

# Final lecture - repetition

Introduction to bioinformatics (MVE510)

Autumn, 2020

# Course organization - examination

**The course is examined in two ways**

1. Written exam, 14 January 2020
2. Approved reports from computer exercises. Exercise 1 is examined during the computer exercise. Exercise 2-4 requires a report that is handed in through PingPong.

**Examination in** MVE510 - Introduction to bioinformatics

**Time:** Thursday January 16 2020, 08.30-12.30

**Examiner:** Erik Kristiansson

**On call:** Felix Held, phone 031-772 5325

~~**Aids:** None~~

Maximum is 40 points. Grade 3 requires 18 points, grade 4 requires 26 points and grade 5 requires 34 points.

1. Data from next-generation DNA sequencing is known to contain a wide range of errors and inaccuracies.
  - (a) Describe the meaning of the following types of errors: 'substitution', 'indel' and 'duplicate'? (2p)
  - (b) What is 'adapter contamination'? Why is this an issue and where does it come from? (2p)

# Common errors: substitutions

## Correct sequence

**GGCGCTGGACTCTACAGCAGATGTGGAACTGGAGA**

CGCTGGGCTCTACATCAG

GGACTCTACAGCAGATGTGG

GACTCTACAGCAGATGTGGA

TCTACATCAGATGTGGAA

CAGCAGATGTGGAACTGGAG

## Sequence reads

# Common errors: Insertions and deletions ('indels')

## Correct sequence

**CTTCATAAGCTAGATGCCAGTTAA-CTGTCGAGAGG**

CTAGATG- CAGTTAA-CTGTC

AGATGCCAGTTAA A CTGTCGA

ATGCCAGTTAA-CTGTCGAGA

TGCCAGTTAA-CTGTCGAGAG

TGCCAGT- AA-CTGTCGAGAG

## Sequence reads

# Common errors: duplicates

- Duplicates are caused by sequencing the same DNA fragment multiple times. These reads all come from the same DNA molecule and does not describe the true diversity in the sample.
- Duplicates typically caused by biases in the amplification steps where certain DNA fragments are amplified with higher efficiency. More amplification often means more duplicates.
- In many applications, duplicates are important to remove to avoid incorrect and misleading results.

1a)

## Substitution

- One nucleotide are incorrectly exchanged for another one

## Indel

- A nucleotide is incorrectly inserted or deleted

## Duplicate

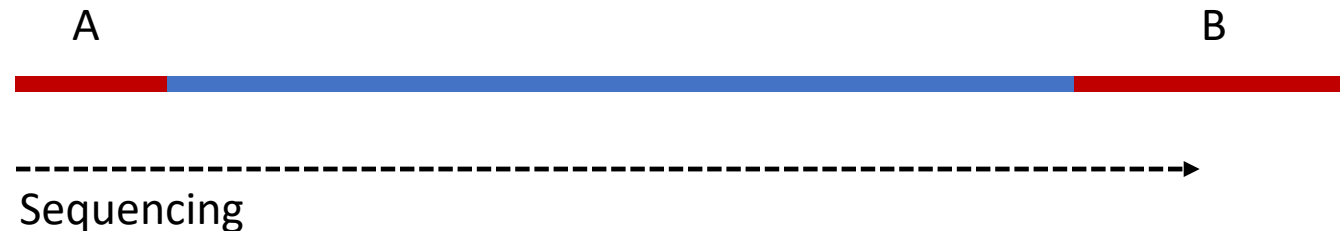
- The same DNA fragment is sequenced several time. Often a result from biases in the amplification before sequencing.
- Errors in the amplification step can appear in a subset of the duplicates and can be mistaken for mutations.



1. Data from next-generation DNA sequencing is known to contain a wide range of errors and inaccuracies.
  - (a) Describe the meaning of the following types of errors: 'substitution', 'indel' and 'duplicate'? (2p)
  - (b) What is 'adapter contamination'? Why is this an issue and where does it come from? (2p)

# Common errors: adapter contamination

- Most sequencing platforms adds adapters to the ends of the reads
- Typically one of the adapters are sequenced and needs to be removed before analysis.
- However, if the DNA fragment is too short, the sequencing process can start to also sequence a part of the other adapter.



1b)

## Adapter contamination

- Result from sequencing DNA fragments that are too short.
- The read will then include parts of the adapter in the end.
- Can be a large issue when working with degraded DNA.
- Needs to be removed in order to avoid errors in the down-stream analysis.

2. Apply the Smith-Waterman algorithm to perform a local alignment between

Sequence 1: GTCTAGT

Sequence 2: CTCAG

Use the scoring matrix

$$S(a, b) = \begin{cases} 7 & \text{if } a = b \\ -3 & \text{otherwise} \end{cases}$$

and a linear gap penalty with  $d = -8$ . Report all optimal alignments and their corresponding scores. (5p)

# Two main forms of alignment

## **Global alignment**

Two sequences are aligned over their full length.

## **Local alignment**

Two sequences are aligned based on their best matching subsequences.

Local alignments are used to match short sequence reads against long reference sequences.

## 2) Two solutions

Alignment 1

CT-AG

CTCAG

Score: 20

Alignment 2

TCTAG

TC-AG

Score: 20

3. In metagenomic amplicon sequencing, a specific genomic region ('barcode') is sequenced and analyzed to assess the taxonomic composition of a microbial community.
- (a) Provide an overview of the analysis of data from amplicon sequencing. Which are the necessary steps and why are they important? (3p)
  - (b) One way to analyze amplicon sequence data is to estimate the diversity. What is the difference between richness and evenness and what do they measure? Describe one way to estimate richness and one way to estimate evenness (3p)

