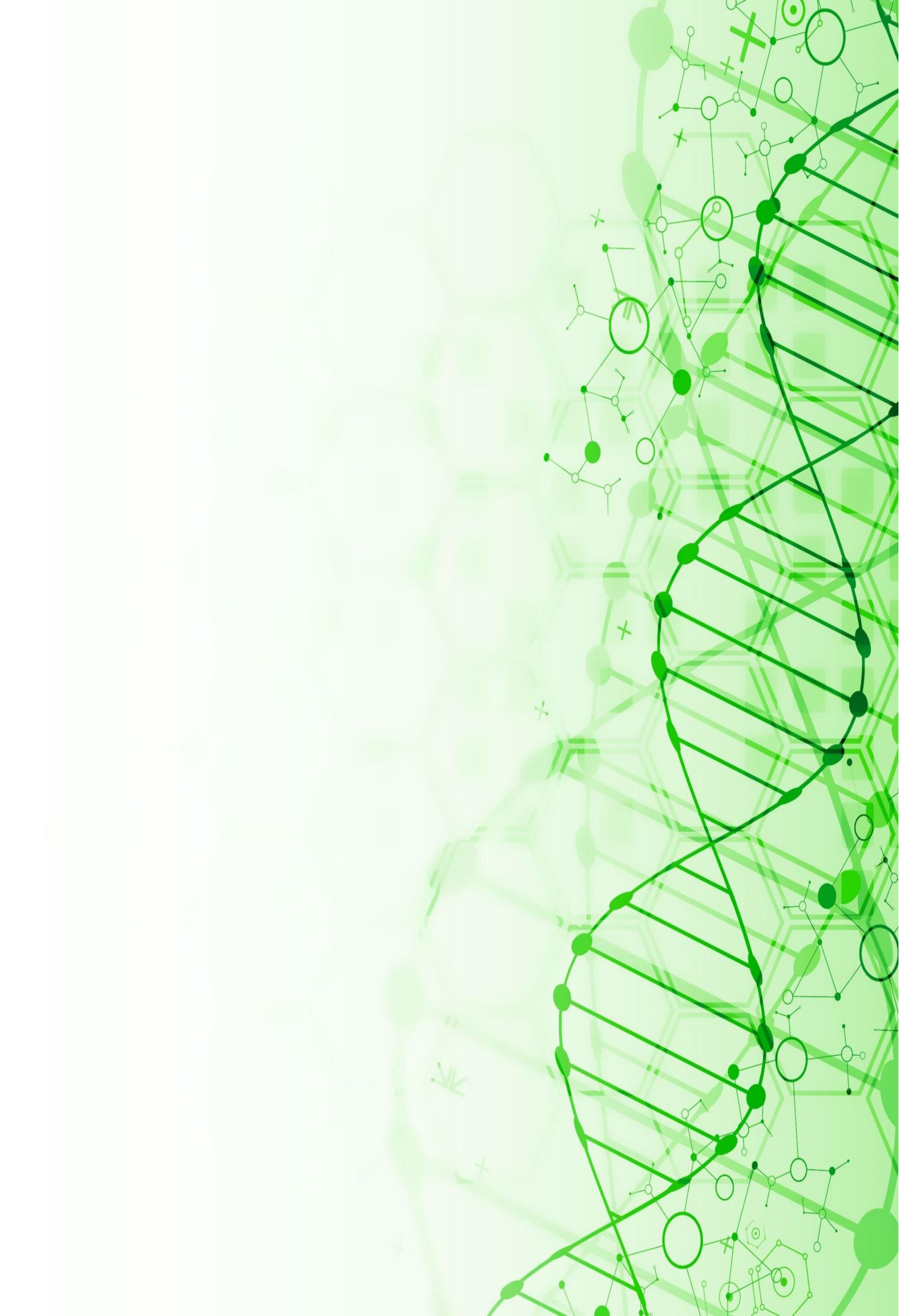


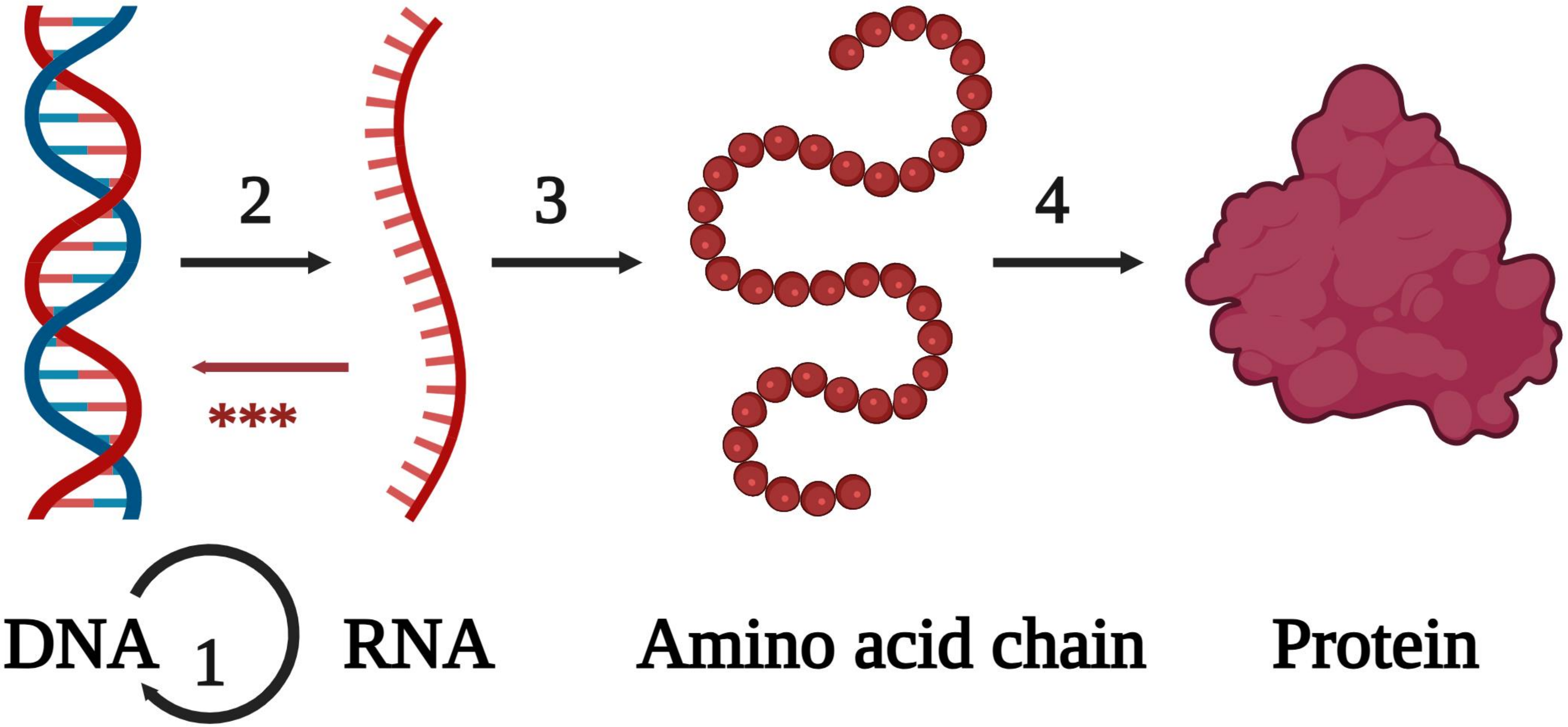
Introduction to microbial community profiling using amplicon sequencing

