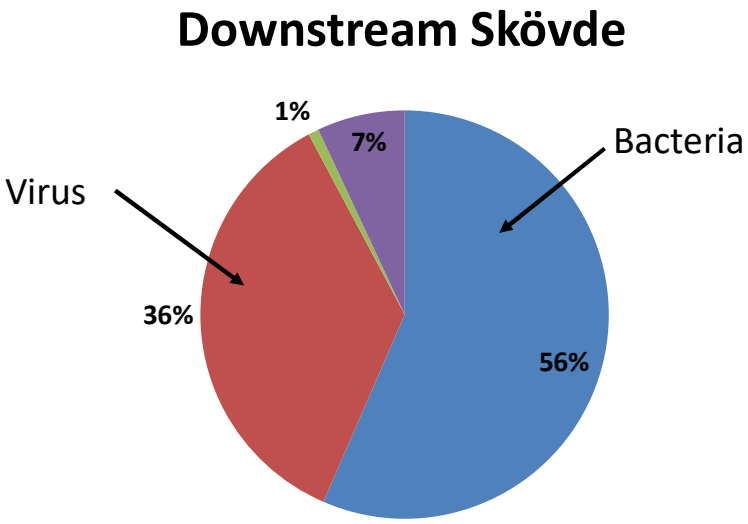
Data from different metagenomic samples are not directly comparable

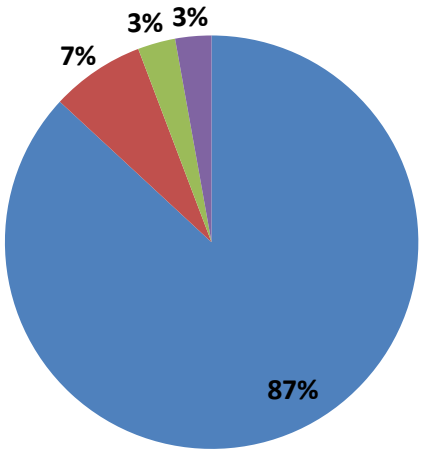
- Differences in sequencing depth
- Systematic effects caused by
  a) Variation in sampling and preparation
  b) The quality sequencing
  c) Systematic errors introduced in the sequencing (e.g. GC-bias)
  d) Taxonomic composition

Normalization is necessary to make samples comparable.

Normalization – example of systematic effects

Metagenomic sequencing of river sediments upstream and downstream of a waste water treatment plant.





Shotgun metagenomics - normalization

Commonly methods for normalization are

- Counts per million reads (CPM)
- Reads per kilobase per million (RPKM)
- Median
- Upper quartile
- TMM
- Reference gene (e.g. 16s)

Methods developed for RNA-seq data may not work as reliable on shotgun metagenomic data.

The nature of shotgun metagenomic data

High dimensional

Many genes present within a single community

- Thousands of functional features (e.g. PFAM, TIGRFAM)
- Millions at a higher resolution (e.g. genes in a catalogue)

Example: Human gut metagenomes from 1000 individuals contains 10 million unique genes. Diversity in many environmental community is estimated to be 10-fold higher.

Few samples

The number of samples are few (often <10) due to

- Costs associated with sequencing and sample preparation
- Sequencing depth is prioritized over biological replication

The nature of shotgun metagenomic data

Very high variability

Metagenomic data has an innate variability form randomly selecting DNA fragments.

Additional technical noise comes from

- DNA extraction and sample preparation
- Sequencing errors
- Binning errors

Biological noise comes from variation between microbial communities

- Variation in species composition
- Variation between genotypes

<u>The variability is typically higher than in RNA-seq data.</u>

Identification of differentially abundant genes

Three approaches

1. Methods based on normal assumptions
   - Can use the flexibility of models based on the normal distribution
   - Will always results in approximation and thus sub-optimal performance
   - Hard to find a transformation that is suitable for all genes
2. Methods based on non-parametric methods
   - Robust against model assumptions
   - Requires a larger number of samples
   - Can be highly sensitive against ties, which are common for gene with few counts.
3. Methods based on count distributions
   - The correct statistical nature of the data is described
   - Complicated models that requires assumption about the mean-variance relationship
   - Computationally more expensive (but usually manageable)