# Metagenomics

## Amplicon sequencing ('metabarcoding')

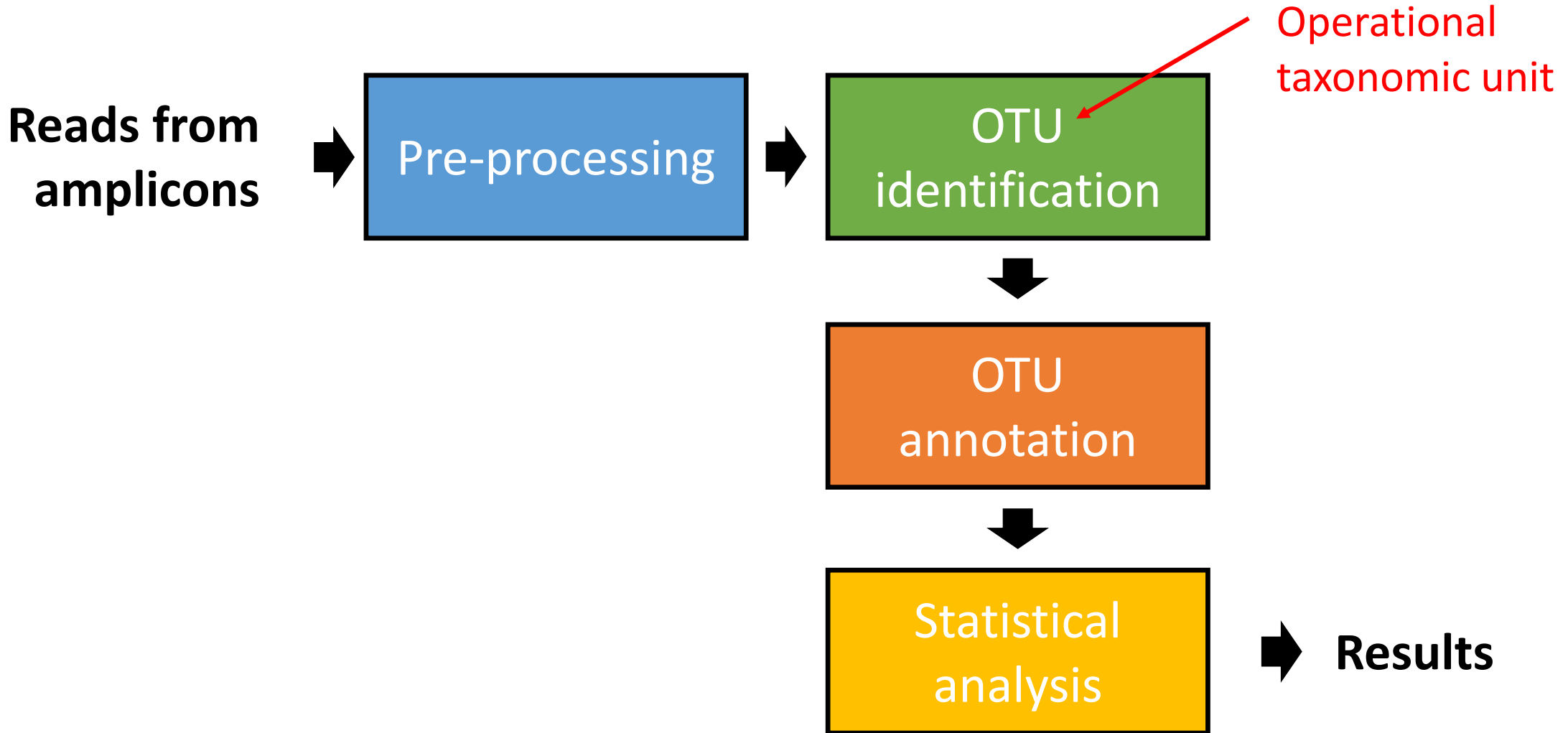
- Sequencing of a specific genomic region of interest ('barcode')
- General idea: Focus on only on a specific genetic marker that provides information about the organism. The marker is sometimes referred to as a barcode
- The marker is randomly sequenced from all microorganisms in the sample.

