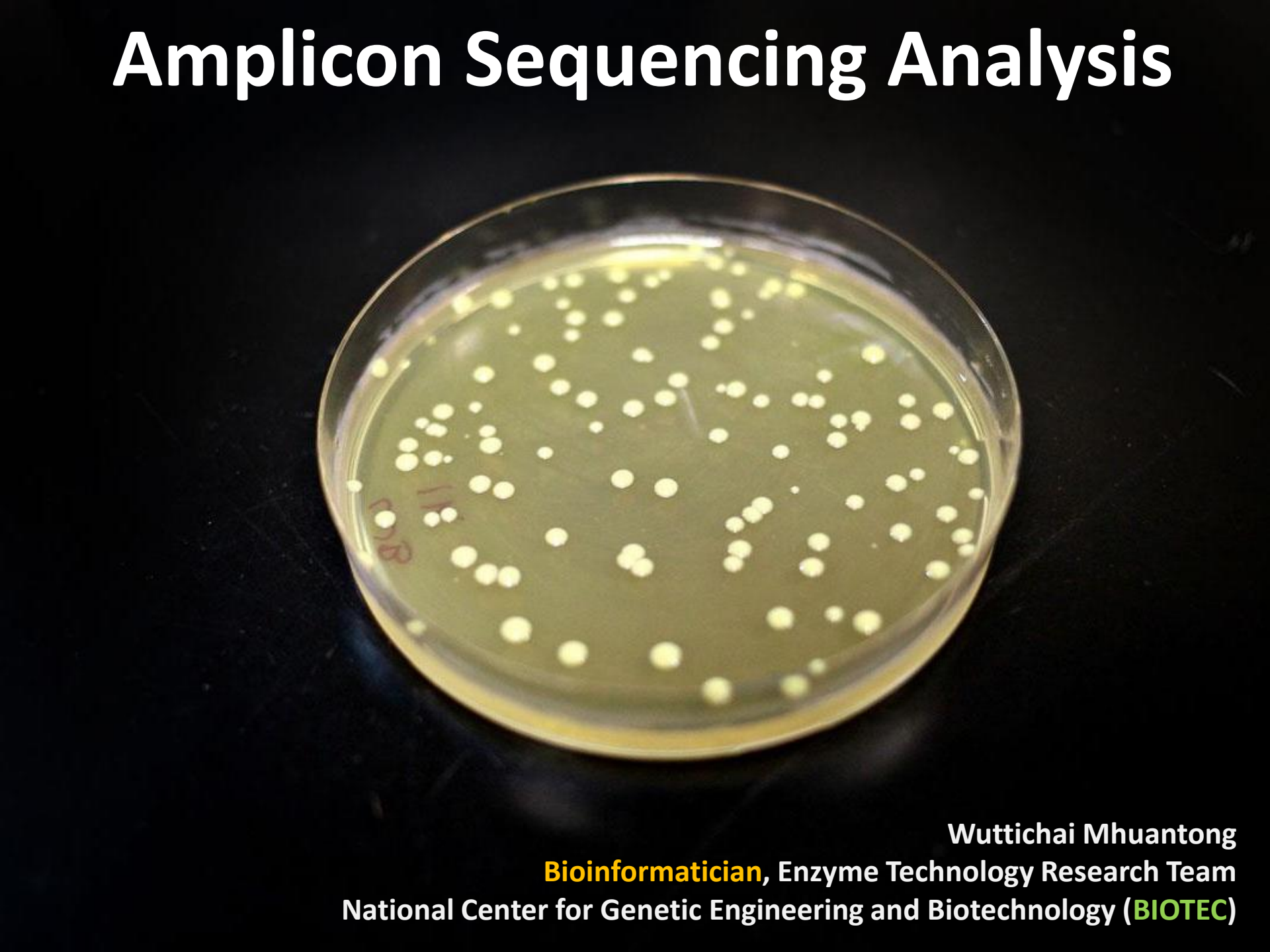
Amplicon Sequencing Analysis



Wuttichai Mhuantong
Bioinformatician, Enzyme Technology Research Team
National Center for Genetic Engineering and Biotechnology (BIOTEC)

Outline

Day1 (Feb/05)

- Introduction to metagenomics and data analysis

- Program installation

- Bacterial amplicon data analysis

Outline

Day2 (Feb/19)

- Fungal amplicon data analysis

- Database construction for amplicon data analysis

- Functional analysis

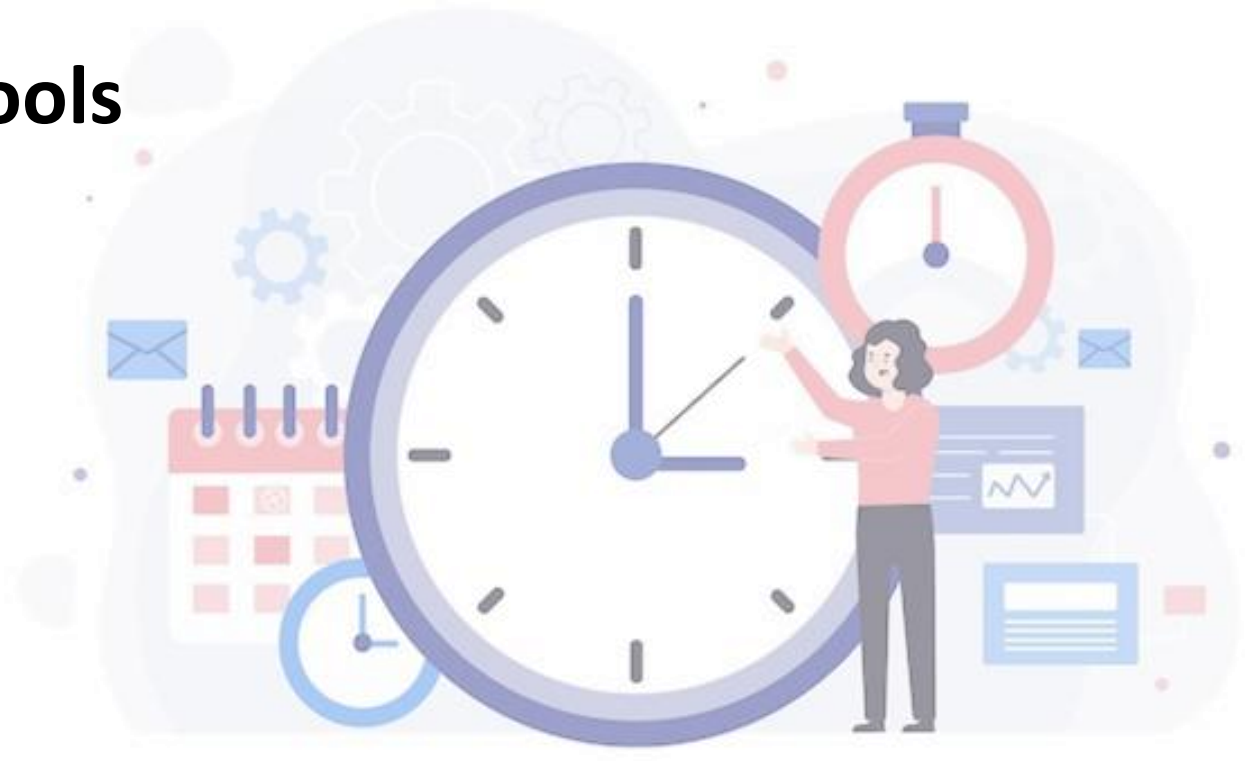
Outline

Day3 (Feb/26)

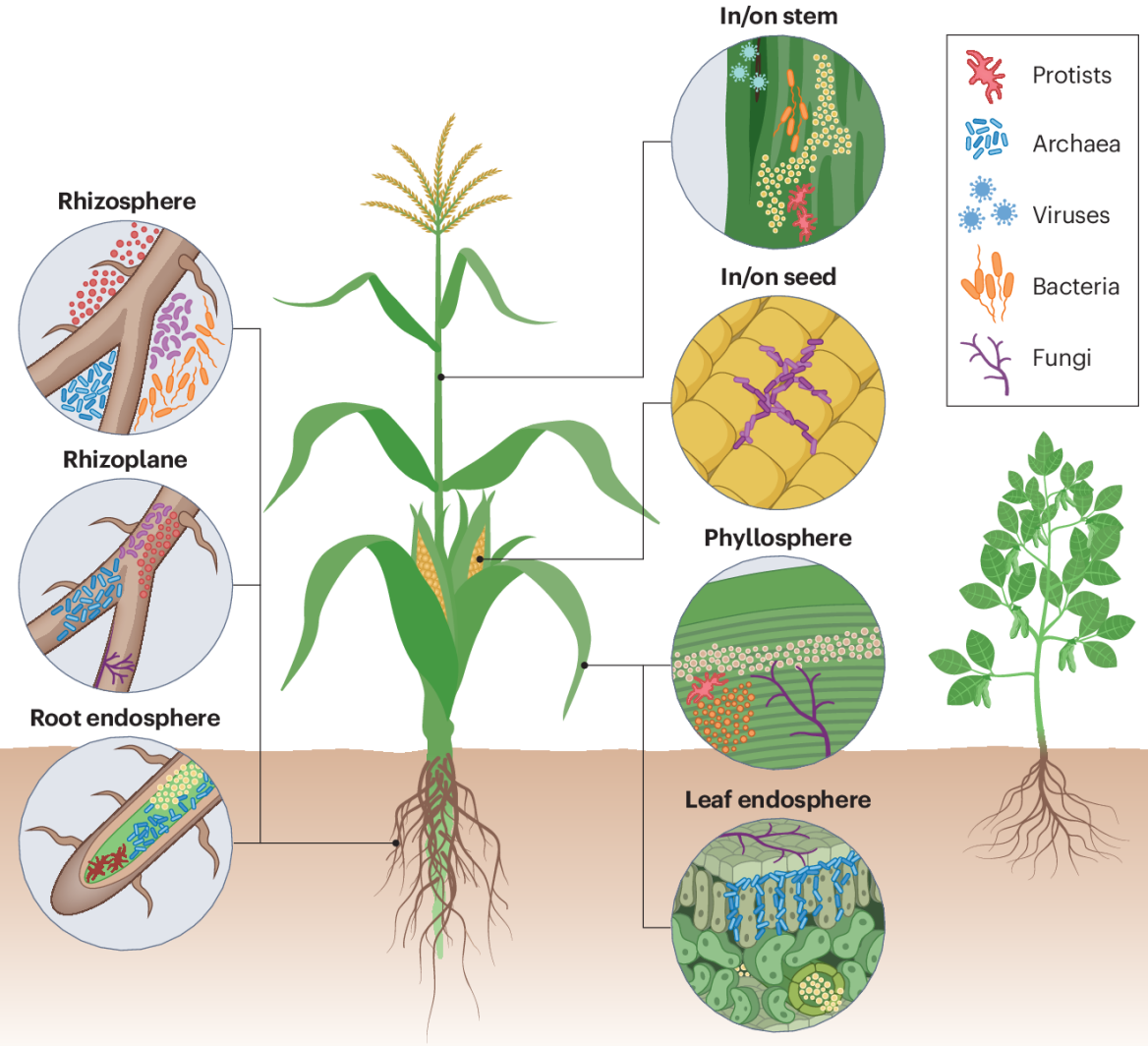
- Data visualization

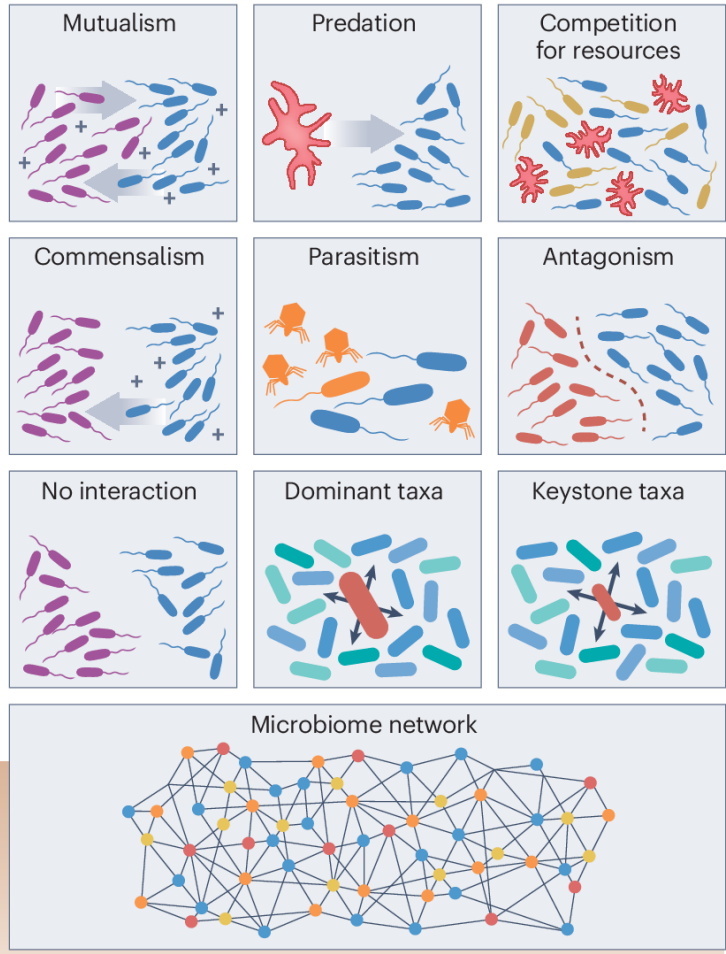
- Statistical analysis

- Online web tools



Why Do We Study Plant Microbiome?