Dr. Pankaj Singh

School of Agriculture and Environment

University of Western Australia

Central Dogma



1 DNA replication
2 Transcription
3 Translation
4 Protein folding

*** Reverse transcription

Population – group of individuals of the same species living in the same area, potentially interacting

Community – group of **populations** of different species living in the same area, potentially interacting

What are some ecological interactions?

Why are ecological interactions important?

Ecological interaction can shape up distribution/ diversity/abundance of any organism within a population

Diversity, richness and evenness

- Diversity indicates like how many different type of species are in present within a community

Alpha diversity--diversity on a local scale, describing the species diversity (richness) within a functional community
Beta diversity--describes the rate at which species composition changes across a region

- Richness quantifies how many species does a population contains.

- Evenness refers to how closer the total number of individual species are present within a population. Lower the dominance of individual species better will the evenness of any populations

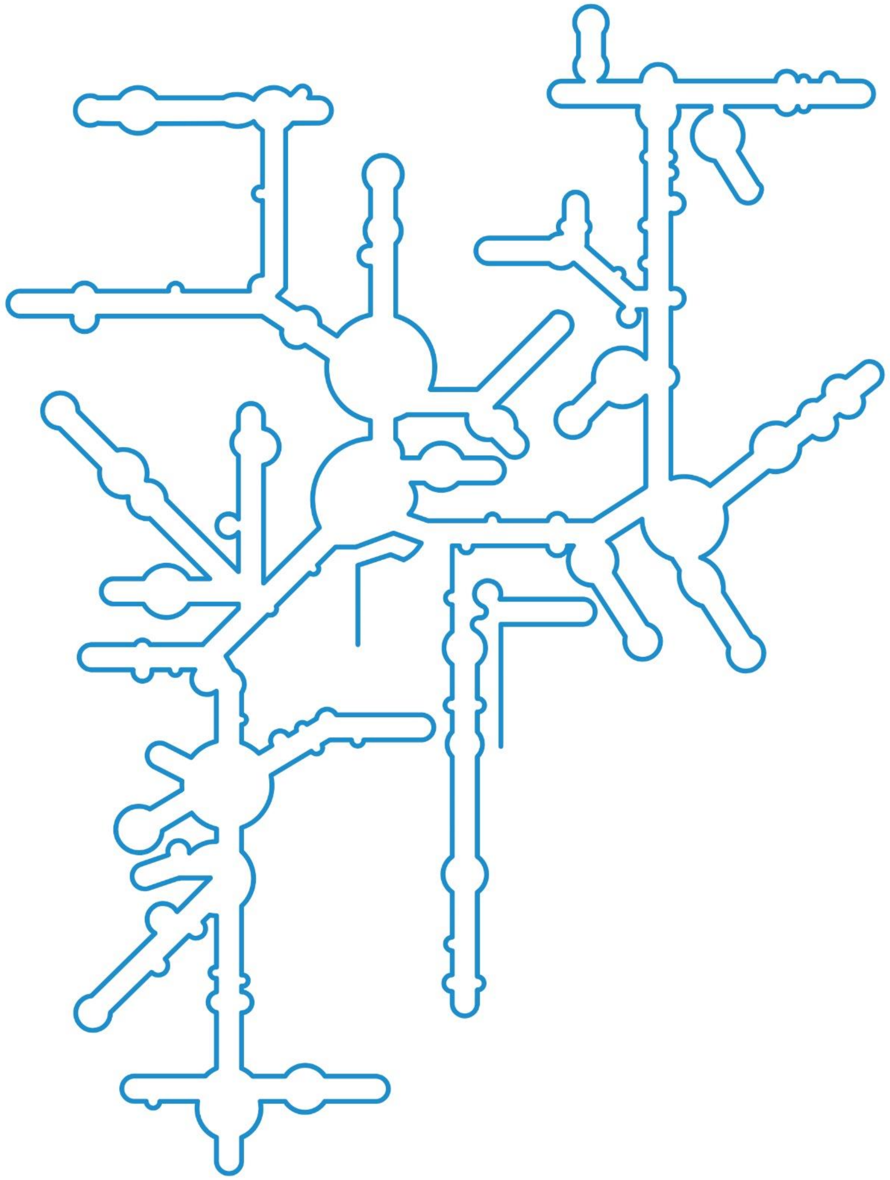
Symbiosis and co-evolution

- ❑ Researchers have challenged Darwinism on the basis of theory of symbiosis and co-evolution, **Lynn Margulis** was one of them.
- ❑ Its based upon interaction of two species (components) and their evolution for countering each other or for existence in a symbiotic way.
- ❑ This contradicts the Darwinian view that evolution occurs mainly as a result of competition between species.
- ❑ The organisms form a symbiotic partnership, typically by one engulfing the other– a process known as endosymbiosis. Dramatic evolutionary changes result.