## 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



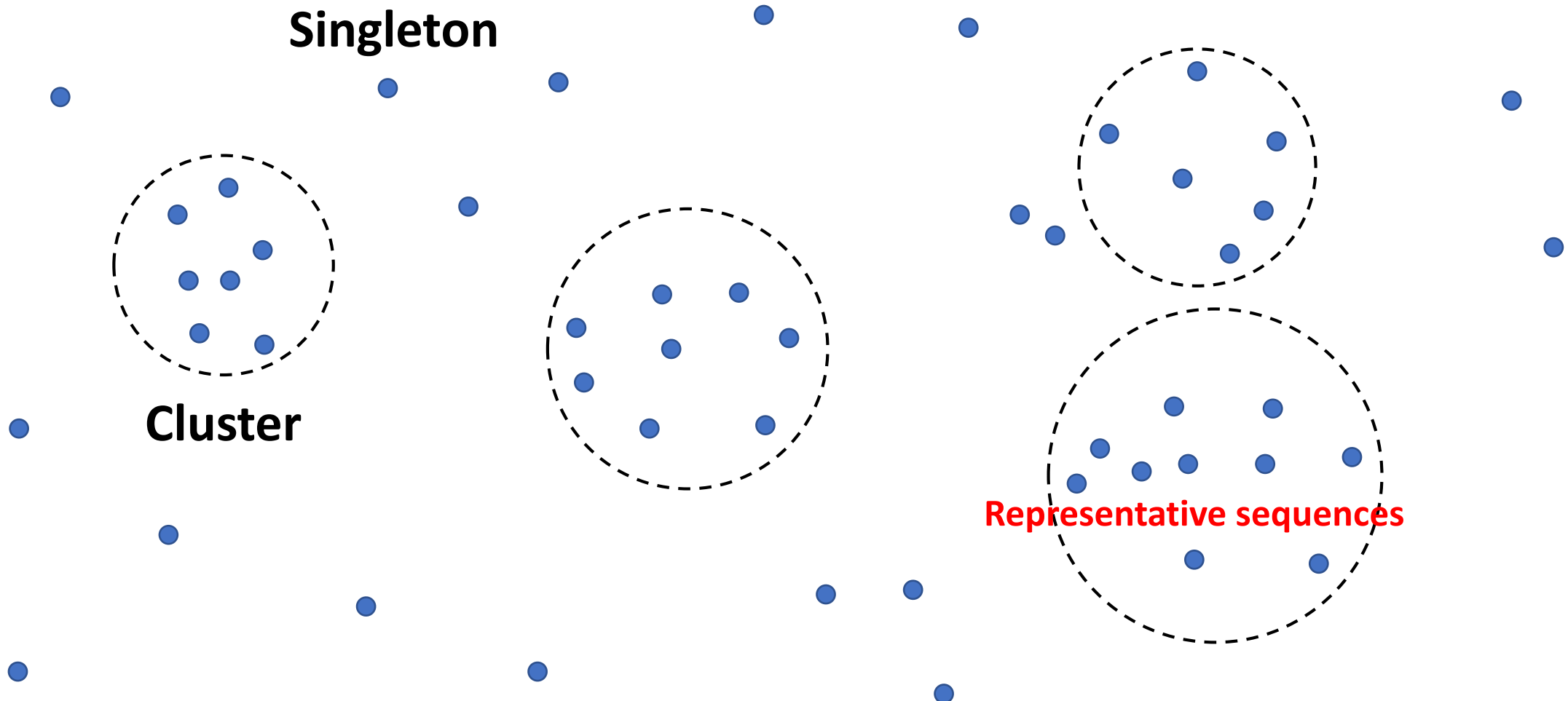
# Amplicon sequencing – data analysis



# 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



# Amplicon sequencing – Annotation of OTUs

## Reference database – 16s

```
>GY194060.4884.6412 Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Unidentified
AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTAGAACGCTGAGGTTTGGTGTGTACACTAGACTGATGAGTTGCGAACGGGTGAGTAACGCGTAGGTAACC
TGCCCTCATAGCGGGGATAACTATTGGAACGATAGCTAATACCGCATAAGAGTAATTAACACATGTTAGTTATTTAAAAGGAGCAATTGCTTCACTGTGAGATGGACCTGCGTTGTATTAGCT
AGTTGGTGAGGTAAAGGCTCACCAAGGCGACGATACATAGCCGACCTGAGAGGGTGATCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTAGGGAATCTTCGGCA
ATGGACGGAAGTCTGACCGAGCAACGCCGCGTGAGTGAAGAAGGTTTTTCGGATCGTAAAGCTCTGTTGTTAGAGAAGAACGTTGGTAGGAGTGGAAAATCTACCAAGTGACGGTAACCTAACCG
AAAGGGACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGGCGGTTCTTTAAGTCTGAAGTTAAAGGCAGTGGCTT
AACCATTGTACGCTTTGGAACTGGAGGACTTGAGTGCAGAAGGGGAGAGTGGAATTCATGTGTAGCGGTGAAATGCGTAGATATATGGAGGAACACCGGTGGCGAAAGCGGCTCTCTGGTCT
GTAAGTACGCTGAGGCTCGAAAGCGTGAGGAGCAAACAGGATTAGATAACCTGGTAGTCCACGCCGTAAACGATGAGTGCTAGGTGTTAGGCCCTTTCCGGGGCTTAGTGCCGCAGCTAACGC
ATTAAGCACTCCGCCTGGGGAGTACGACCGCAAGGTTGAACTCAAAGGAATTGACGGGGGCGCACAAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACCTTACCAGGTCTT
GACATCCTTCTGACCGGCCCTAGAGATAGGCTTTCTCTTCGAGCAGAAAGTGACAGGTGGTGCATGGTTGTCTGTCAGCTCGTGTCTGAGATGTTGGGTAAAGTCCCGCAACGAGCGCAACCCCT
ATTGTTAGTTGCCATCATTAAGTTGGGCACTCTAGCGAGACTGCCGGTAATAAACCGGAGGAAGGTGGGGATGACGTCAAATCATCATGCCCTTATGACCTGGGCTACACACGTGCTACAATG
GTTGGTACAACGAGTCGCAAGCCGGTGACGGCAAGCTAATCTCTTAAAGCCAATCTCAGTTTCGGATTGTAGGCTGCAACTCGCCTACATGAAGTCGGAATCGCTAGTAATCGCGGATCAGCACG
CCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCACGAGAGTTTGTAAACACCCGAAGTCGGTGAGGTAACCTTTTAGGAGCCAGCCGCTAAGGTGGGATAGATGATTGGG
GTGAAGTCGTAACAAGGTAGCCGTATCGGAAGGTGCGGCTG
```

```
>AC201869.46386.47908 Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Candidatus
Regiella;Candidatus Regiella insecticola
AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGCAGCGGGGAGTAGCTTGCTACTCTGCCGGCGAGCGGCGGACGGGTGAGTAAAGTCTGGGGAT
CTGGCTTATGGAGGGGGGTAACCATTTGGAACGGTGGCTAATACCGCATGATGTCGCAAGACCAAAGTGGGGGACTTTAGGGCCTCACGCCATAAGATGAACCCAGATGAGATTAGCTAGTAGG
TGCGGTAATGGCGTACCTAGGCGACGATCTCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAAGTGAAGAACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGC
GGAAGCCTGATGCAGCCATGCCGCGTGTGTGAAGAAGGCCTTAGGGTTGTAAAGCACTTTCAGCGGGGAGGAAGGCGGTAAGAGTAATATGCTTATCGATTGACGTTACCCGCAAAAGAAGCAC
CGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCGAGCGTTAATCGGAATAACTGGGCGTAAAGGGCATGTAGGCGGTAAGTTAAGTCAGATGTGAAATCCCCGAGCTTAACTTGG
GAACGGCATTGAGACTGGCGGGCTAGAGTTTTGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACAAGGACT
GACGCTGAGGTGCGAAAGCGTGGGTAGCAAACAGGATTAGATAACCTGGTAGTCCACGCTGTAAACGATGTCGATTTGGAGGTTGTGGCCTAGTGTTATGGCGTCCGAAGCTAACGCGATAAAT
CGACCGCCTGGGGAGTACGGCCGCAAGGTTAAACTCAAATGAATTGACGGGGGCCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCTACTCTTGACATCC
ACGGAATTGAGCAGAGATGCTTAAGTGCCCTTCGGGAGCCGTGAGACAGGTGCTGCATGGCTGTCGTGAGCTCGTGTGTTGTGAAATGTTGGGTAAAGTCCCGCAACGAGCGCAACCCCTTGTCTTT
GTTGCCAGCGCGTGATGGCGGGAACCTCAAAGGAGACTGCCGGTGATAAACCGGAGGAAGGTGGGGATGACGTCAAGTCATCATGGCCCTTACGAGTAGGGCTACACACGTGCTACAATGGCGCA
TACAAAGAGCGGCGAGCGTGCGAACGTAAGCGAACCTCATAAAGTGCGTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATGAAGTCGGAATCGCTAGTAATCGTGGATCAGAATGCTACG
GTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTGGGCGGCAAAAGAAGTAGGTAGCTTAACCTTCGGGGGGGCGCTTACCACCTTGTGGTTCATGACTGGGCTGAAG
TCGTAACAAGGTAACCGTAGGGGAACCTGCGGTTG
```

# 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?

3a)

The analysis of amplicon sequencing data includes 1) preprocessing, 2) OTU identification, 3) OTU annotation and 4) statistical analysis.

1. The preprocessing is used to remove and/or trim reads with low quality. Reads with a high number of sequencing errors can otherwise be interpreted as new species.

2. In the OTU identification step, reads are clustered based on similarity to form OTUs which constitutes potential species. A cut-off of 3% sequence similarity is often used and sequences that have a similarity of 97% or more are thus considered to come from the same species. Reads that do not cluster are called singletons. A representative sequence (e.g. the longest sequence) is used to represent each cluster.

3a)

3. In the third step, the OTUs are annotated by comparing them (e.g. the representative sequence) against a reference database. The reference database consists of sequences with a known taxonomic affiliation, i.e. we have some knowledge from which species the sequence come from. Dependent on the match to the reference we can annotate each OTU to various taxonomic ranks. Some OTUs may not have a match and are therefore considered to be unknown.

4. In the final steps, the data is statistically analyzed. This can include comparison of the counts of OTUs between samples (comparative metagenomics) or to investigate the diversity in one or several samples. This steps requires that the data is normalized to make it comparable (e.g. using rarefaction).

3. In metagenomic amplicon sequencing, a specific genomic region ('barcode') is sequenced and analyzed to assess the taxonomic composition of a microbial community.
- (a) Provide an overview of the analysis of data from amplicon sequencing. Which are the necessary steps and why are they important? (3p)
  - (b) One way to analyze amplicon sequence data is to estimate the diversity. What is the difference between richness and evenness and what do they measure? Describe one way to estimate richness and one way to estimate evenness (3p)