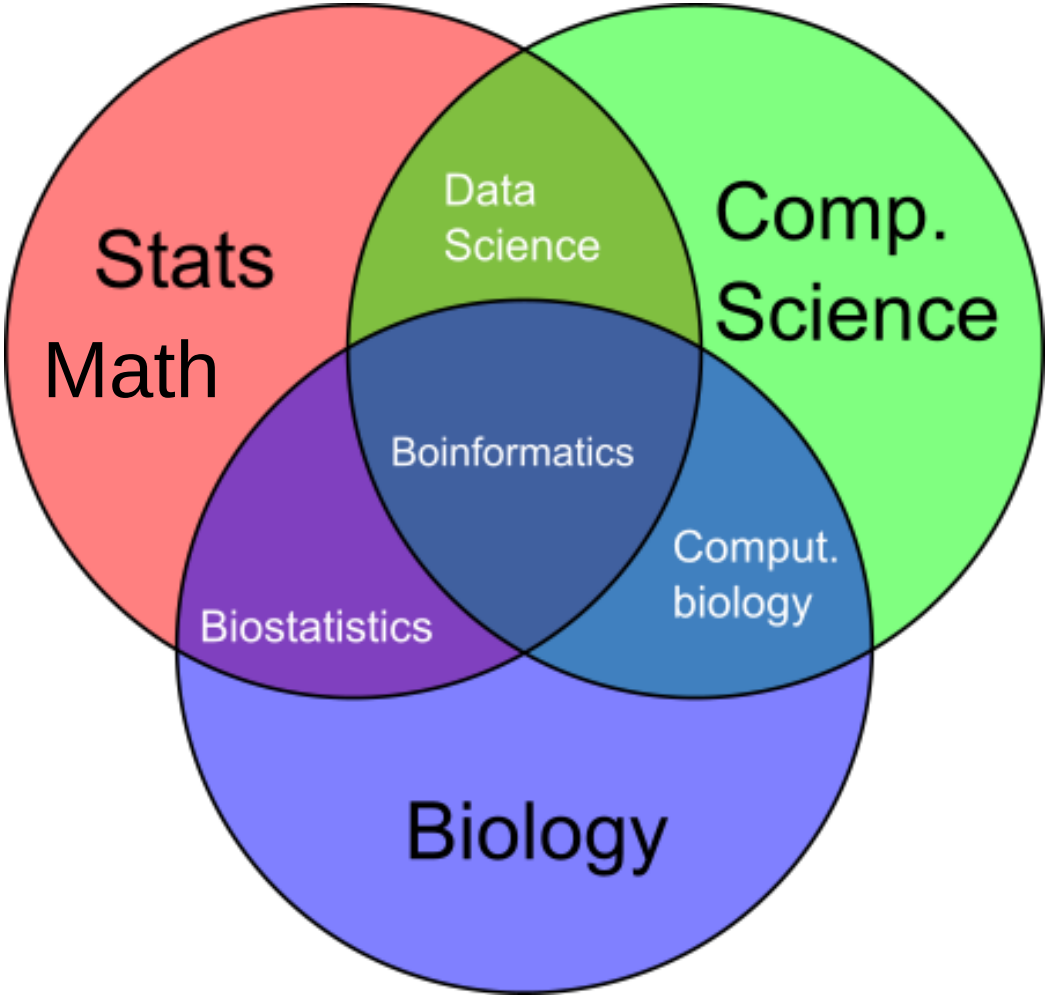
Taxonomy, Genomics and Useful of Plant-Associated Bacteria

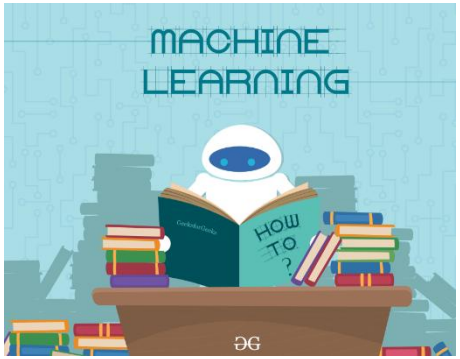
Source: Compant, S., Cassan, F., Kostić, T. *et al.* Harnessing the plant microbiome for sustainable crop production. *Nat Rev Microbiol* **23**, 9–23 (2025).

Job Description as A Bioinformatician

What's the bioinformatics?

Definition: Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines **biology**, **computer science**, information engineering, mathematics and **statistics** to analyze and interpret biological data.





Introduction: Enzyme Technology Research Team@BIOTEC

NSTDA | NECTEC a member of NSTDA | BIOTEC a member of NSTDA | MTEC a member of NSTDA | NANOTEC a member of NSTDA | ENTEC a member of NSTDA

NSTDA: driving force for Thailand's S&T

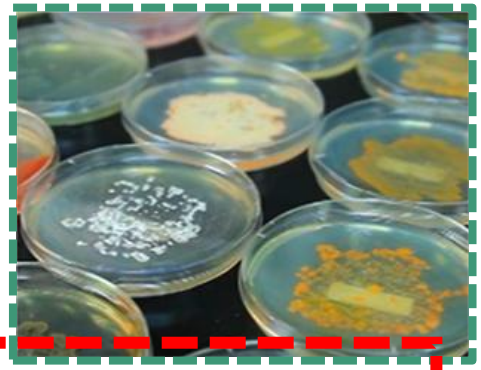
BIOTEC

- Genomic & Omics Technology
- Smart Agriculture
- Biorefinery & Biospecialties
- Food Innovation
- Tropical & Emerging Diseases

Enzymes for the green industry

Thailand Science Park: Technology incubator for industry

Introduction: Enzyme for Green Industry



1. Pure Culture

2. Environmental Metagenome

Enzyme library

Animal Feed

Enhancing nutrition/ digestibility

Pentozyme E8 Multi-enzymes for animal use

ENZBOOST

Biofuels & Biorefinery

Sugar platform conversion

ENZcas VHGF enzyme

Green processing

ENZbleach ENZEASE

Pulp/Textile
Reducing chemicals and energy

Specialty enzymes

Healthcare products

Serinsilkzyme

Introduction: Enzyme for Green Industry

TBRC
THAILAND BIORESOURCE RESEARCH CENTER

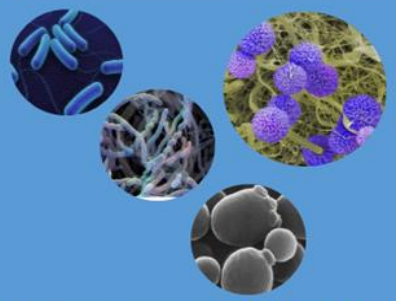
Core facilities: Microbial collection network

TBRC network

- BIOTEC-TISTR-DOA-MOPH
- ASEAN culture collection network

THAI SCIENCE BIODIVERSITY

TBRC
THAILAND BIORESOURCE RESEARCH CENTER



Microbe: > 80,000
Microcosms
Metagenomes
Biomaterials

1st in ASEAN
6th in World

BIOTEC High-throughput facility

HTP workflow for
Screening and identification
Genomic-phenotypic characⁿ
Strain improvement
Media-process optimization
Product formulation

Good strain
Non-GMM
GMM

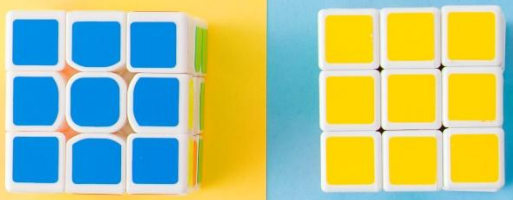
Good process
Lab-to-pilot workflow

- Enzymes
- Biochemicals
- Bioactive cpds & Bio-pharma
- Functional ingredients
- Biopolymers
- Biocontrol agent
- Specialized cells & biocatalysts
- Functional microcosms

Introduction: Enzyme for Green Industry

1% is culturable

99% is unculturable



Introduction: Enzyme for Green Industry

1. Culture Collection

2. Environmental Metagenome

Enzyme library

Animal Feed

Enhancing nutrition/ digestibility

Pentozyme E8 Multi-enzymes for animal use

ENZBOOST

Biofuels & Biorefinery

Sugar platform conversion

ENZcas VHGF enzyme

Green processing

ENZ bleach ENZEASE

Pulp/Textile
Reducing chemicals and energy

Specialty enzymes

Healthcare products

Serinsilkzyme

Introduction: Metagenomics

- is a study of genetic materials recovered directly from environmental samples
- also known as environmental genomics

Isolate

Genomics

Community

Metagenomics

Introduction: Metagenomics

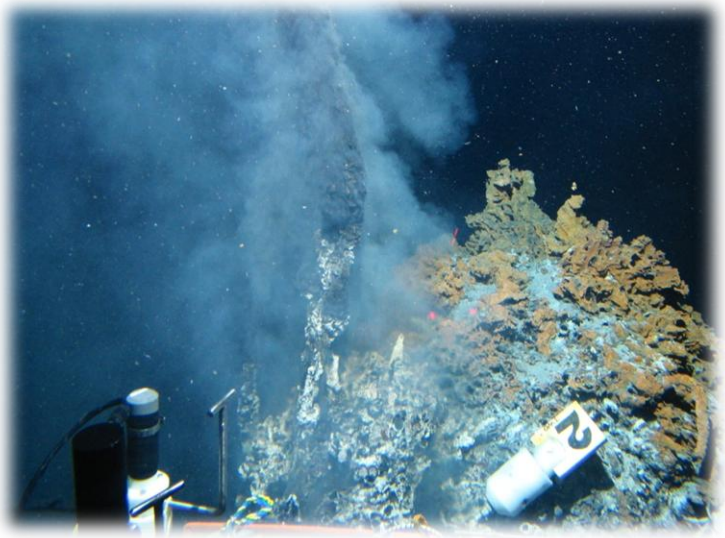
Introduction: Metagenomics

Why Metagenomics?