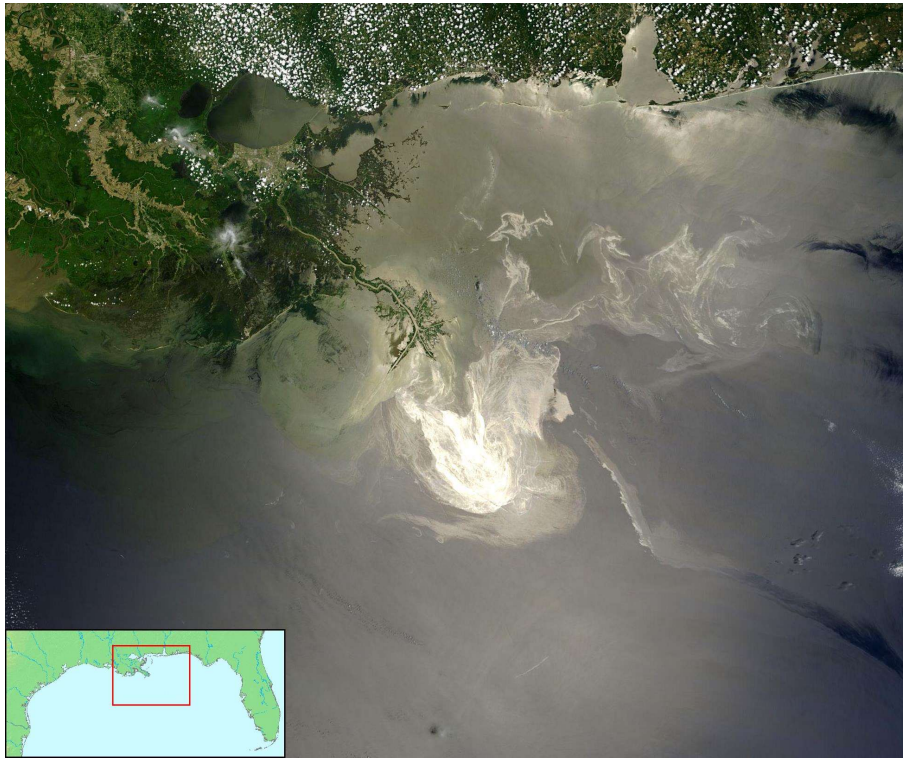
Statistical analysis: DESeq2

- An R-package for analysis of overdiserpsed count data. Originally developed for RNA-seq data.
- Implements
  - Robust normalization method called REL that have high similarities to TMM
  - A statistical model that have many similarities to the negative binomial distribution.
  - Robust estimation of gene-specific overdispersion (shrinkage)
- Can handle many experimental designs, including comparisons of groups, regression and ANOVA-like analyses.
- Easy-to-use, well-tested and have reliable performance for many datasets.
- https://www.bioconductor.org/packages/release/bioc/html/DESeq2.html

Computer exercise 4: Analysis of sediments exposed to oil

Computer exercise 4

- Deepwater Horizon oil spill in the Mexican Gulf
- Samples taken from sediment close and far away from the borehole
- Overall goal is to assess the environmental impact of oilspill



Computer exercise 4

Aims

1. Which species/OTUs are present in the samples?
2. How does the species composition and diversity compare between samples?
3. Are there biological functions that differ between polluted and clean environments?

Computer exercise 4

Study design

- 6 sampling sites
  - 3 highly contaminated
  - 3 less contaminated
- 16S rRNA amplicon data
- Shotgun data

Summary

- Metagenomics is the study of the metagenome by random sampling of DNA fragments.
- Metagenomics can be performed by
  - Sequencing of specific genomic regions that are amplified by PCR (amplicons)
  - Sequencing of random fragments form the entire metagenome (shotgun)
- Amplicon data is analyzed by clustering the reads into OTUs which are then annotated.
- The OTU abundance distribution can be used for comparisons between samples and for estimation of the biodiversity.

Summary

- Analysis of shotgun metagenomic data includes
  - Gene quantification
  - Normalization
  - Identification of differentially abundant genes
- Count models can be used to specifically model the discrete nature of metagenomics (and RNA-seq) data. These models are however often non-standard and implemented in specific R-packages.
- Many methods used in RNA-seq can be applied to shotgun metagenomics and vice versa.