# 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

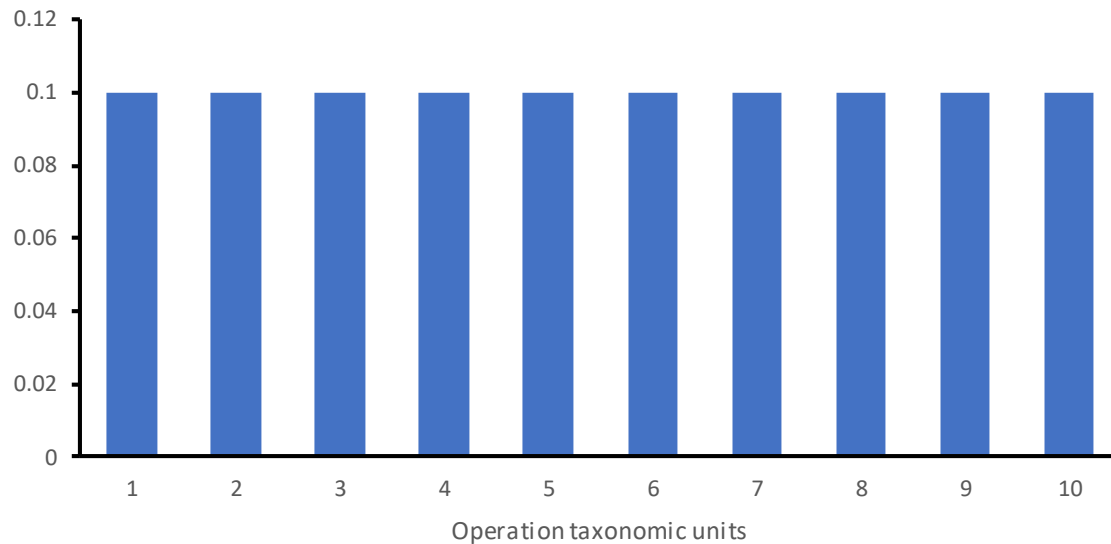
## Evenness

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

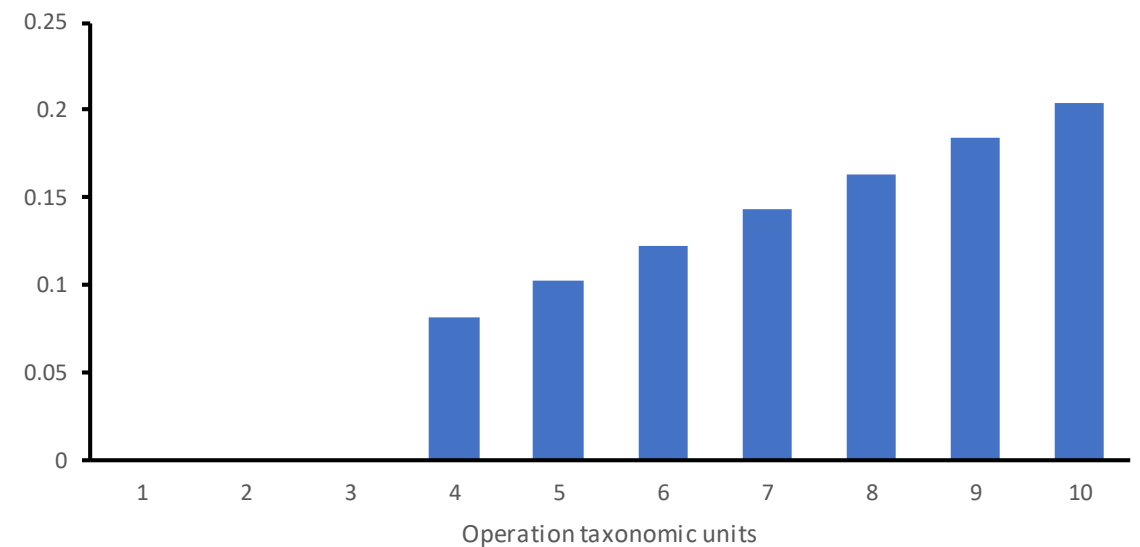
$$\lambda = \sum_{i=1}^n p_i^2$$

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

3b)

Richness measures how many species that are present in a sample and evenness how even their abundances are distributed. A sample can have high richness and low evenness and, vice versa, low richness but high evenness.

Richness can be estimated by counting the number of different species (OTUs) that are found in a sample. Evenness can be estimated using Shannons index (measure the 'distance' to a uniform distribution). A high Shannon means a more uniform abundance distribution. Another way to measure evenness is to use the Simpson index, which measures the probability that we get the same species if we take two reads. Note that all these diversity indices are influenced by sequencing depth and rarefication of the data is therefore necessary.

4. In bioinformatics, a large number of tests is often performed simultaneously. This gives rise to the multiple testing problem.
- (a) Describe the multiple testing problem. What can happen if it is not properly addressed? (2p)
  - (b) What is the family-wise error rate (FWER)? Also, describe the Bonferroni method and how it can be used to control the FWER (3p).
  - (c) Assume that we search for mutations in a genome consisting of 10,000,000 positions. The five most significant genes have the p-values  $7 \times 10^{-12}$ ,  $2 \times 10^{-11}$ ,  $3 \times 10^{-9}$ ,  $3 \times 10^{-8}$  and  $7 \times 10^{-7}$ . Use the Bonferroni method to adjust the p-values. How many positions are significant if we require the FWER to be less than 0.01? (2p)

# Multiple testing

Assume that we perform  $m$  tests.

Performing  $m$  tests at a level  $\alpha$  where  $H_0$  is true result in  $m \times \alpha$  false positives (in average).

Correcting for multiple testing means controlling the number of false positives!

4a)

The multiple testing problem arises when several statistical hypotheses are tested. For each test, the probability for false positives will increase. In fact, if  $m$  hypotheses are tested at a significance level of  $\alpha$  we expect to have  $m \times \alpha$  false positives in average. The false positives can, if not properly handled, lead to incorrect interpretation of the results.

4. In bioinformatics, a large number of tests is often performed simultaneously. This gives rise to the multiple testing problem.
- (a) Describe the multiple testing problem. What can happen if it is not properly addressed? (2p)
  - (b) What is the family-wise error rate (FWER)? Also, describe the Bonferroni method and how it can be used to control the FWER (3p).
  - (c) Assume that we search for mutations in a genome consisting of 10,000,000 positions. The five most significant genes have the p-values  $7 \times 10^{-12}$ ,  $2 \times 10^{-11}$ ,  $3 \times 10^{-9}$ ,  $3 \times 10^{-8}$  and  $7 \times 10^{-7}$ . Use the Bonferroni method to adjust the p-values. How many positions are significant if we require the FWER to be less than 0.01? (2p)



# Multiple testing - FWER

## Family-wise error rate (FWER)

The probability of at least one false positive, i.e.

$$\text{FWER} = \text{Prob}(V > 0)$$

Under the null hypotheses ( $H_0$  is true for the  $m$  tests)

$$\text{Prob}(V > 0) = \text{Prob}(\text{at least one false positive}) = 1 - (1 - \alpha)^m$$

# The Bonferroni correction

## **Algorithm:** Bonferroni correction

Divide the significance level  $\alpha$  (p-value cut-off) by the number of performed test (i.e.  $m$ ).

A Bonferroni adjusted p-value can be calculated by multiply each p-value with the number of performed tests (i.e.  $m$ ).

It can be shown that Bonferroni corrected p-values always control the Family Wise Error Rate (using Boole's inequality).

Interpretation: If we perform  $m$  tests and use the significance cut-off  $\alpha/m$ , then  $\text{Prob}(V > 0) \leq \alpha$ .

4b)

The family-wise error rate (FWER) is the probability that we have at least one false positive. By controlling the FWER (make sure that it is below a certain value), we will limit the number of false positives.

The Bonferroni correction method controls the FWER. The method can either modify the significance level by dividing it by the number of tests. Alternatively, the p-values can be adjusted by multiplying them with the number of tests. The Bonferroni method ensures that the FWER is then below the significance level. Note, however, that this method often is considered to be too conservative.