- Great source of novel enzymes and metabolic capabilities

Hot spring

Heat stable catalase - to reduce the cost of wastewater management (catalase breaks down H_2O_2)



Deep sea thermal vent

Found an **ultra heat stable cellulase** – that works at 109 ^{o}C !!!! (production of biofuels from plant fiber)

Introduction: Metagenomics > case study









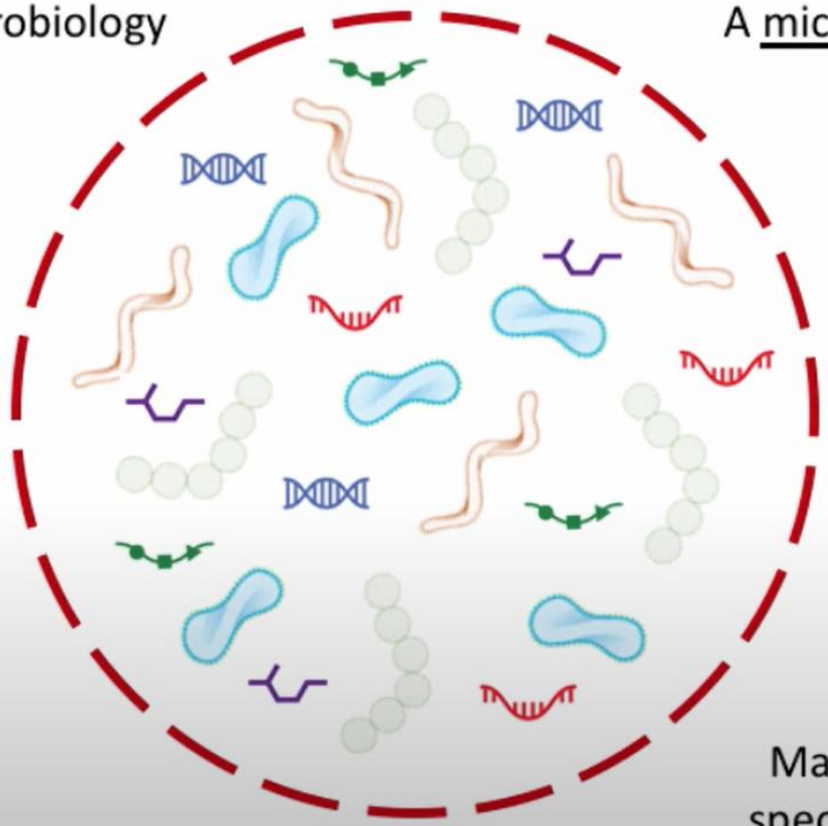


Definition

What is a microbiome?

The "micro" is from microbiology

- **Bacteria**
- Archaea
- Fungi (e.g. yeast)
- Viruses
- Protozoa



A microbiome is a community *in an environment*

Some definitions include community genes and gene products as well (RNAs, proteins, metabolites)

Cells from more than one species make a microbial community

Could be two, or three, or 100s…

Many cells of one microbial species make a monoculture

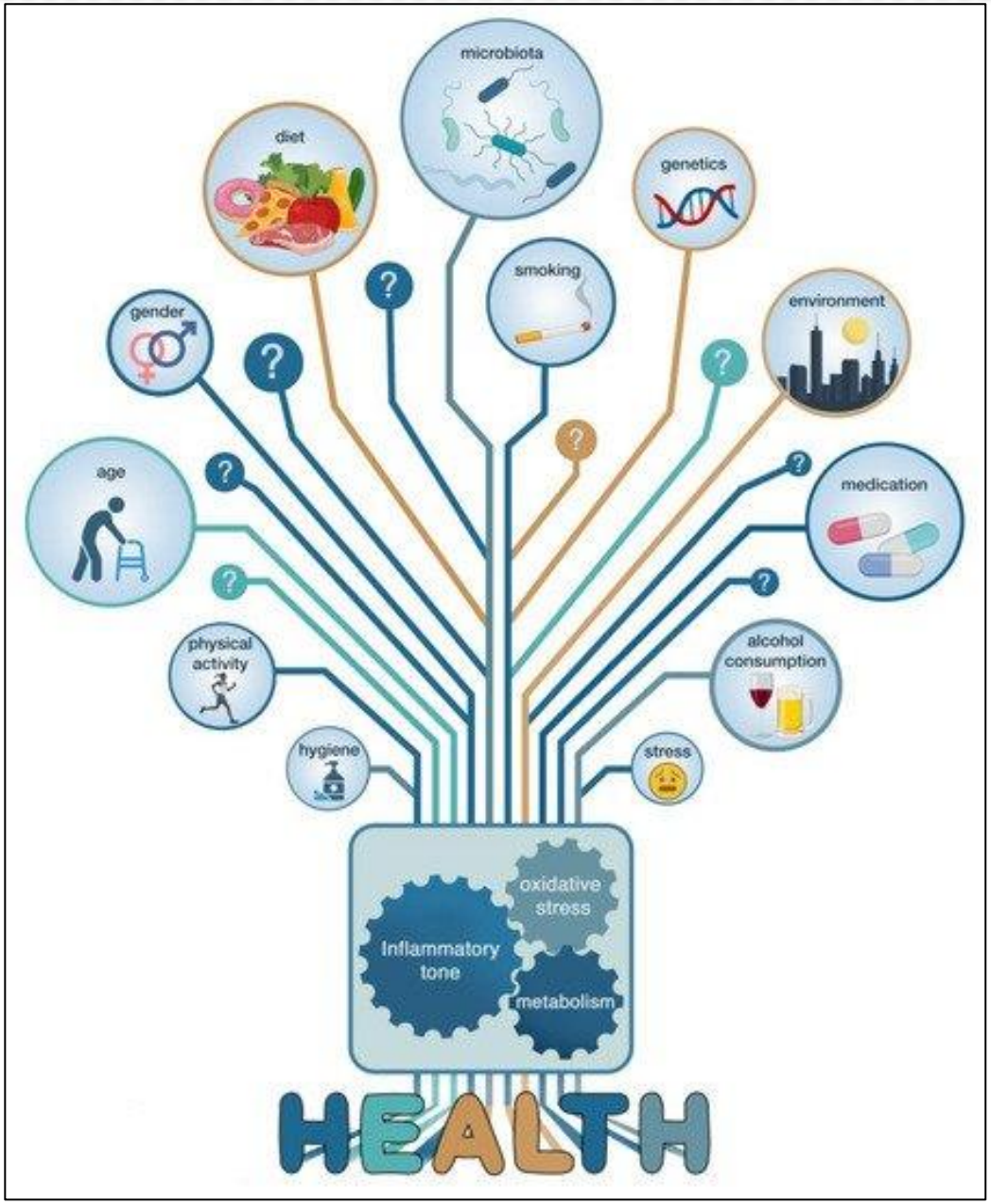
Slide adapted from Huttenhower, http://huttenhower.sph.harvard.edu/bst281

Definition

Metagenomics is the application of modern genomic techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species

Why It Matters?

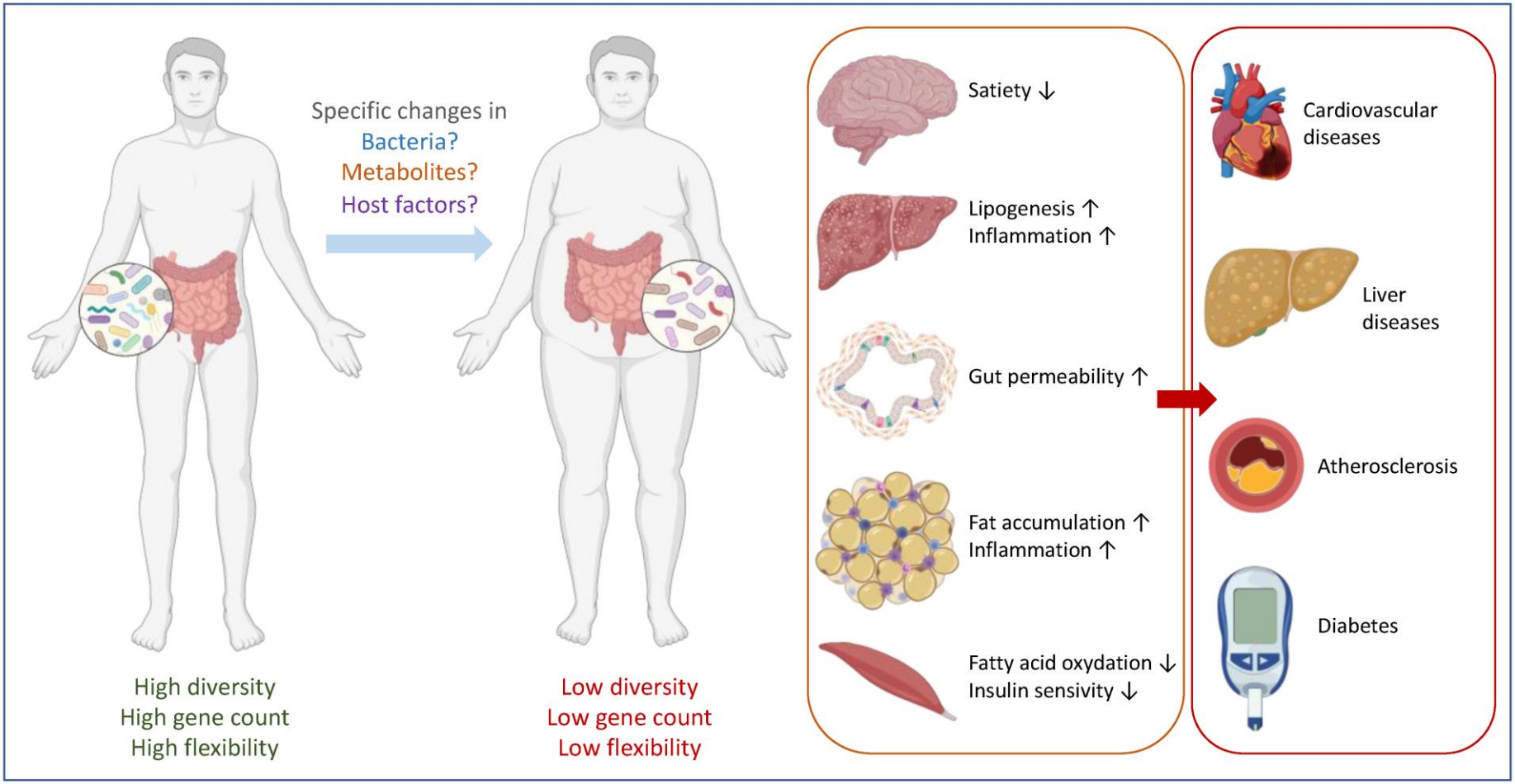
Why It Matters?



https://www.youtube.com/watch?v=RcTdTWpoWcs

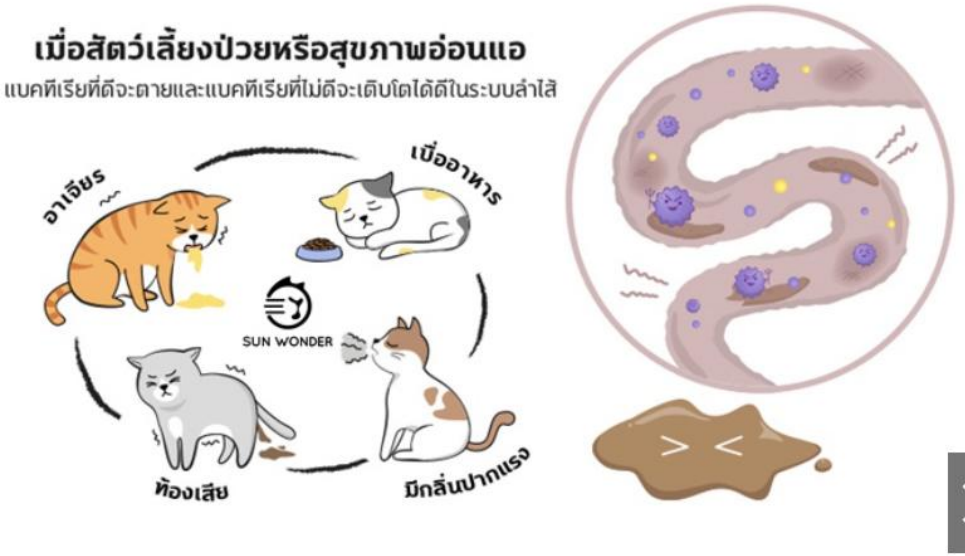
Source: Gut Microbiota and Host Metabolism: From Proof of Concept to Therapeutic Intervention