Why is amplicon sequencing exciting ?

- Only about 1-2 % of microbes can be cultured using conventional laboratory practices
- That means majority of microbial flora remains unidentified. Hence their role and functions remain unresolved .
- NGS techniques allow the reading of DNA from uncloned samples.

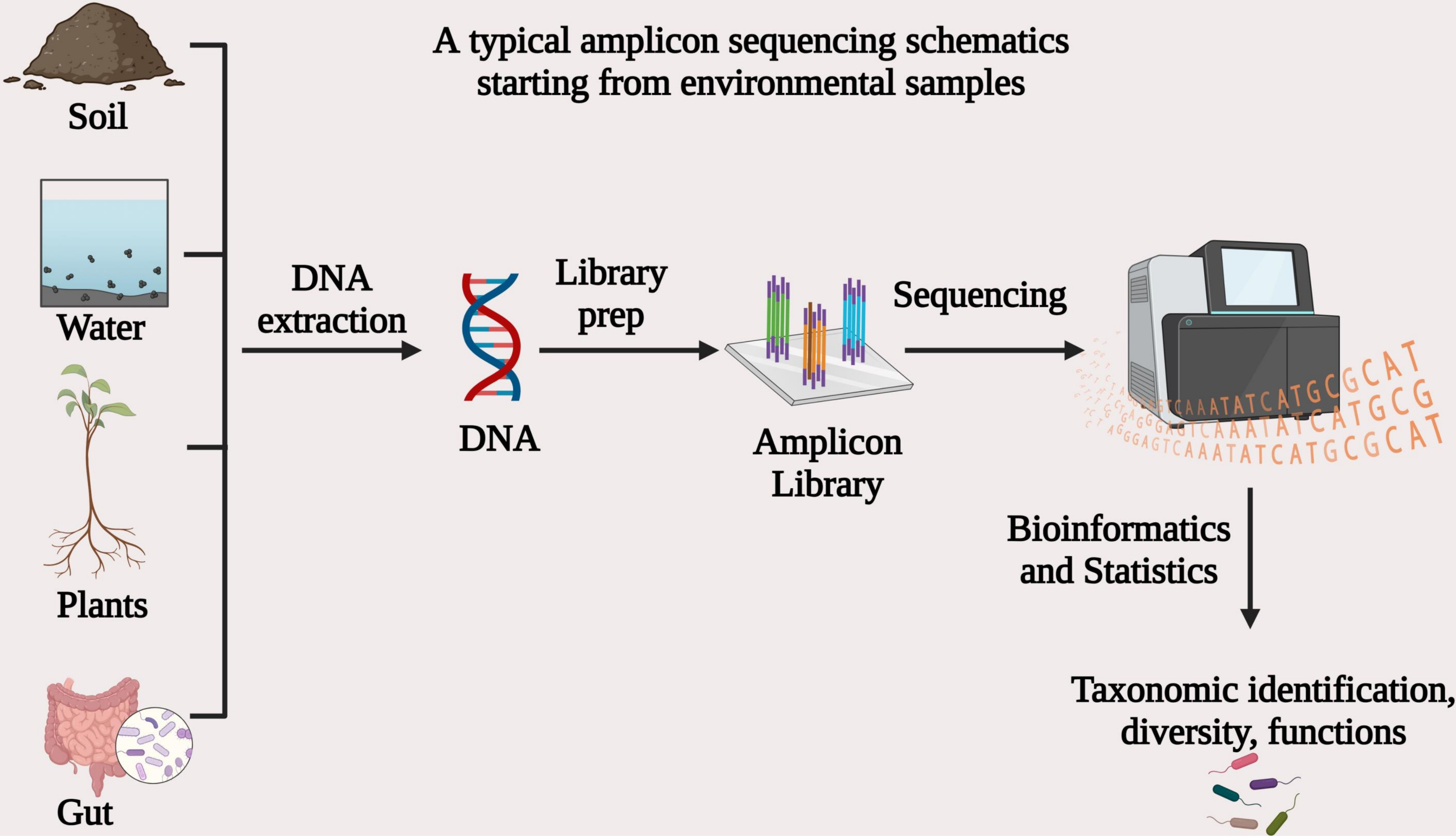
What to use for bacterial identification



16S rRNA a molecular barcode

- Universal
- Undergone less mutation
- Horizontal gene transfer is not an issue
- Conserved and has multiple variable regions for targeted amplification (V1 to V9)