4. In bioinformatics, a large number of tests is often performed simultaneously. This gives rise to the multiple testing problem.
- (a) Describe the multiple testing problem. What can happen if it is not properly addressed? (2p)
  - (b) What is the family-wise error rate (FWER)? Also, describe the Bonferroni method and how it can be used to control the FWER (3p).
  - (c) Assume that we search for mutations in a genome consisting of 10,000,000 positions. The five most significant genes have the p-values  $7 \times 10^{-12}$ ,  $2 \times 10^{-11}$ ,  $3 \times 10^{-9}$ ,  $3 \times 10^{-8}$  and  $7 \times 10^{-7}$ . Use the Bonferroni method to adjust the p-values. How many positions are significant if we require the FWER to be less than 0.01? (2p)

4c)

Original p-values:

$7 \times 10^{-12}$ ,  $2 \times 10^{-11}$ ,  $3 \times 10^{-9}$ ,  $3 \times 10^{-8}$  and  $7 \times 10^{-7}$ .

Number of tests: 10,000,000

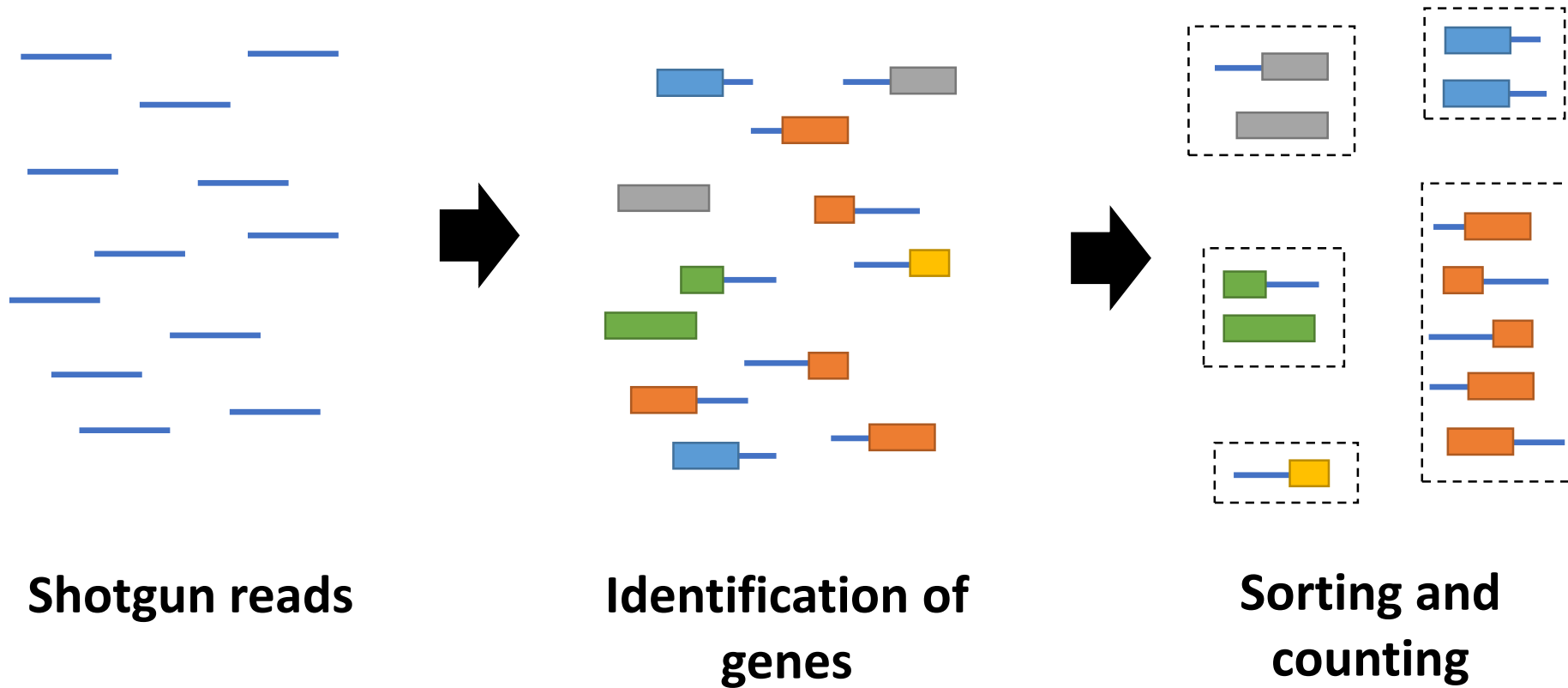
Adjusted p-values:

$7 \times 10^{-5}$ ,  $2 \times 10^{-3}$ ,  $3 \times 10^{-2}$ ,  $3 \times 10^{-1}$  and 1.

Number of significant tests after adjustment: 2

5. Shotgun metagenomics is used to study bacterial communities by sequencing random fragments of their collective genome (metagenome).
- (a) What does binning mean? What is the difference between direct and reference-guided binning? Which approach would you recommend if you have very short sequence reads? Why?(3p)
  - (b) Why do we normalize data from shotgun metagenomics? Also, mention at least one type of systematic error that may be present in metagenomic data but is not present in transcriptomic data. (3p)