Gut Microbiome



Source: Gut Microbiota and Host Metabolism: From Proof of Concept to Therapeutic Intervention


มัดรวมสินค้า
Probiotics
ดูแลสุขภาพลำไส้ง่ายๆ
ได้ทุกวัน
ซื้อได้ที่
Bio
Bio

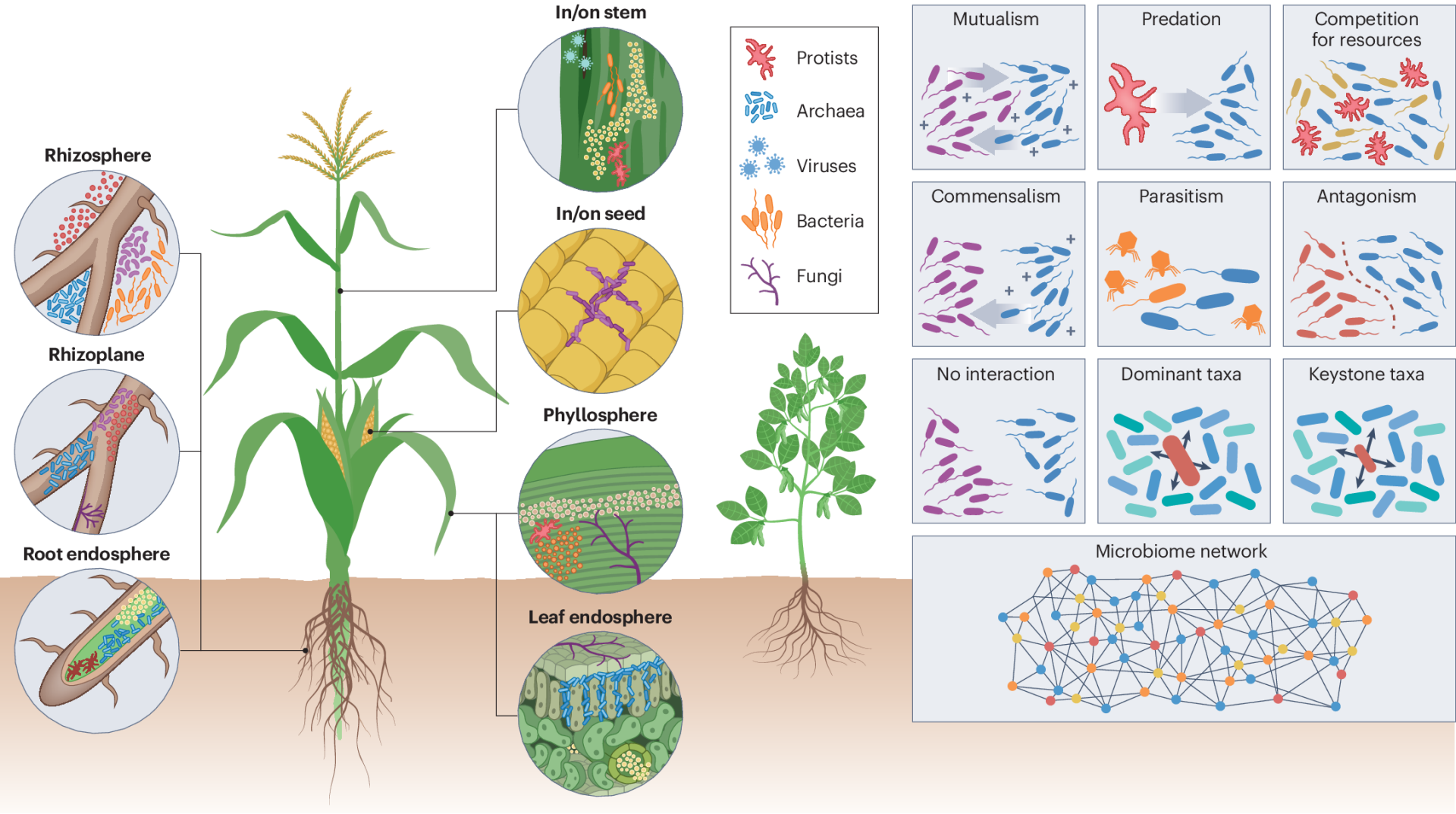


Source: Gut Microbiota and Host Metabolism: From Proof of Concept to Therapeutic Intervention

Gut Microbiome on Netflix

Gut Microbiome on Netflix

Plant Microbiome

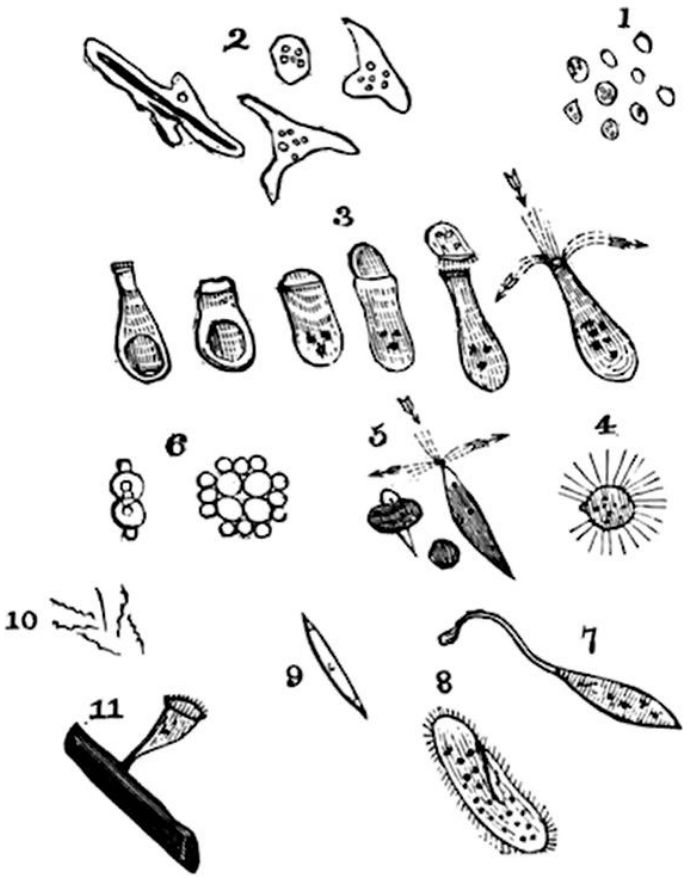


Source: Compant, S., Cassan, F., Kostić, T. *et al.* Harnessing the plant microbiome for sustainable crop production. *Nat Rev Microbiol* **23**, 9–23 (2025).

How to get data?

How to get data?

Van Leeuwenhoek's "Animalcules"



Van Leeuwenhoek, *Communication to the British Royal Society* (1676)

Antonie van Leeuwenhoek

A portrait of Antonie van Leeuwenhoek (1632–1723) by Jan Verkolje

Born	24 October 1632 Delft, Dutch Republic
Died	26 August 1723 (aged 90) Delft, Dutch Republic

What Is Genome Sequencing and Why Does It Matter?

ผมคือ DNA

ผมเป็นสารพันธุกรรม (จีโนม) ที่อยู่ในทุกเซลล์ของสิ่งมีชีวิตบนโลกนี้ ตัวผมมีหน้าที่กำหนดการทำงานต่างๆของร่างกาย และกำหนดลักษณะต่างๆของสิ่งมีชีวิต

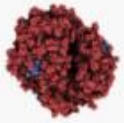
โครงสร้างของ DNA ประกอบด้วย "รหัสลับ"

Adenine
Thymine
Cytosine
Guanine

รหัสทั้ง 4 ตัวเรียกว่า "เบส"

ลำดับ DNA เหล่านี้จะสั่งการทำงานของสิ่งมีชีวิต โดยการสร้างเอนไซม์

ATG
GTC
ATC
GTA
TTG
AAT

→  →

เอนไซม์

สีผิว
เส้นผม
ส่วนสูง
สีตา
การตอบสนองต่อโรค

เมื่อแต่ละคนมี "รหัสลับ" ต่างกัน

ATTGGCATGATTG
ACGGCTAGGCCTT
AGAAATGGCTCTG

GGCATGATTGAC
GGCTAGTGCGCC
ATAGAAATGGCT

TTATAGCGAACG
TCATGATTGACG
GCTAGTGCGCCA

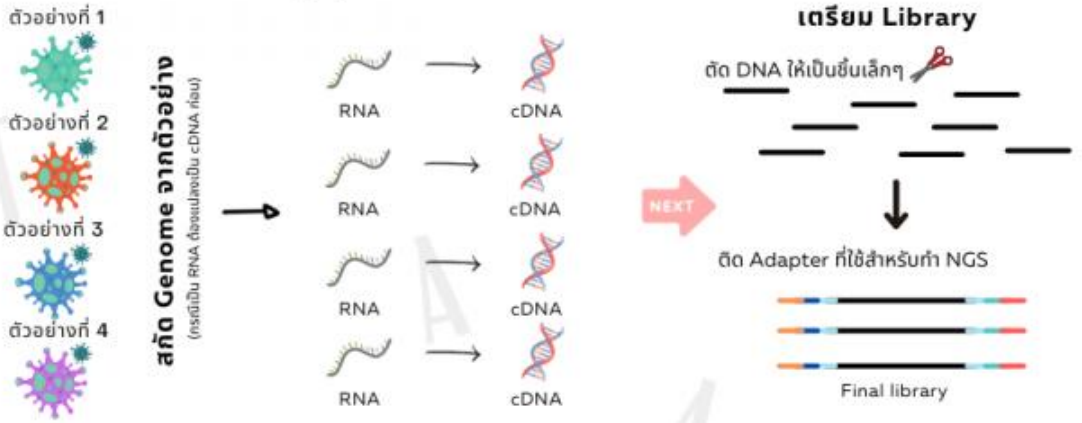
ในสิ่งมีชีวิตประเภทเดียวกัน เช่น มนุษย์
เมื่อมี**ลำดับดีเอ็นเอที่แตกต่าง**กัน ก็จะ**มีลักษณะแตกต่าง**กัน

การรู้ลำดับ DNA ในสิ่งมีชีวิต
คือ การไขรหัสลับที่ซ่อนอยู่

What Is Genome Sequencing and Why Does It Matter?

Library preparation

ขั้นตอนการเตรียม DNA ตัวอย่างให้อยู่ในรูปแบบที่พร้อมทำ NGS



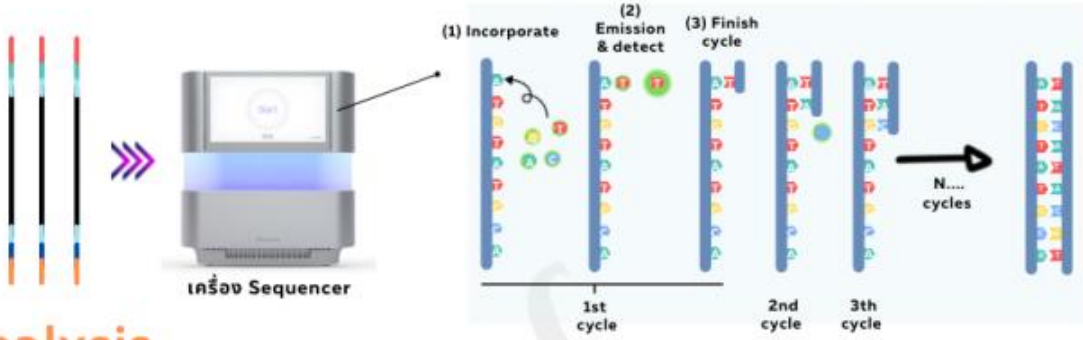
Sequencing การค้นหาลำดับเบส

วิธีไขรหัสลับ

*เป็นการยกตัวอย่างการตรวจสายพันธุ์ของเชื้อ SARS-Co-V2 ด้วย NGS

Sequencing

ขั้นตอนการทำ Sequencing เพื่อหาลำดับ DNA ด้วยวิธีการเติมเบสเข้าไปในสาย DNA จากตัวอย่าง



Analysis

การอ่านลำดับ DNA ของตัวอย่างจากเครื่อง Sequencer

การแปรผล/วิเคราะห์ผล

What Is Genome Sequencing and Why Does It Matter?