A typical amplicon sequencing schematics starting from environmental samples



Soil

Water

Plants

Gut

DNA extraction

DNA

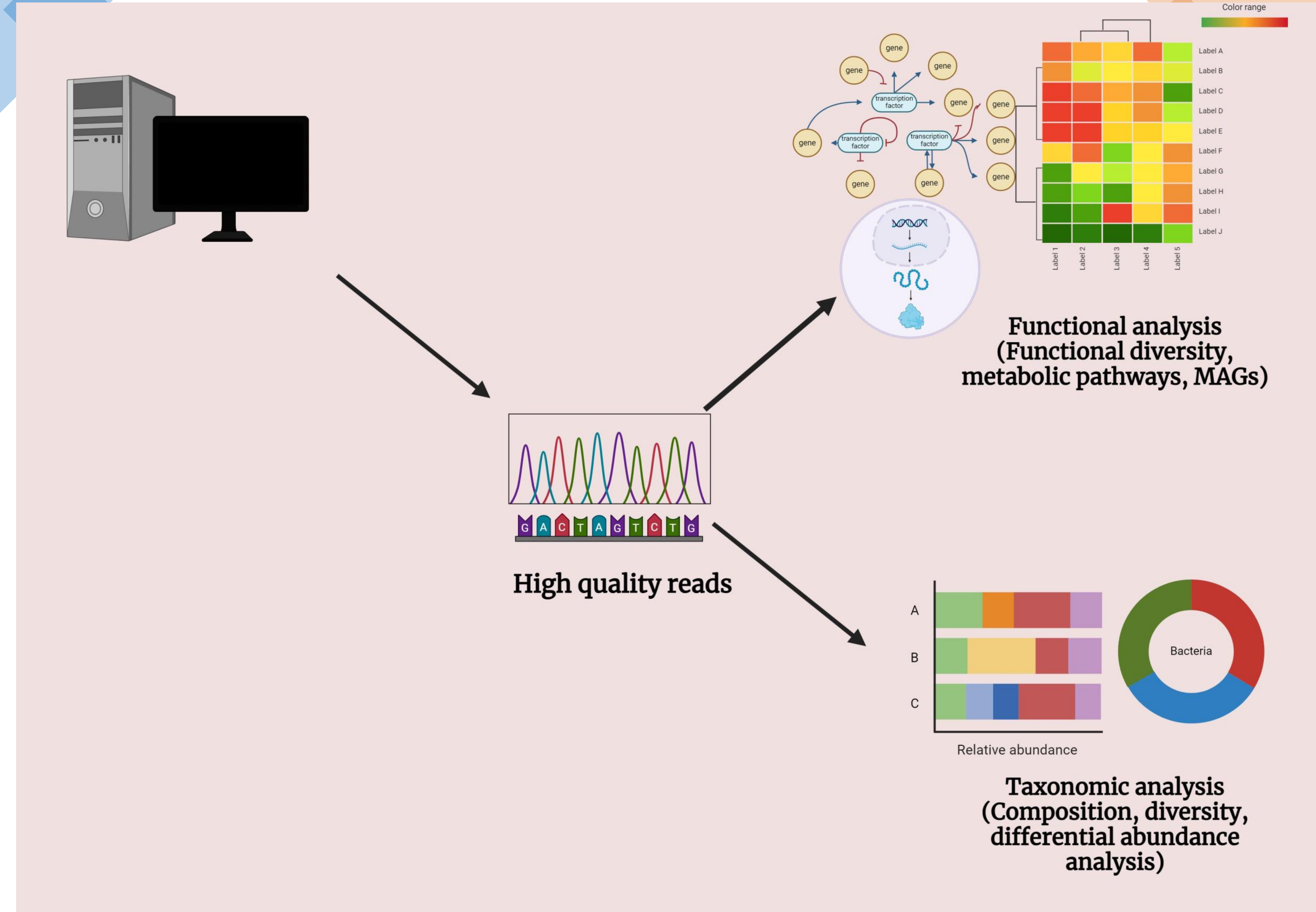
Library prep

Amplicon Library

Sequencing

Bioinformatics and Statistics

Taxonomic identification, diversity, functions



High quality reads

Functional analysis (Functional diversity, metabolic pathways, MAGs)

Taxonomic analysis (Composition, diversity, differential abundance analysis)

File formats in NGS

□ SRF

□ HDF5

➢ Helicos

➢ PacBio, Applied Biosystems, Oxford Nanophore

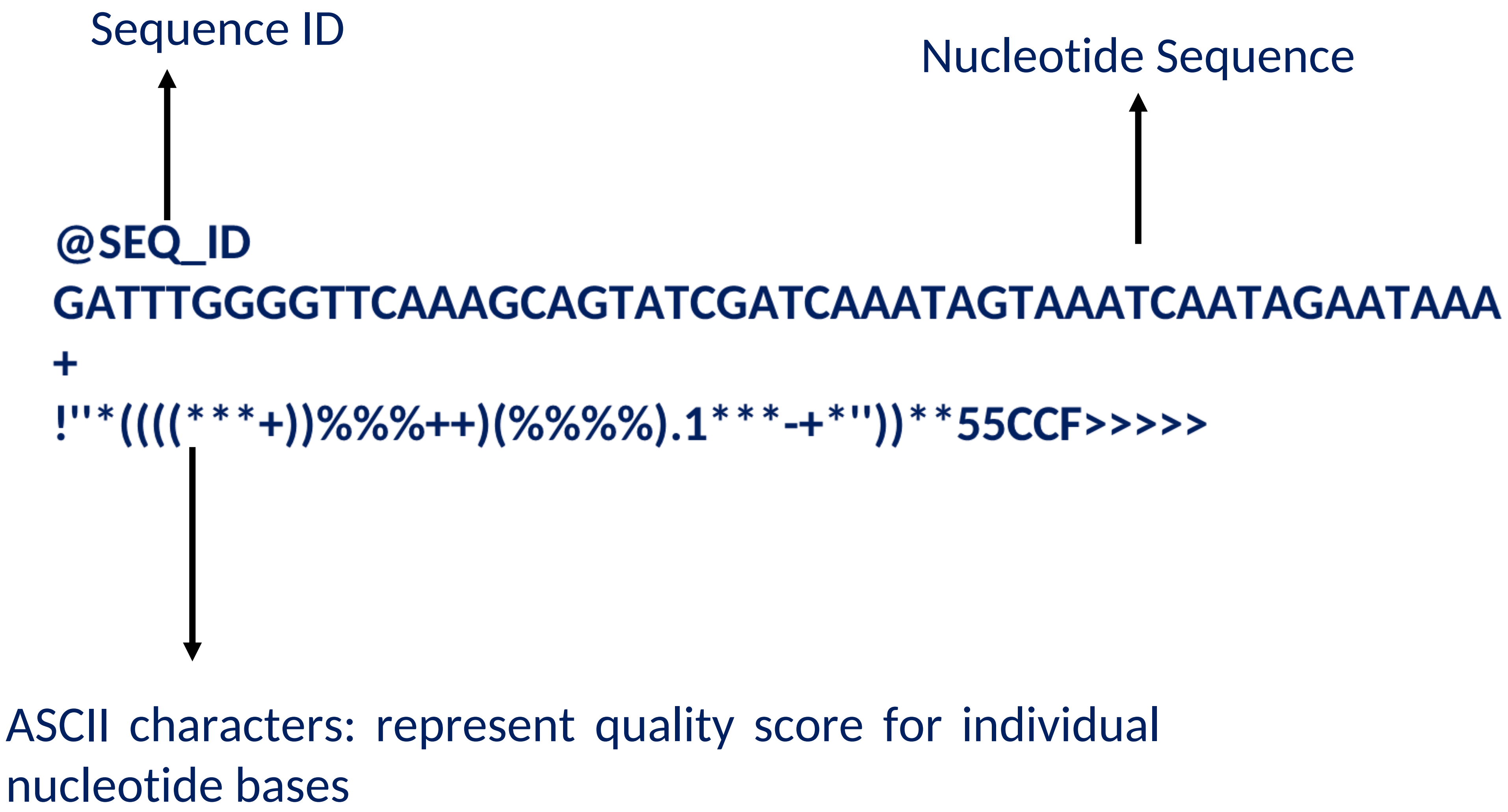
□ FASTQ

□ FASTA

➢ Most common(Illumina and others)