# Direct binning



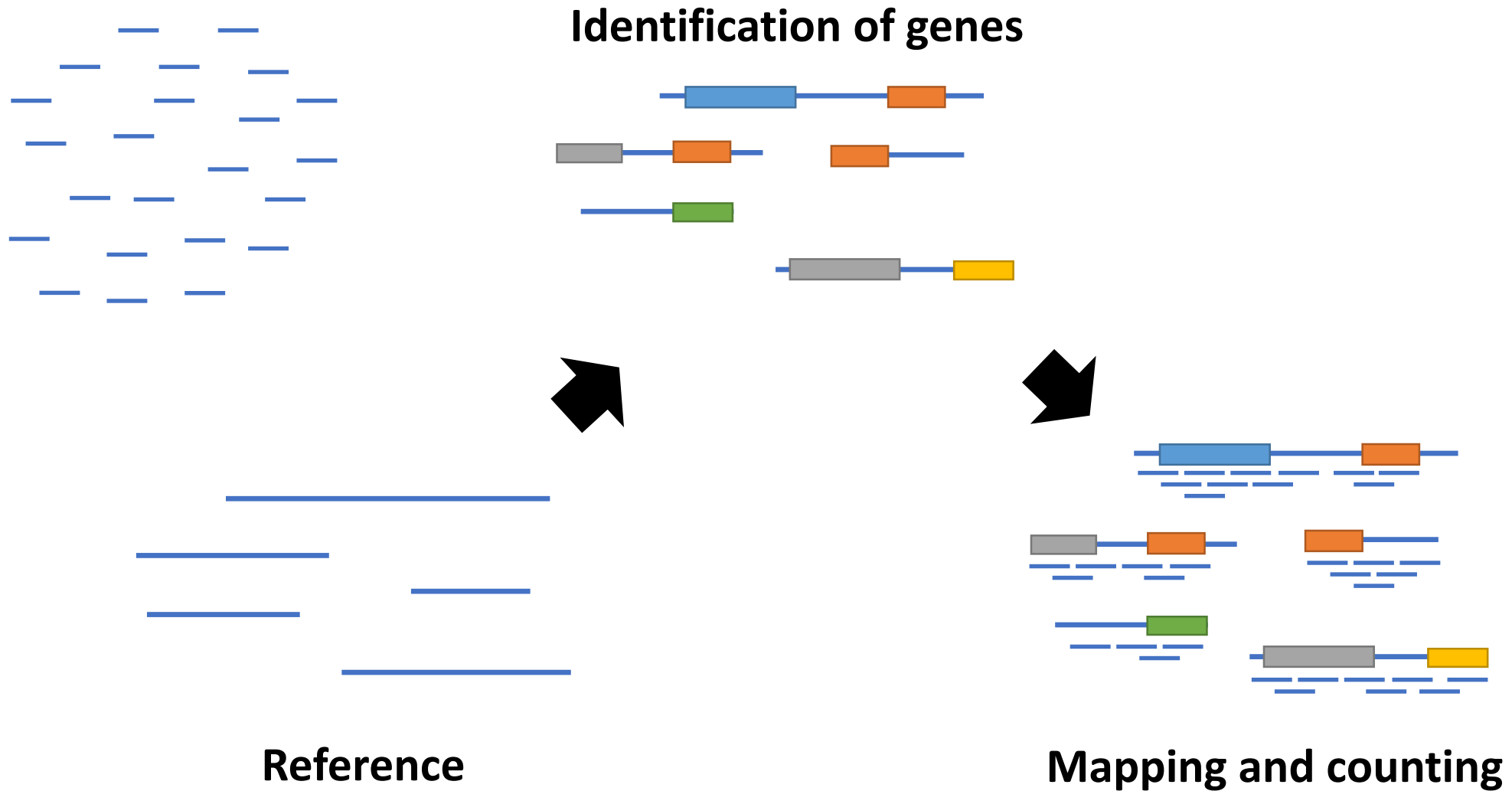
# Gene quantification

## 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.



# Reference-guided binning



# 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).

5a) The purpose of binning is to quantify gene abundances. This is done by placing reads in different 'bins' depending on their gene or gene function. The number of genes in each bin is a measure of the gene abundance.

Direct binning tries to identify genes directly from the reads. This approach is only applicable to reads that are of substantial length so they contain a substantial part of the gene.

Reference-guided binning uses a reference. The approach is similar to RNA-seq, where each read is mapped to the reference. By using an annotation of the reference, we can see what gene it contains and, based on the number of matching reads, quantify the abundances of these genes in the metagenome.

If you have very short reads, reference-guided binning should be used.

5. Shotgun metagenomics is used to study bacterial communities by sequencing random fragments of their collective genome (metagenome).
- (a) What does binning mean? What is the difference between direct and reference-guided binning? Which approach would you recommend if you have very short sequence reads? Why?(3p)
  - (b) Why do we normalize data from shotgun metagenomics? Also, mention at least one type of systematic error that may be present in metagenomic data but is not present in transcriptomic data. (3p)

# 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

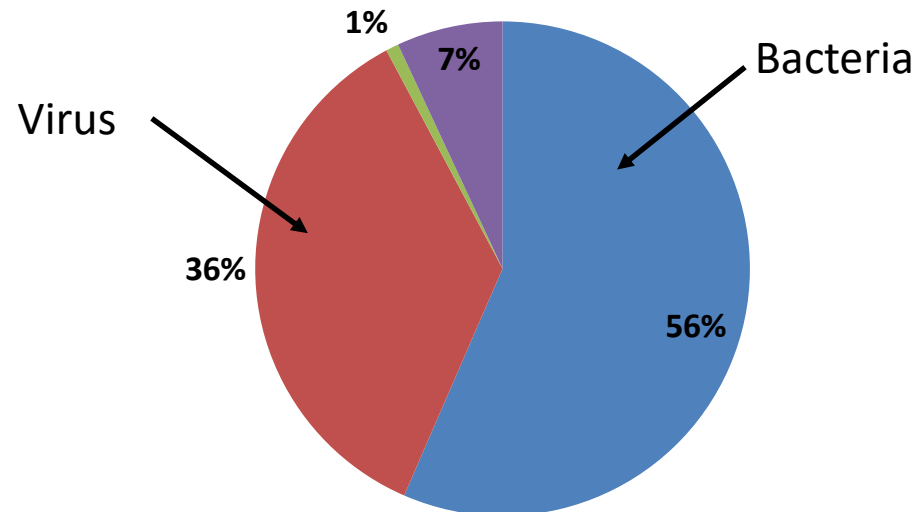
**Normalization is necessary to make samples comparable.**

# Normalization – example of systematic effects

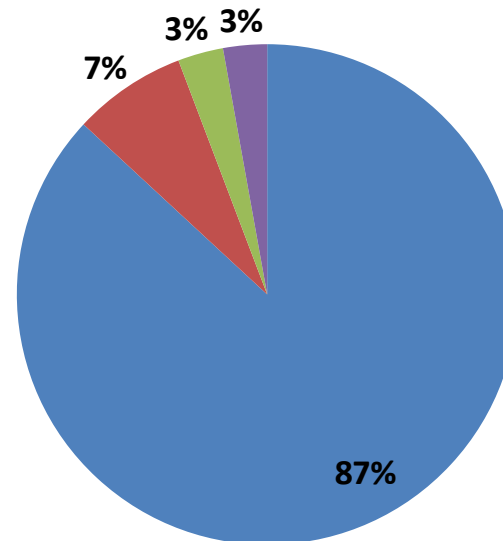
Metagenomic sequencing of river sediments upstream and downstream of a waste water treatment plant.



**Downstream Skövde**



**Upstream Skövde**



5b) Shotgun metagenomic data is affected by a wide range of systematic bias that make it hard to compare gene abundances, both between and within samples. Such systematic errors include for example, differences in sequencing depth and gene length. Normalization is used to remove these biases and thereby reduce the overall data variability. This makes it easier to identify differentially abundant genes.

One type of systematic error present in metagenomic data but typically not in transcriptomic data is differences in taxonomic composition. The proportion of eukaryotes, viruses and bacteria can differ between samples. These differences in taxonomic composition, will then affect the relative abundance of all bacterial genes. If not removed, this can significantly increase the variability thus reducing the statistical power.

6. Unsupervised methods are often used to find patterns in high-dimensional data.
- (a) What is the main difference between clustering and principal component analysis (PCA)? Give examples on when to use clustering and when to use PCA. (3p)
  - (b) How does hierarchical cluster analysis work? Describe the different steps of the algorithm and explain meaning of a 'distance metric' and a 'linkage criterion'. (3p)



# Unsupervised methods

- In this course we will look at two main methods for unsupervised data analysis
- Clustering
  - Aims find groups of samples or genes ('clusters') that are 'similar'.
  - Two important methods: Hierarchical clustering and K-means clustering.
- Principal component analysis (PCA)
  - Used for exploration and visualization of high-dimensional data
  - Identifies the dimensions with highest variability. These are used to represent the data at a lower dimension (e.g. in two dimensions).