ใช้ไขรหัสลับให้ใครได้บ้าง?

Human health

การตรวจความเสี่ยงโรคมะเร็ง

โรคทางพันธุกรรม

การแพทย์แบบเฉพาะเจาะจง

Reproductive health

ตรวจคัดกรองทางพันธุกรรมก่อนการตั้งครรภ์

การตรวจคัดกรองความผิดปกติของโครโมโซมทารกในครรภ์

Microbiology

ศึกษากลุ่มประชากรของจุลชีพ

ระบุการระบาดของไวรัส

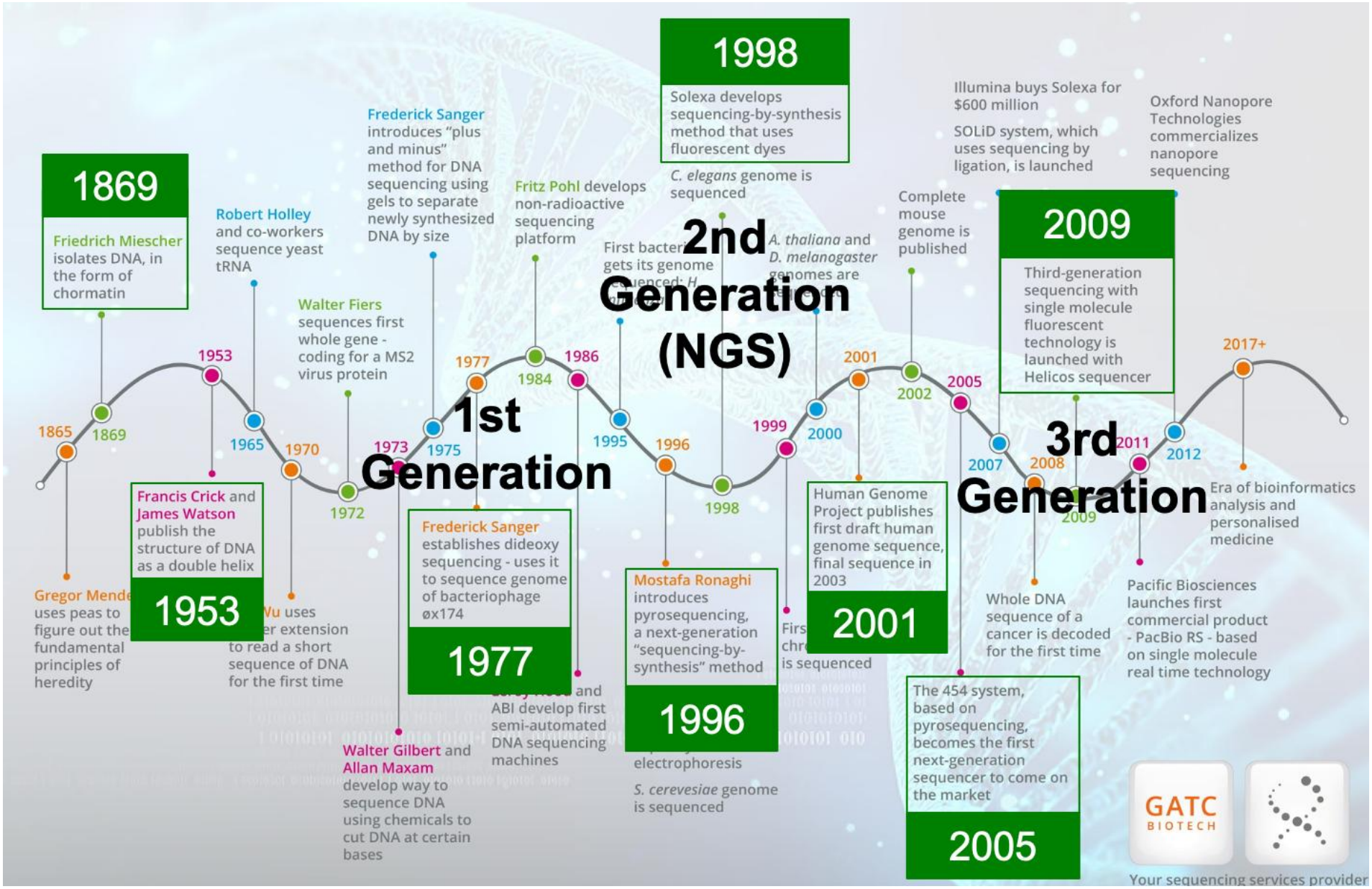
เฝ้าระวังไวรัสกลายพันธุ์

Agriculture

คัดเลือกสายพันธุ์/ตรวจโรคในสัตว์และพืชเศรษฐกิจ

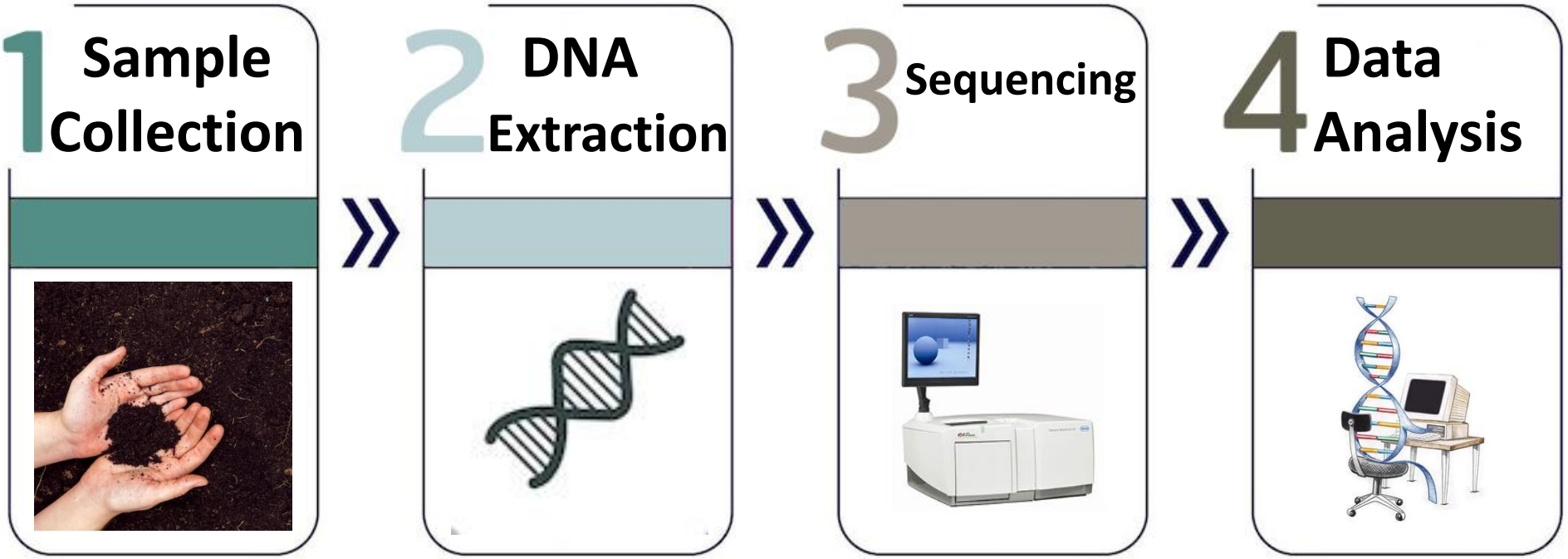
Source: https://www.bio-active.co.th/2021/07/19/next-generation-sequencing/

History of Bacterial Genome Sequencing

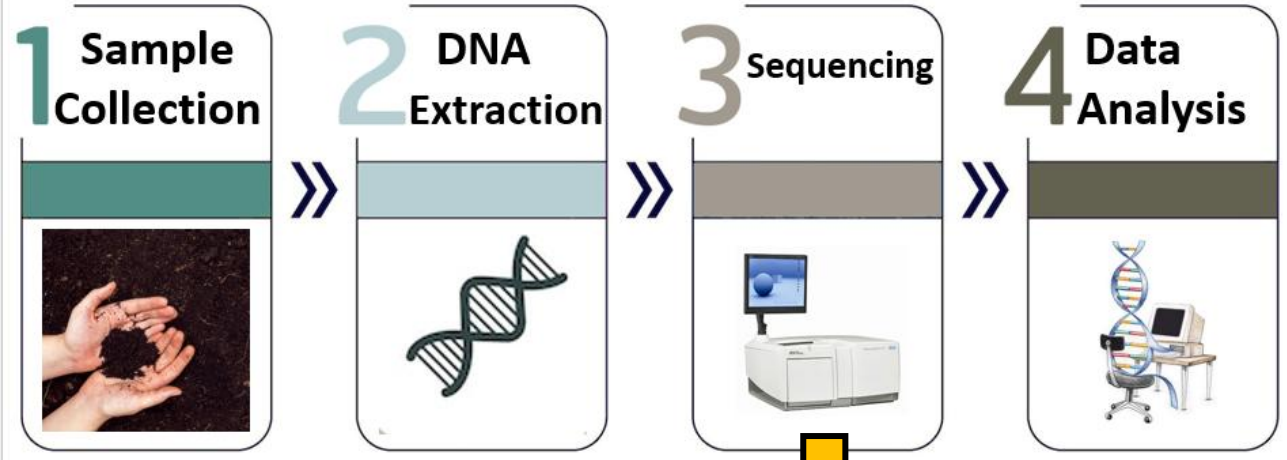


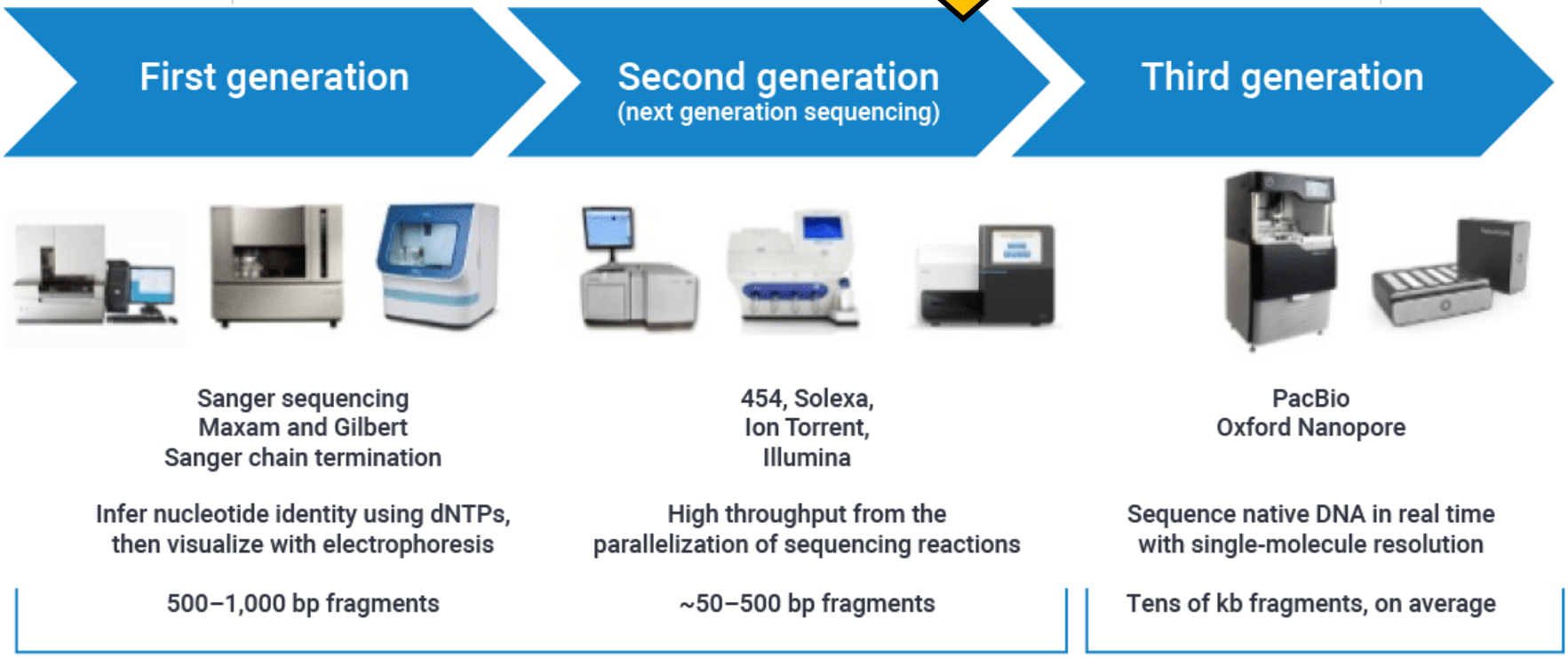
Metagenome data analysis: workflow

Simple workflow for metagenome study



Metagenome data analysis: How to get data?