➢ No quality information

fastq sequence format



Sequence ID

Nucleotide Sequence

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCAATAGAATAAA
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>
```

ASCII characters: represent quality score for individual nucleotide bases

Phred quality (Q) Scores

❑ Quality (Q) scores represent probability of erroneous of a base call

For e.g. $Q = -10 log_{10} P \longrightarrow P = 10^{-Q/10}$

Q= 20 means error probability of $P = 10^{-2}$ = 1 in 100

Q= 30 means error probability of $P = 10^{-3}$ = 1 in 1000

❑ Better the Q Score lesser the chances of error better will be the data quality

❑ Q Score ranges from 0-93

FASTA FILE FORMAT

Just like @ indicates start of new sequence

Sequence nameC

```
>ERR010482.1 FT9FZH301B6YPS/3
ATCAACACATTAGGACTTACACGAATCAGGCATTCGTTACCAT
CAGTATGTCGAT

>ERR010482.2 FT9FZH301ARSRC/3
ATGCTTGCTCGGCCGACGTGAGCGTTATTCGAGCAGGGCTCG
GATGGTAGTTAGCGATCCAAAGGGGAGTC
```

There is no quality information

Quality Control

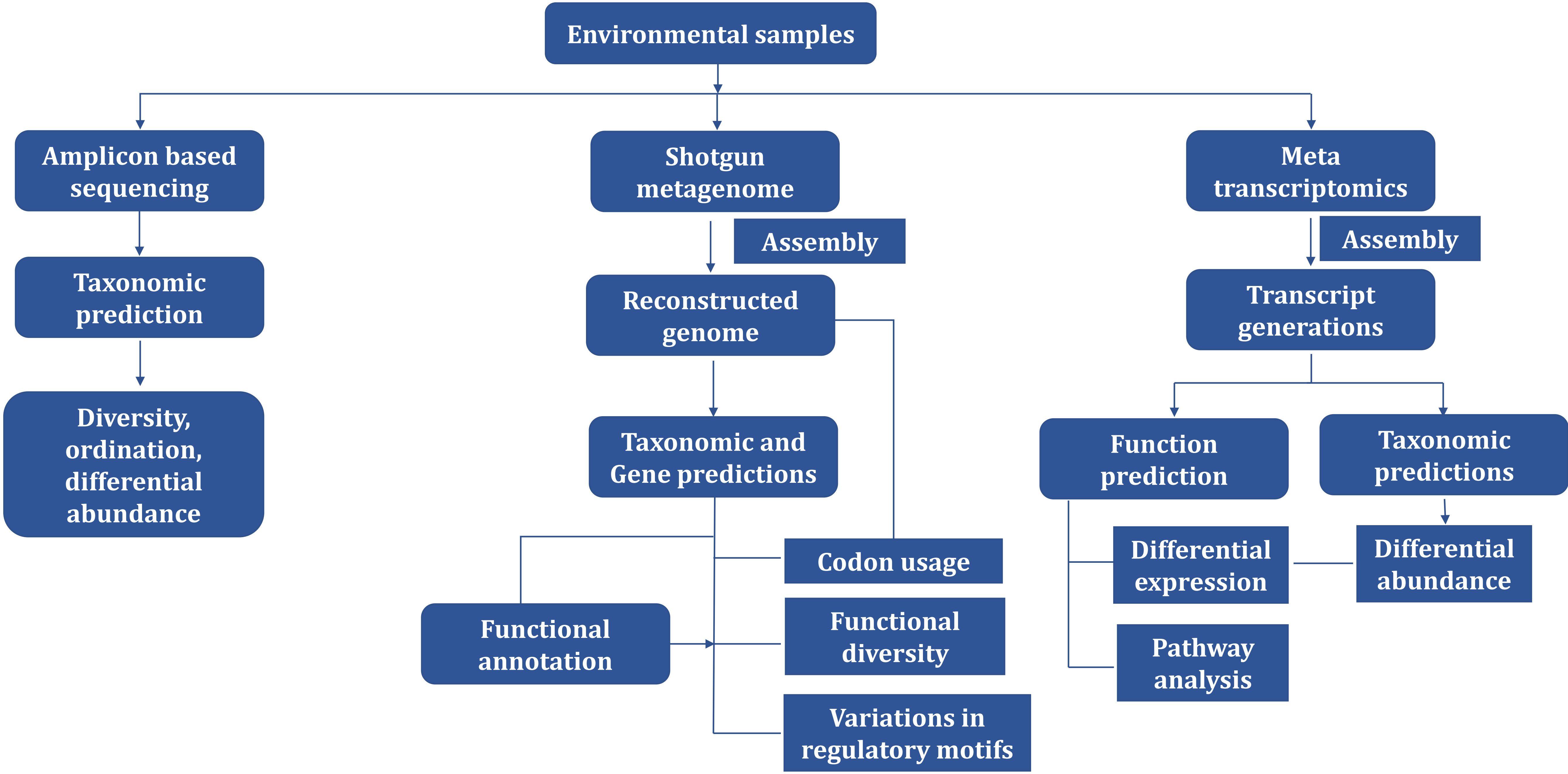
- Quality control is an important step in data processing
- Low quality reads below a defined threshold are removed from the dataset
- Only high-quality reads are processed further for bioinformatics
- General threshold is defined as Q=30, but it varies according to the source of samples and sequencing results

Garbage in

Garbage out

Quality of the data matters the most

Amplicon sequencing, metagenomics and meta-transcriptomics



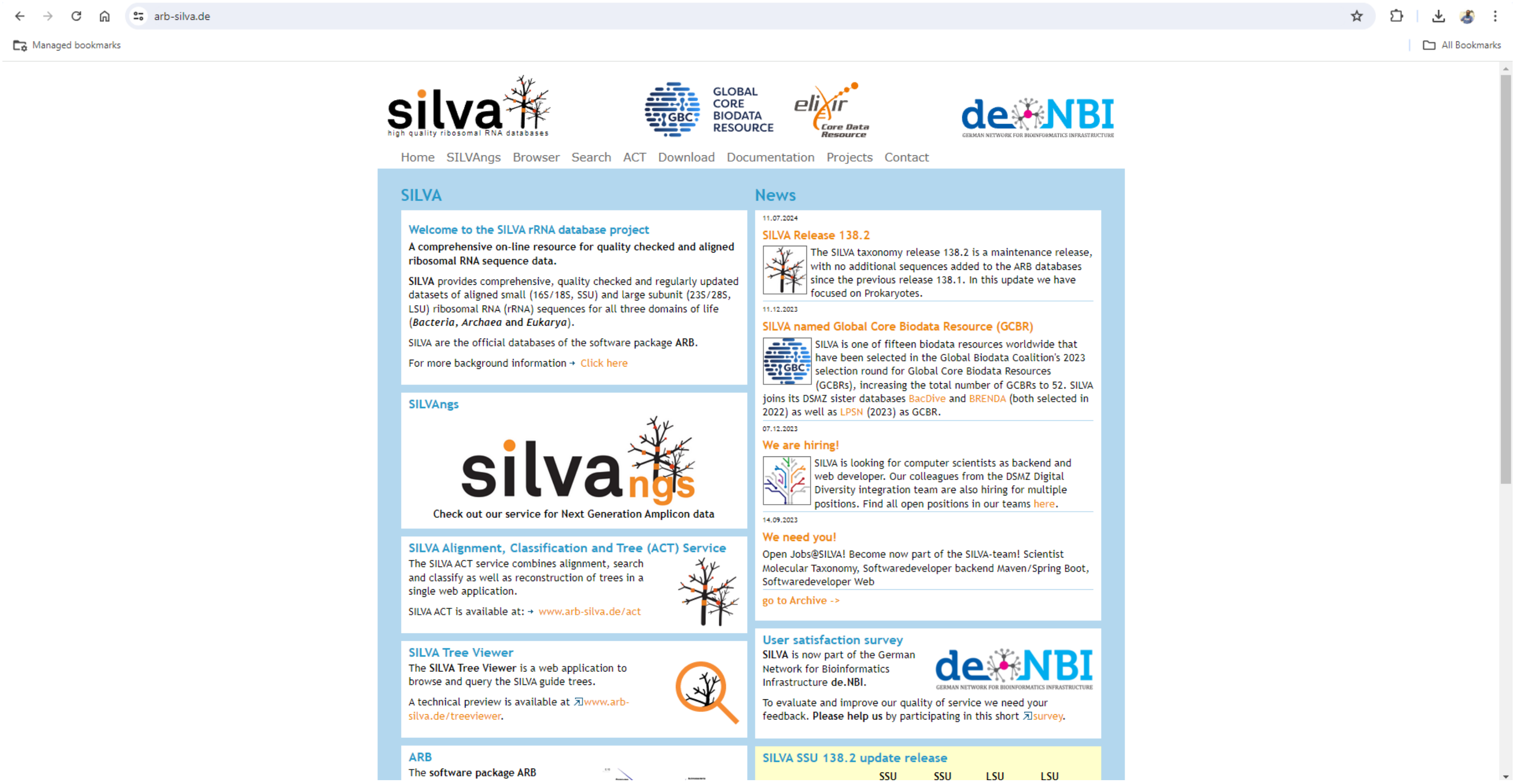
Taxonomic identification of processed reads

- ❑ The general rational behind taxonomic classification of sequences is based on the sequence similarity /homology
- ❑ Prokaryotes and eukaryotes are classified on different molecular barcodes such as 5S rRNA, 16S rRNA (Prokaryotes) 18SrRNA ,23S rRNA ITS (Internal Transcribed Spacer)
- ❑ Taxonomic annotation is performed against reliable databases. For eg
  Bacteria : SILVA, RDP
  Fungi : UNITE
  Specific databases based on samples: Anaerobic digestors : MIDAS database
  https://www.midasfieldguide.org/guide

Molecular barcode and proper primer selection for sequencing is very critical.