# Hierarchical clustering

- Hierarchical clustering builds a hierarchy describing the similarity between clusters
- Agglomerative: All samples starts as individual clusters which are merged together (aggregated) according to
  - A distance measure describing the separation between data points
  - A linkage criterion measuring the distance between clusters
- Divisive: All samples starts in a single large cluster which is then successively divided
- The number of clusters are then defined based on a minimum similarity threshold

# Principal component analysis (PCA)

- Method used for exploration and visualization of high-dimensional data
- The main principle: Identify the dimensions of the data that explains as much variability as possible. These dimensions are defined by principal components.
- The data can then be projected on a set of the to principal components (often two) which results in a reduction of the dimensionality
- PCA is often used to identify patterns in the data. However, PCA does not assign each data point to a cluster.

6a) Clustering and PCA are both unsupervised methods. The aim of clustering is to identify clusters, i.e. group of genes or samples that shows similar patterns in the data. The aim of PCA is to produce a low-dimensional representation of the data. This can, for example, be used to visualize high-dimensional data in 2d- or 3d- plots. Since high variability sometimes corresponds to separation of data-points, PCA is useful to identify heterogeneity of the data. Note however, that PCA does not explicitly identify clusters.

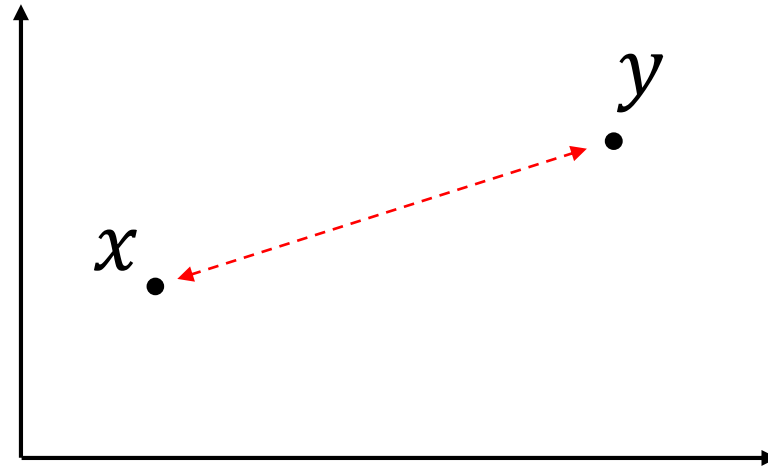
One example when clustering is used is to identify subgroups of samples in cancer research based on transcriptomics data. Here, each subgroup can correspond to different properties of the tumors.

One example where PCA is used is to analyze patterns from amplicon sequence data. Here, the PCA can be used to visualize the relationship between samples and thus demonstrate their relationships.

6. Unsupervised methods are often used to find patterns in high-dimensional data.
- (a) What is the main difference between clustering and principal component analysis (PCA)? Give examples on when to use clustering and when to use PCA. (3p)
  - (b) How does hierarchical cluster analysis work? Describe the different steps of the algorithm and explain meaning of a 'distance metric' and a 'linkage criterion'. (3p)

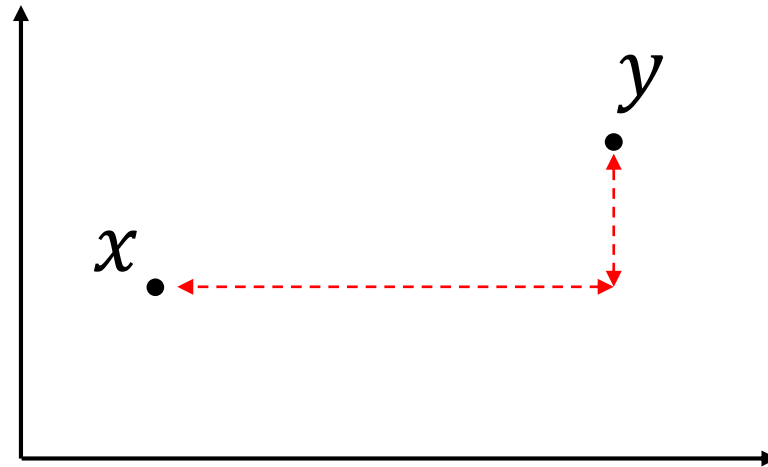
## *Euclidean distance*

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$



## *Manhattan distance*

$$d(x, y) = \sum_{i=1}^n |x_i - y_i|$$

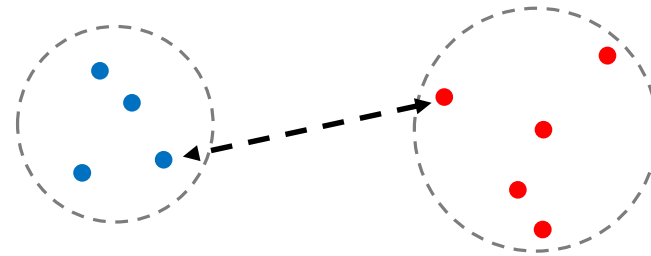


# Linkage criteria

The linkage criterion calculates the similarity between clusters.

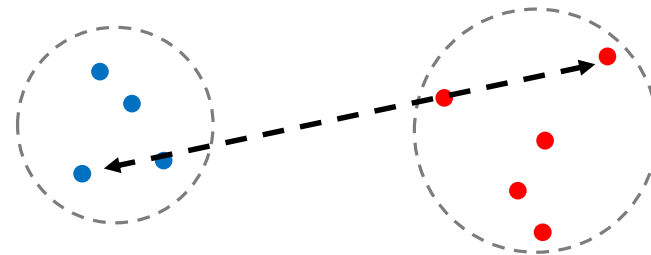
*Single linkage*

$$D(X, Y) = \min_{x \in X, y \in Y} d(x, y)$$



*Complete linkage*

$$D(X, Y) = \max_{x \in X, y \in Y} d(x, y)$$



# Hierarchical clustering

## Algorithm: Agglomerative hierarchical clustering

1. Calculate a distance matrix containing the pair-wise distance between all data point
2. Let each data point be a cluster
3. Identify the most similar clusters  $X$  and  $Y$  according to a *linkage criterion*.
4. Merge  $X$  and  $Y$  into a new cluster. Update the distance matrix.
5. Goto 3 if the number of clusters are more than 2.