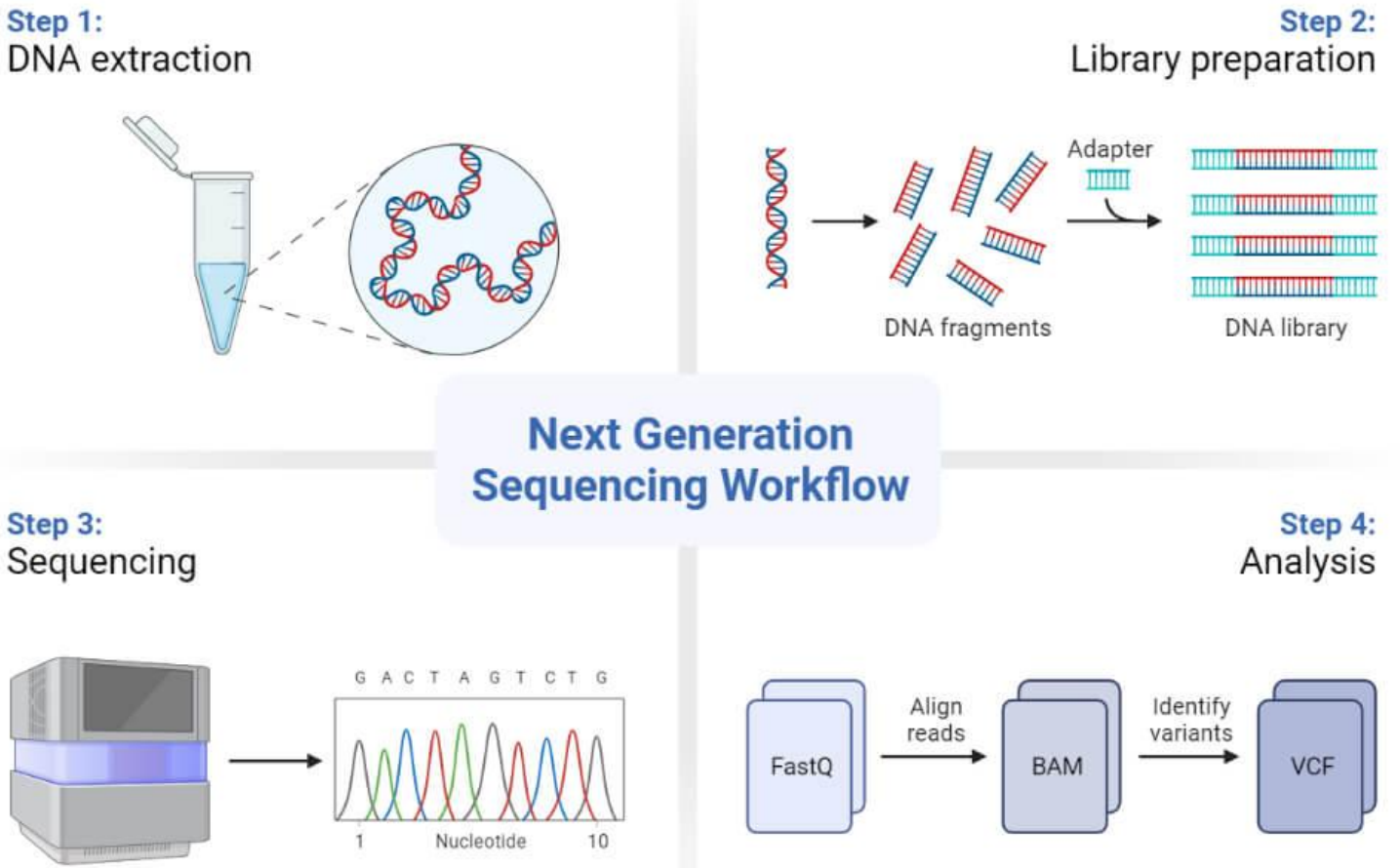




Source: https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

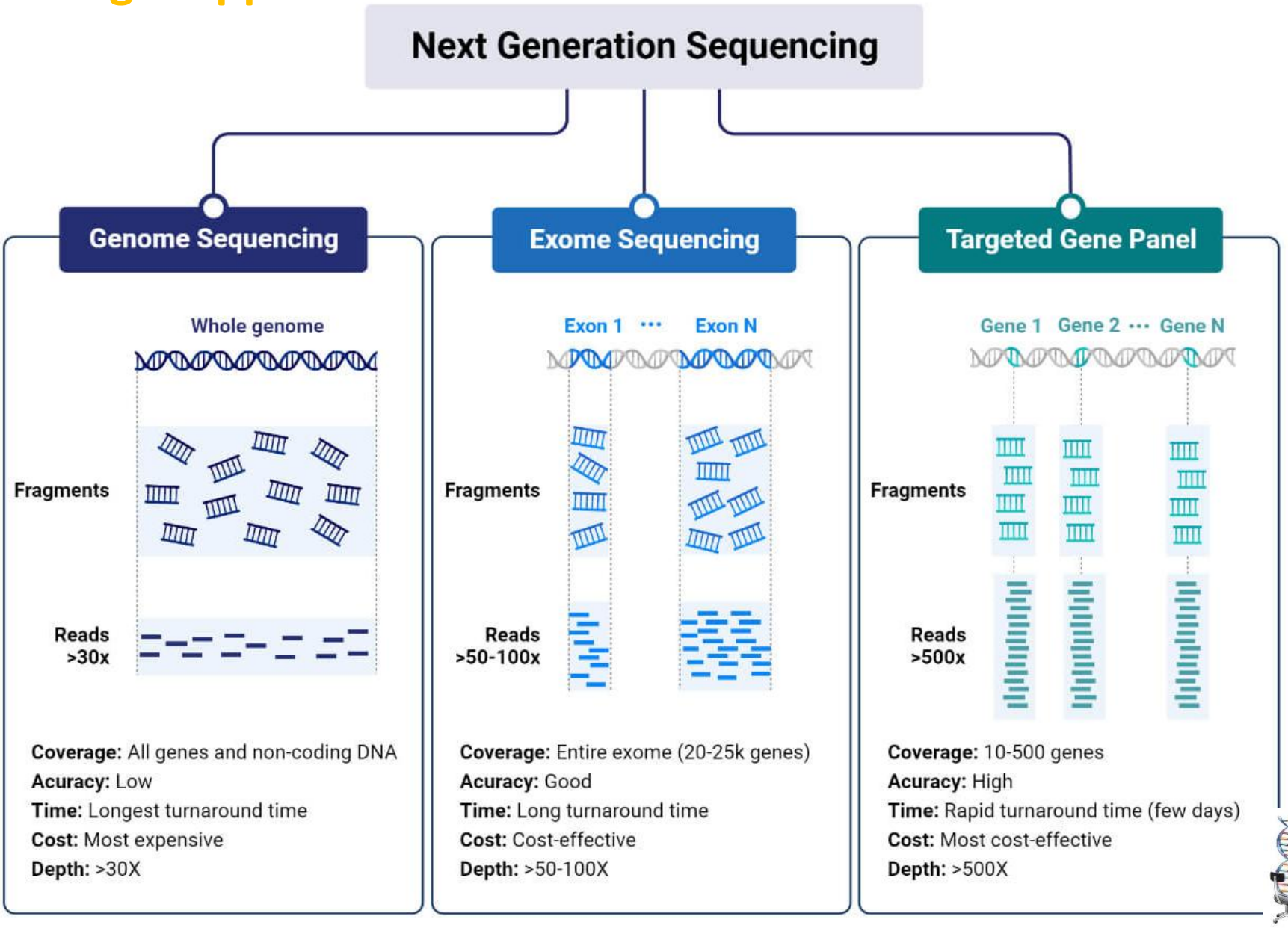
Next-Generation Sequencing (NGS)

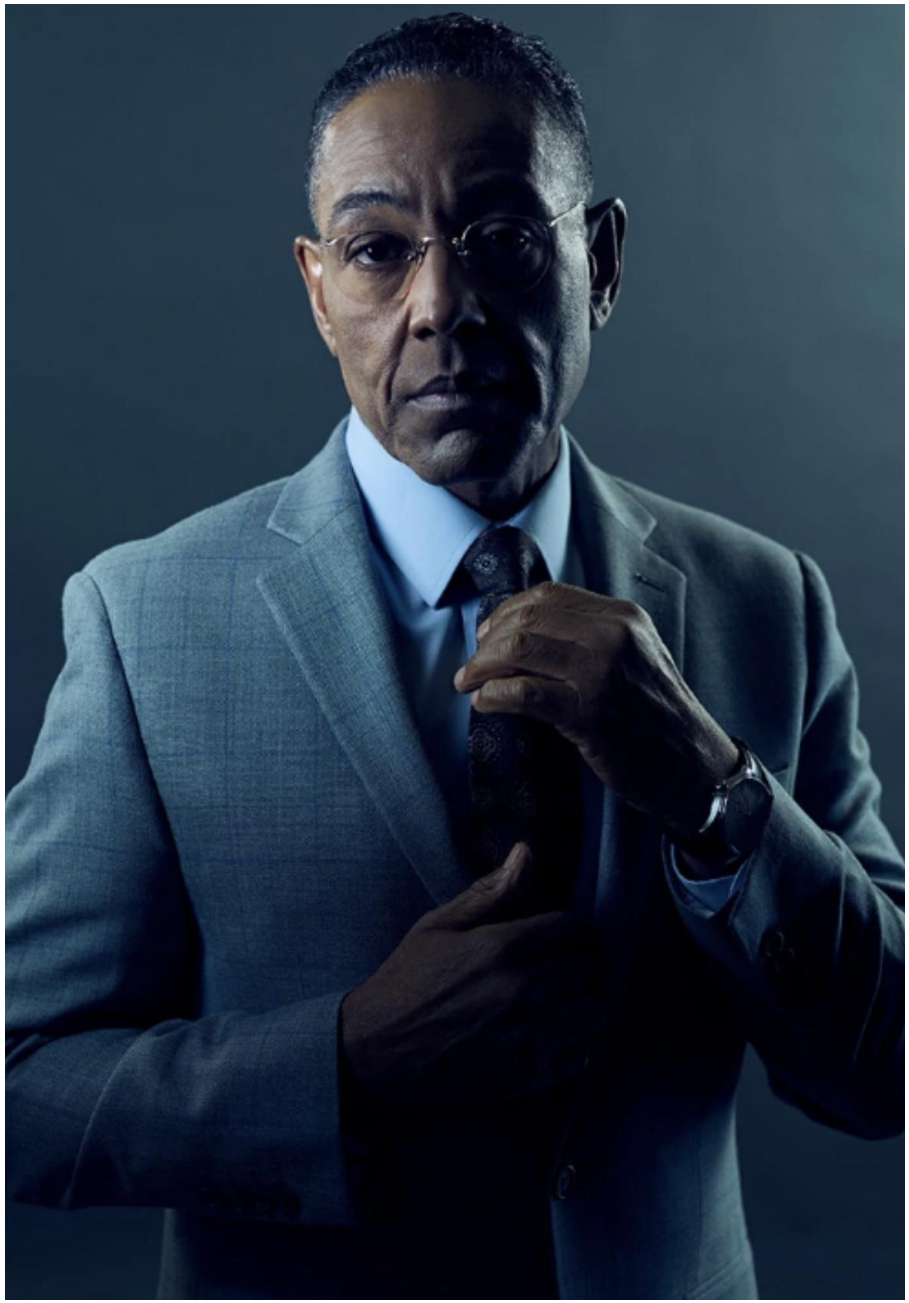
NGS workflow



Next-Generation Sequencing (NGS)