Databases for taxonomic annotation

https://www.arb-silva.de/


arb-silva.de
Managed bookmarks
All Bookmarks
silva
high quality ribosomal RNA databases
GLOBAL CORE BIODATA RESOURCE
elixir Core Data Resource
de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE
Home SILVAngs Browser Search ACT Download Documentation Projects Contact

SILVA

Welcome to the SILVA rRNA database project

A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.

SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (*Bacteria*, *Archaea* and *Eukarya*).

SILVA are the official databases of the software package **ARB**.

For more background information → Click here

SILVAngs

Check out our service for Next Generation Amplicon data

SILVA Alignment, Classification and Tree (ACT) Service

The SILVA ACT service combines alignment, search and classify as well as reconstruction of trees in a single web application.

SILVA ACT is available at: → www.arb-silva.de/act

SILVA Tree Viewer

The **SILVA Tree Viewer** is a web application to browse and query the SILVA guide trees.

A technical preview is available at www.arb-silva.de/treeviewer.

ARB

The software package ARB

News

11.07.2024

SILVA Release 138.2

The SILVA taxonomy release 138.2 is a maintenance release, with no additional sequences added to the ARB databases since the previous release 138.1. In this update we have focused on Prokaryotes.

11.12.2023

SILVA named Global Core Biodata Resource (GCBR)

SILVA is one of fifteen biodata resources worldwide that have been selected in the Global Biodata Coalition's 2023 selection round for Global Core Biodata Resources (GCBRs), increasing the total number of GCBRs to 52. SILVA joins its DSMZ sister databases BacDive and BRENDA (both selected in 2022) as well as LPSN (2023) as GCBR.

07.12.2023

We are hiring!

SILVA is looking for computer scientists as backend and web developer. Our colleagues from the DSMZ Digital Diversity integration team are also hiring for multiple positions. Find all open positions in our teams here.

14.09.2023

We need you!

Open Jobs@SILVA! Become now part of the SILVA-team! Scientist Molecular Taxonomy, Softwaredeveloper backend Maven/Spring Boot, Softwaredeveloper Web