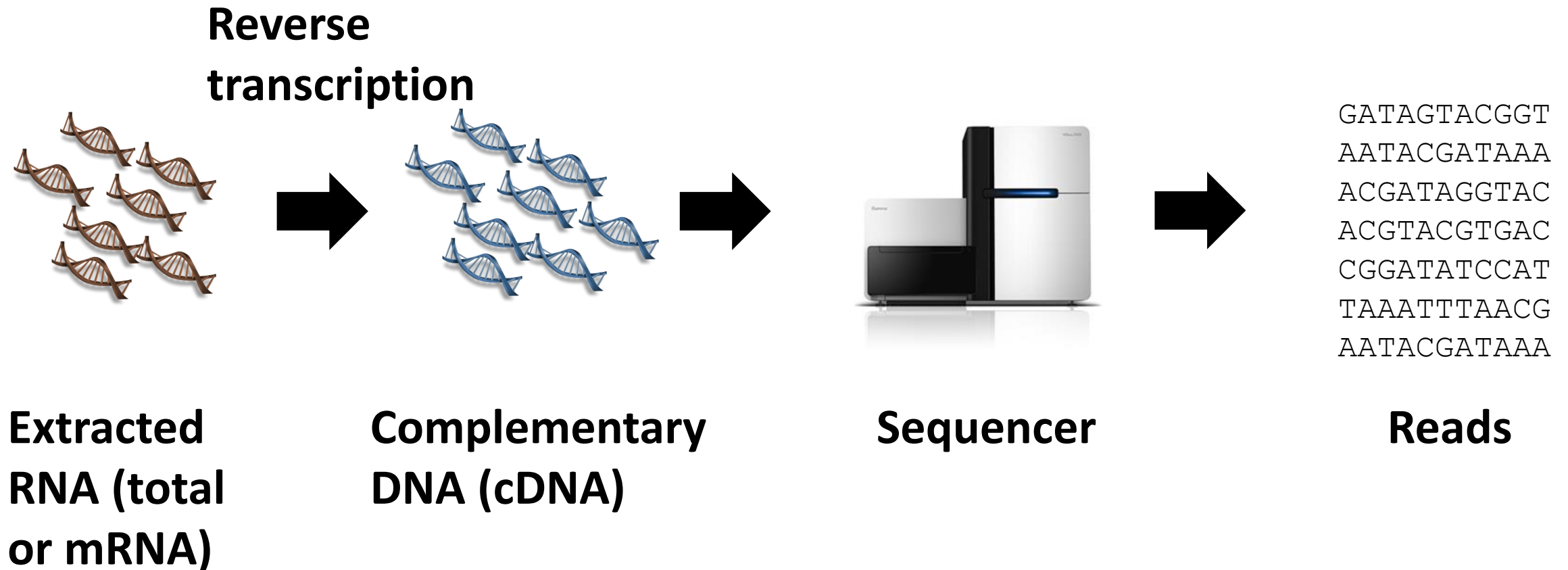


6b) Hierarchical clustering can be either agglomerative or divisive. In agglomerative clustering, each data-point starts as its own cluster. These are then merged successively. The distance between data-points are measured using a distance metric (e.g. Euclidean or correlation). The distance between clusters is measured using a linkage criterion (e.g. single linkage or complete linkage).

First a distance matrix, describing the pair-wise distances between all clusters are calculated. In each step, the algorithms identifies the clusters with the lowest distance. These clusters are then merged and the distance matrix is updated. This is repeated until all data-points are in one single cluster. The process can be visualized as a dendrogram, which can be used to select a suitable number of clusters.

7. In RNA-seq, RNA is converted to cDNA which is then sequenced. This results in reads that are randomly selected from the RNA molecules in the sample.
- (a) Briefly describe how reads from RNA-seq can be used to estimate gene expression (i.e. RNA abundance). (2p)
  - (b) Explain the difference between quantifying the expression of genes, exons and isoforms. (2p)
  - (c) Data from RNA-seq is often said to be semiquantitative. Explain what this means and its implications, i.e. what can we not measure using RNA-seq because of that the data is semiquantitative? (2p)

# The RNA-seq process



7a) RNA-seq works by randomly selecting RNA-fragments, which are reverse-transcribed and sequenced. The resulting reads corresponds thus to random positions of the RNA expressed in the sample. For a particular gene, the number of reads is a measure of its abundance. This is derived by mapping the reads to a reference, and counting the number of reads matching each read.

7. In RNA-seq, RNA is converted to cDNA which is then sequenced. This results in reads that are randomly selected from the RNA molecules in the sample.
- (a) Briefly describe how reads from RNA-seq can be used to estimate gene expression (i.e. RNA abundance). (2p)
  - (b) Explain the difference between quantifying the expression of genes, exons and isoforms. (2p)
  - (c) Data from RNA-seq is often said to be semiquantitative. Explain what this means and its implications, i.e. what can we not measure using RNA-seq because of that the data is semiquantitative? (2p)

# Quantification of gene abundances

## Common references

- Genome
  - Requires high quality genome assembly
  - The mapper needs to be able to handle splicing (splice-aware)



- *De novo* assembled transcriptome
  - Construction of the transcripts from the sequence data
  - Hard to identify splicing. Representing mainly mature spliced transcripts.

# Quantification of gene abundances

## *Gene quantification*

- Count the total number of fragments for each gene

## *Exon quantification*

- Counting the number of fragments for each exon.
- Splicing can be identified from fragments mapping junctions

## *Isoform quantification*

- Estimates the abundance for each isoform.
- Requires mathematical models and complementary data such as a database of known splice variants.

7b) In RNA-seq, the quantification can be done at different resolution. When quantifying the overall expression of a gene, we count all reads matching any part of the gene.

It is also possible to quantify individual exomes. In this case, we calculate the number of reads for each exome. This enables identification of specific splicing patterns.

It is finally possible to quantify the expression of specific isoforms. An isoform is a protein variant created by combining specific exons. By using a database of known isoforms, it is possible to based on information where the reads are mapping, predict abundance of the individual isoforms. This process is however complex and not possible to do accurately in species that lack information about the isoforms.



7. In RNA-seq, RNA is converted to cDNA which is then sequenced. This results in reads that are randomly selected from the RNA molecules in the sample.
- (a) Briefly describe how reads from RNA-seq can be used to estimate gene expression (i.e. RNA abundance). (2p)
  - (b) Explain the difference between quantifying the expression of genes, exons and isoforms. (2p)
  - (c) Data from RNA-seq is often said to be semiquantitative. Explain what this means and its implications, i.e. what can we not measure using RNA-seq because of that the data is semiquantitative? (2p)

# RNA-seq data is semiquantitative

Counts			Proportions		
	Sample 1	Sample 2		Sample 1	Sample 2
Gene 1	10	10	Gene 1	0.2	0.2
Gene 2	10	10	Gene 2	0.2	0.2
Gene 3	10	10	Gene 3	0.2	0.2
Gene 4	10	10	Gene 4	0.2	0.2
Gene 5	10	10	Gene 5	0.2	0.2
<b>Total</b>	<b>50</b>	<b>50</b>	<b>Total</b>	<b>1</b>	<b>1</b>

# RNA-seq data is semiquantitative(!!)

Counts				Proportions			
	Sample 1	Sample 2			Sample 1	Sample 2	
Gene 1	10	30	Up-regulated gene	Gene 1	0.2	0.6	↑
Gene 2	10	5		Gene 2	0.2	0.1	↓
Gene 3	10	5		Gene 3	0.2	0.1	↓
Gene 4	10	5		Gene 4	0.2	0.1	↓
Gene 5	10	5		Gene 5	0.2	0.1	↓
Total	50	50		Total	1	1	

7c) RNA-seq data is semiquantitative, which means that even if we can quantify gene abundances, we can only do this in relation to the abundances of other genes. Information about the absolute abundance, i.e. how many RNA molecule present in the sample, is not measured.

This has important consequences in how we interpret RNA-seq data. In particular, an increase of gene expression identified in RNA-seq can either be a result on a large number of RNA molecules for that specific genes or, alternatively, a decrease of RNA molecules of one of several other genes. Similarly, a decrease in measured gene expression could corresponds to a lower number of RNA molecules for the specific gene or an increase in RNA-molecules in one or several other genes. This means that we can not relate the differences that we see to absolute changes in RNA-molecules, which makes RNA-seq data hard to interpret.