NGS target application





DNA ของเราไม่เหมือนกัน

DNA ของคุณ

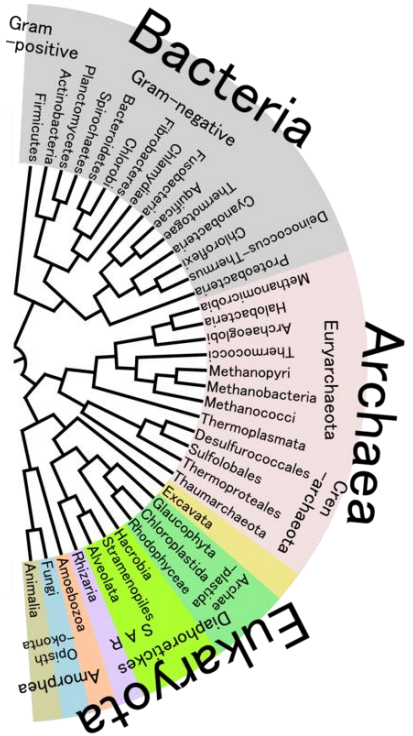
DNA ของผม

```
>CDX67397 gene=BnaA07g14370D
ATGGGATCTCCACCATCTGTATTATGTCTAATCTTATTACTCTTGTTTCATCATATTCCTTGTGTTACAA
GCCAGCGAGCACCACGTGGGTTTTGTGATACCCTAATCCACTGTGGCAACATCTCAGTTGGCTTCCCTTT
CTGGGGAGAGAATCGTCATCAAAGATGCGGTCATCCTTCTCTGAAACTTAACTGCAACAAACACTCAAAC
ACAACCACTCTTTTCATCTCAGGCTACAACTACAGTCTCCTCCATATAGACAACACGACCAACATCATTC
GACTTTTCAGACAAGATTTCTCAACTTCTTTCTGCTCCGCTTCATTCTCCTCCGCACCTCTGCCTTCTGT
ACTCTTCCAGAATTTGCCTTCTTACAAAAGCCTCACCGTCTACTACTATTGTGACCCTCGCCGCAATTTA
CTTGGGAACTTCACATGTCCATATCCAGAGAAAGGTCTTGGCTCGCTGGTTCAGTTCTCCAAATATCGTA
AGCTTTGTGAGAAGAGTTTCAAAGTCACAGTTCCTACAAGCTACGTCCCAGATGAACAAGCTTTGAGTTT
GACCCATTTGGAGAGTGTTCTCAAAAAAGGATTTGAGGTGAAGATGACGATTGGTGAAAAGTTTTGCCAA
GACTGTATTATCCCTGGTGGACACTGTGGCTTAATTTGCAGTGATACAATGTATACAACTGGTGGAACTT
ACCGCATGAATCGATCCATTGTGATCACATTTGGTTCAGTTGTAGGGATAATATTAACTGTATATGTTGC
GGTTCTTCTTGCATTTGTTTTAAGCGAAAGAAGGGCAATCAATGCTGGGAGAGATCATAACCTCGAGGCA
```

Metagenome data analysis: What do we want to know?

1

Who is in there?



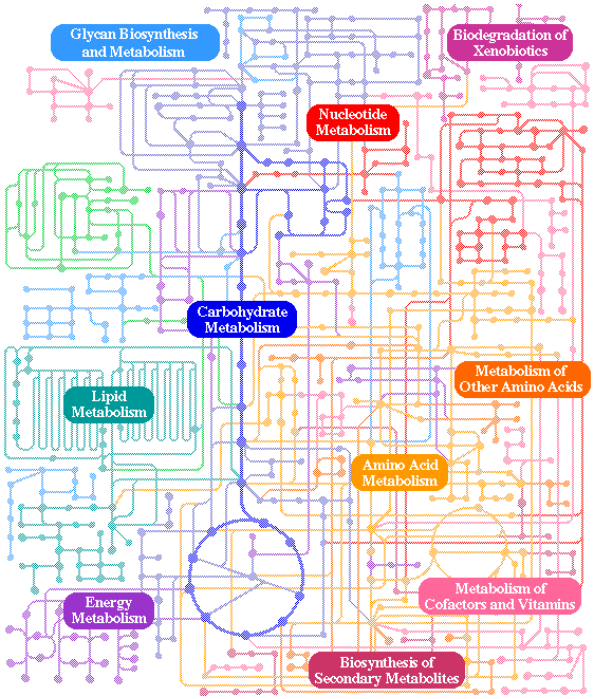
Microbial Diversity

⇩

Amplicon sequencing

2

What are they doing?



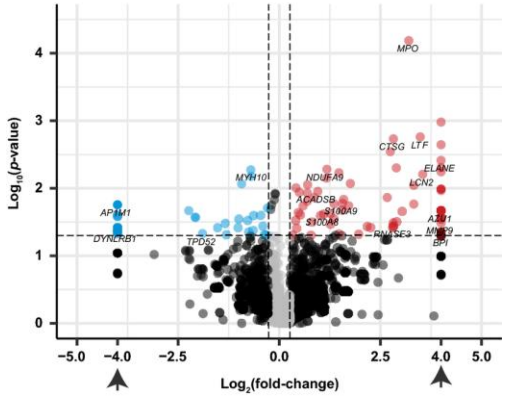
Metabolic genes

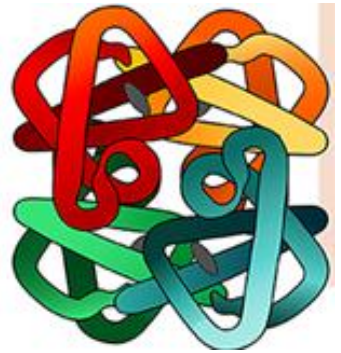
⇩

Shotgun sequencing

3

How are they doing it?

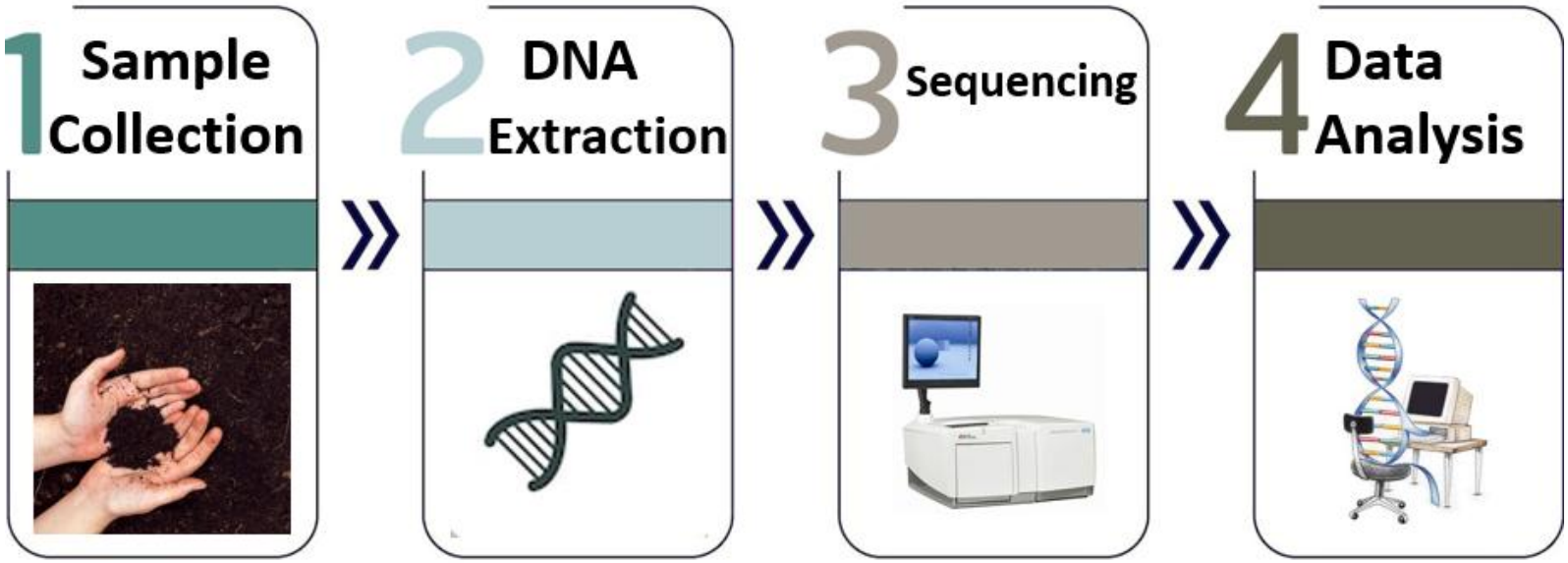




Meta-transcriptome
Meta-proteome

Types of Metagenome study: 1. Amplicon Sequencing

Evolutionary markers as a tool for microbial diversity (Amplicon sequencing or Targeted Metagenomics)

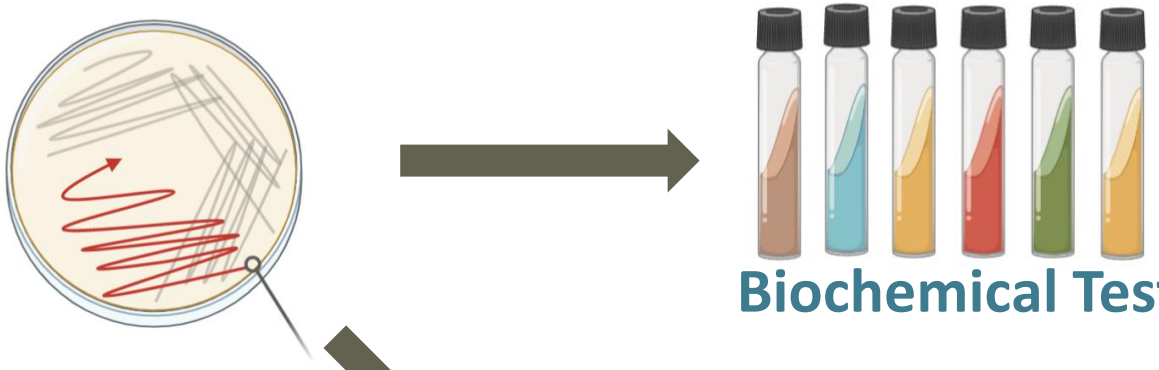


PCR amplified based on regions of interest

Source: https://www.nature.com/articles/nrmicro3330

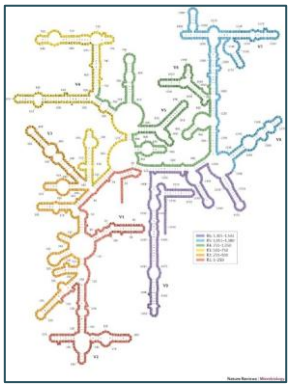
Types of Metagenome study: 1. Amplicon Sequencing

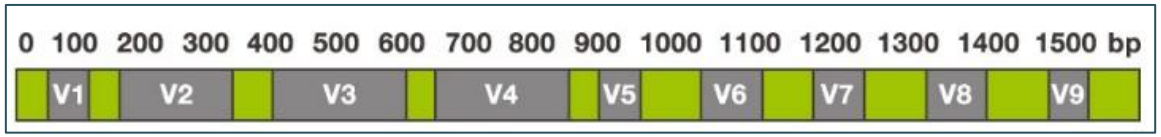
16S rRNA gene: a marker gene for Prokaryotic microorganisms