go to Archive ->

User satisfaction survey

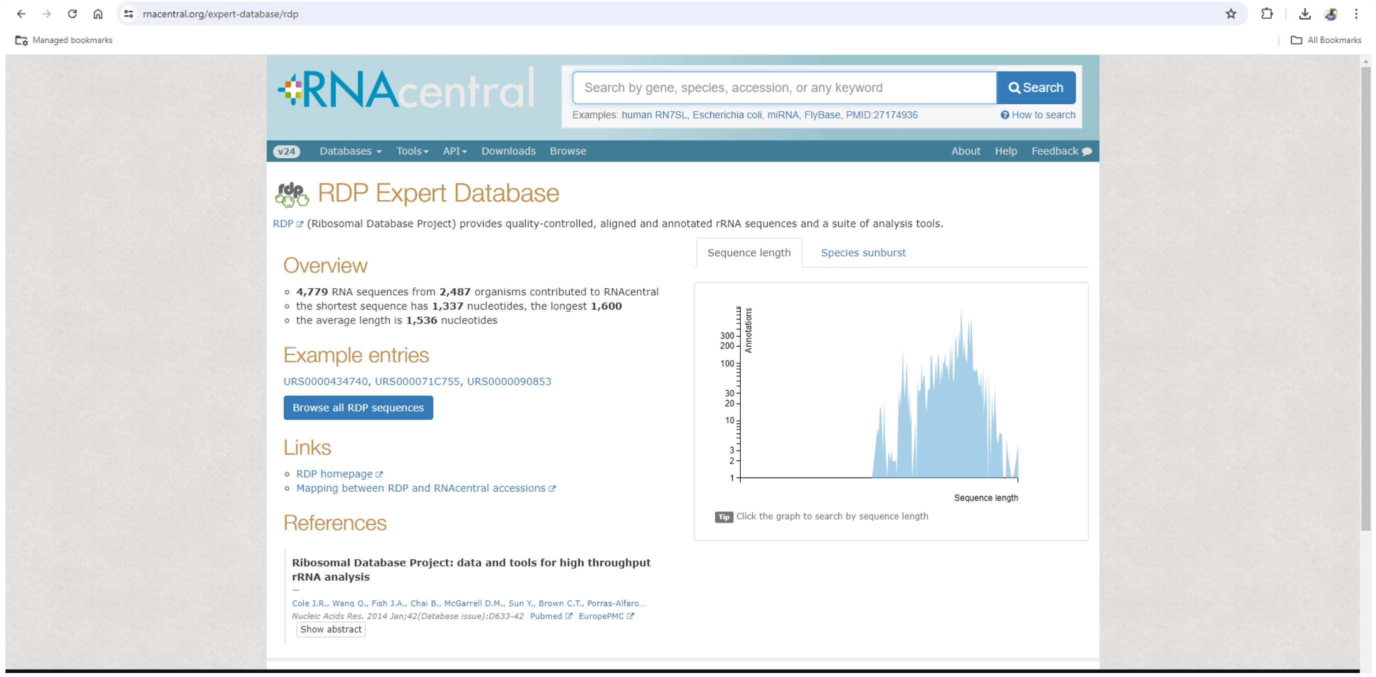
SILVA is now part of the German Network for Bioinformatics Infrastructure **de.NBI**.

To evaluate and improve our quality of service we need your feedback. **Please help us** by participating in this short survey.

SILVA SSU 138.2 update release

SSU SSU LSU LSU

https://rnacentral.org/expert-database/rdp



Examples: human RN7SL, Escherichia coli, miRNA, FlyBase, PMID:27174936

RDP Expert Database

RDP (Ribosomal Database Project) provides quality-controlled, aligned and annotated rRNA sequences and a suite of analysis tools.

Overview

- **4,779** RNA sequences from **2,487** organisms contributed to RNAcentral
- the shortest sequence has **1,337** nucleotides, the longest **1,600**
- the average length is **1,536** nucleotides

Example entries

URS0000434740, URS000071C755, URS0000090853

Browse all RDP sequences

Links

- RDP homepage
- Mapping between RDP and RNAcentral accessions

References

Ribosomal Database Project: data and tools for high throughput rRNA analysis

—

Cole J.R., Wang Q., Fish J.A., Chai B., McGarrell D.M., Sun Y., Brown C.T., Porras-Alfaro...

Nucleic Acids Res. 2014 Jan;42(Database issue):D633-42 Pubmed EuropePMC

Show abstract

Click the graph to search by sequence length

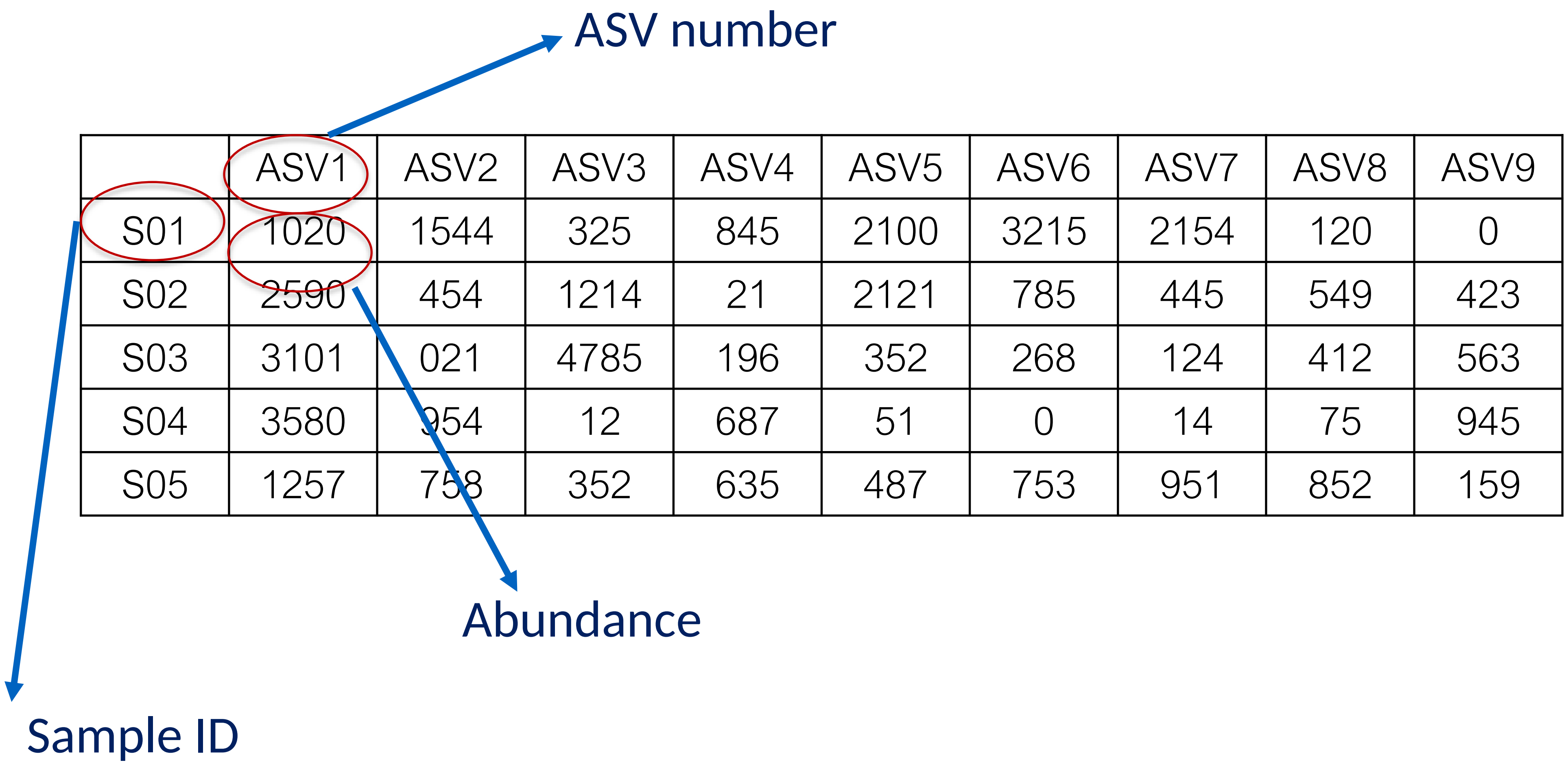
p.s. Greengenes has not been updated for longtime avoid using it

Common bioinformatics software for amplicon sequencing data analysis

- ❑ QIIME2 : Quantitative Insights Into Microbial Ecology 2
  - One of the popular tools
  - Unix/Linux depend, cannot use straightway on Windows
  - Need Docker or Virtual Machine on Windows
  - Good for visualizations
  - Requires good knowledge of Bash scripting and python
- ❑ USEARCH
  - Unix/Linux depend, cannot use straightway on Windows
  - Need Docker or Virtual Machine on Windows
  - Need good skill of Bash scripting
- ❑ Mothur
  - Platform independent
  - Light and works on Windows
  - Does not require installation

What do we get out of bioinformatics analysis

- ❑ OTU file : Classifying sequenced reads into Amplicon Sequencing Variants(ASVs) or Operational Taxonomic Units
- ❑ Information regarding how many time every sequence in the form of OTU/ASV has appeared-----this relates to abundance



ASV number

	ASV1	ASV2	ASV3	ASV4	ASV5	ASV6	ASV7	ASV8	ASV9
S01	1020	1544	325	845	2100	3215	2154	120	0
S02	2590	454	1214	21	2121	785	445	549	423
S03	3101	021	4785	196	352	268	124	412	563
S04	3580	954	12	687	51	0	14	75	945
S05	1257	758	352	635	487	753	951	852	159

Abundance

Sample ID

What do we get out of bioinformatics analysis cont….

Taxonomy file

❑ Taxonomic information for every OTU/ASV corresponding to the OTU file

❑ Taxonomic information starting from Domain to Genus (sometimes species)

Taxonomic information of individual ASVs

ASV number

	Kingdom	Phylum	Class	Order	Family	Genus
ASV1	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	NA
ASV2	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	NA
ASV3	Bacteria	Firmicutes	Bacilli	Bacillales	Planococcaceae	Sporosarcina
ASV4	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	NA
ASV5	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	NA
ASV6	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus
ASV7	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas

Most important file : Metadata file

- Most important part, designed by user
- Has all variable and factors from the experiment
- Can include plant, soil parameters, etc.

SampleID	SampleType	Plant	Stress	Timepoint
S01	Bulk Soil	Ryegrass	Control	T1
S02	Rhizosphere	Ryegrass	Heat	T1
S03	Root	Ryegrass	Control	T2
S04	Root	Lucerne	Control	T2

What basic stats to do post analysis

- ❑ Taxonomic composition of dataset
- ❑ Alpha diversity, richness indices based on metadata(experimental factors and variables)
- ❑ Beta diversity : Ordination plots
- ❑ Responsive ASVs/bacteria to factors : Indicator species analysis
  Differential abundance analysis
- ❑ Statistical modelling
- ❑ Application of machine learning : Random Forest

- ❑ Many more

All depends on your dataset and hypothesis

obrigado
Dank U
Merci
mahalo
Köszi
спасибо
Grazie
Thank you
mauruuru
Takk
Gracias
Dziękuję
Děkuju
danke
Kiitos