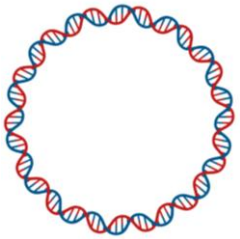
Biochemical Test

16S rRNA genes





Whole genome sequencing

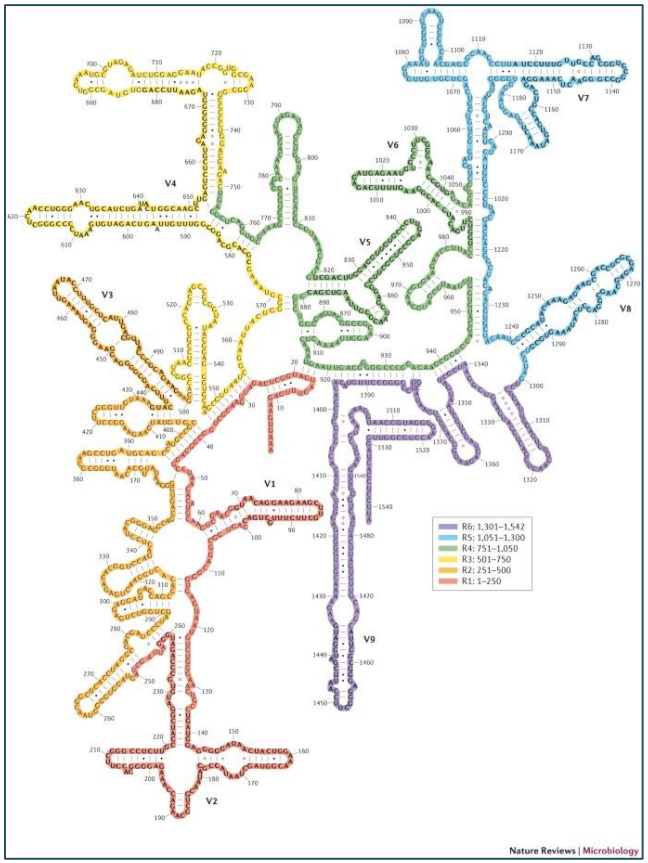


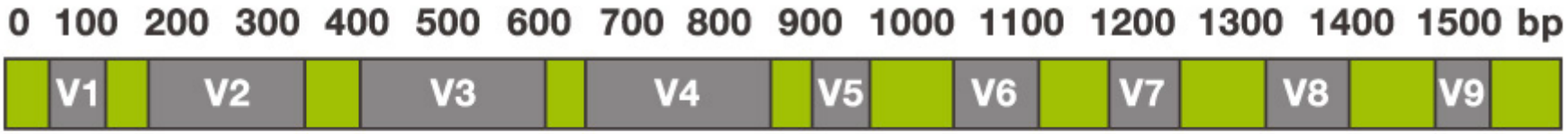
PACIFIC BIOSCIENCES®

Types of Metagenome study: 1. Amplicon Sequencing

16S rRNA gene: a marker gene for Prokaryotic microorganisms

- Universal phylogenetic marker
- Present in all bacteria and archaea
- Currently, the only widely used taxonomic marker that is sufficiently informative





CONSERVED REGIONS: unspecific applications

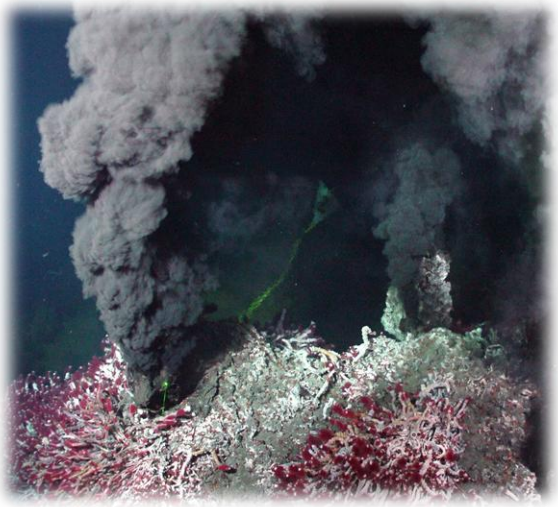
VARIABLE REGIONS: group or species-specific applications

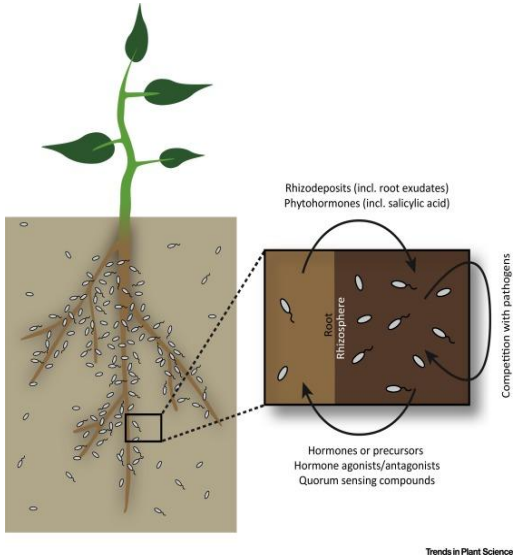
Source: https://www.nature.com/articles/nrmicro3330

Types of Metagenome study: 1. Amplicon Sequencing

16S rRNA gene: a marker gene for Prokaryotic microorganisms

- DNA isolation can be a technical challenge
  - Some microorganisms, especially thermophiles, are **difficult to lyse** under standard conditions
  - Release of very stable nucleases upon cell lysis





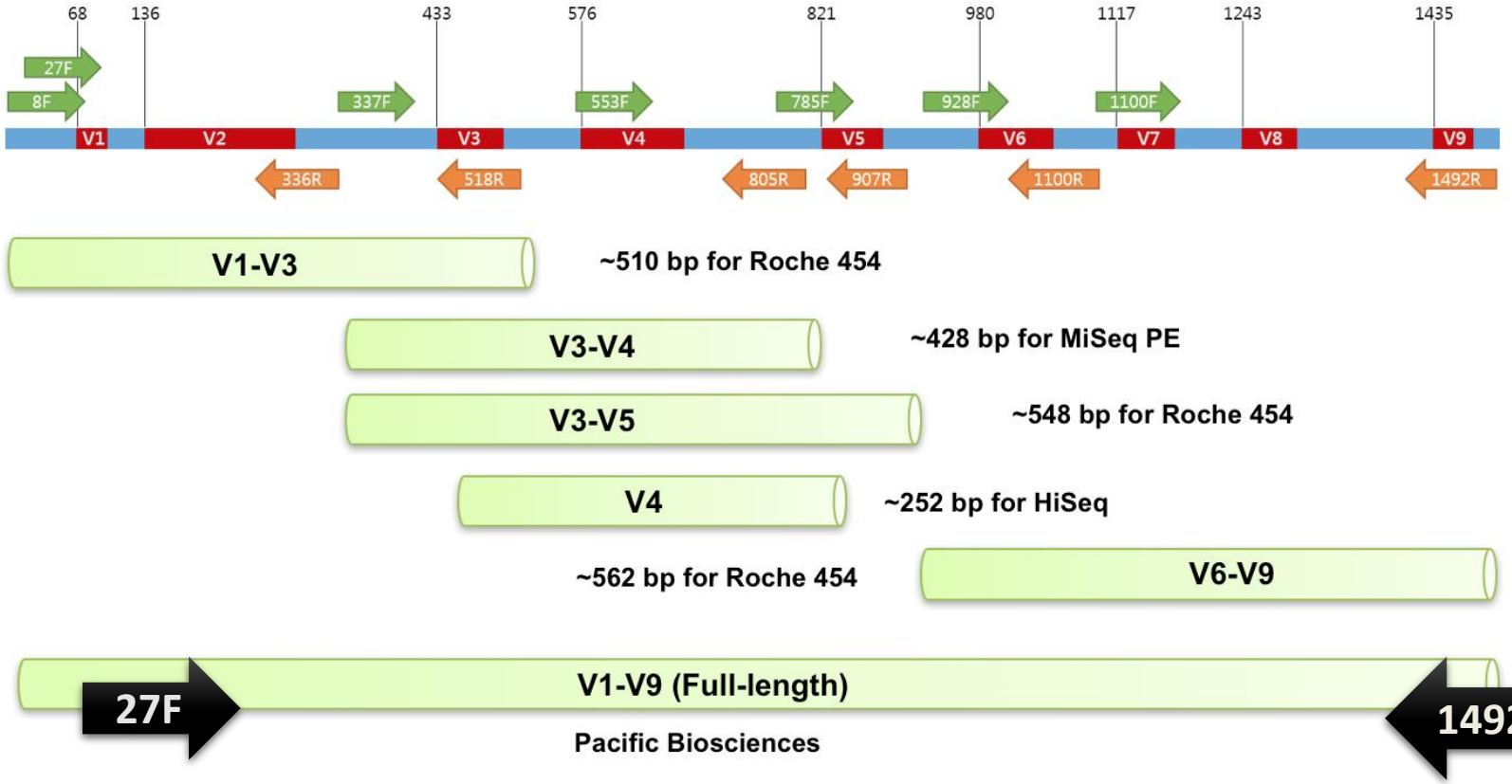
- If the target community is associated with a host (e.g. plants or invertebrates), then either fractionation or selective lysis might be suitable to ensure that **minimal host DNA is obtained**.

Source: https://www.nature.com/articles/nrmicro3330

Types of Metagenome study: 1. Amplicon Sequencing

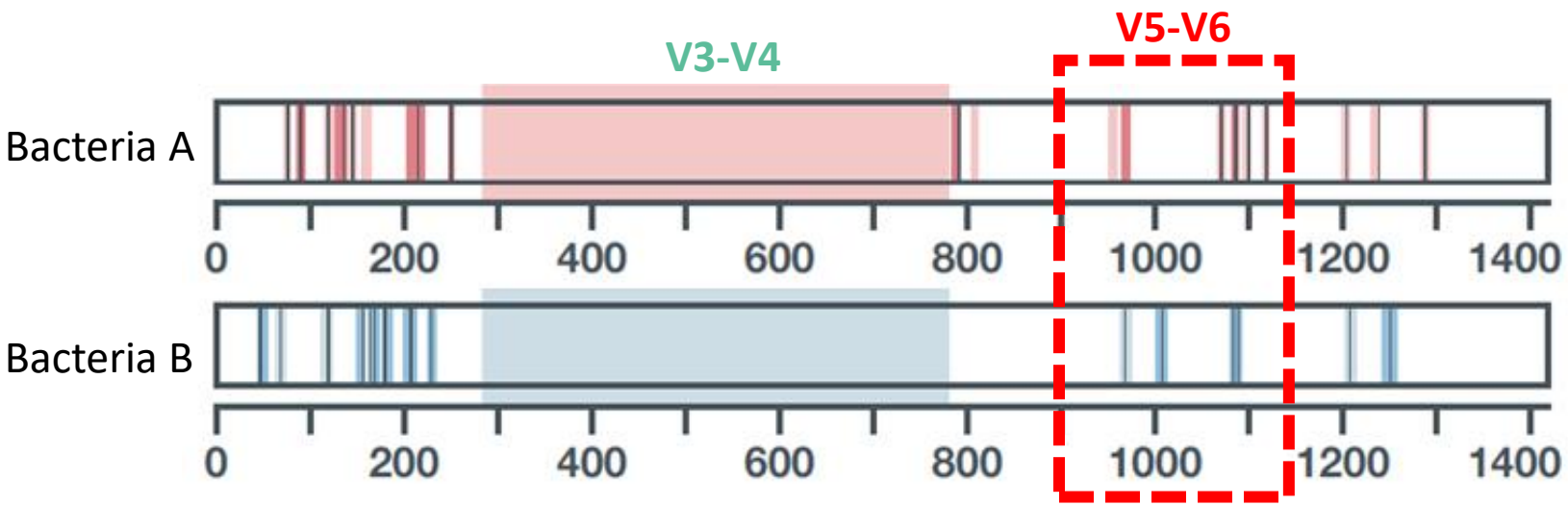
16S rRNA gene: a marker gene for Prokaryotic microorganisms

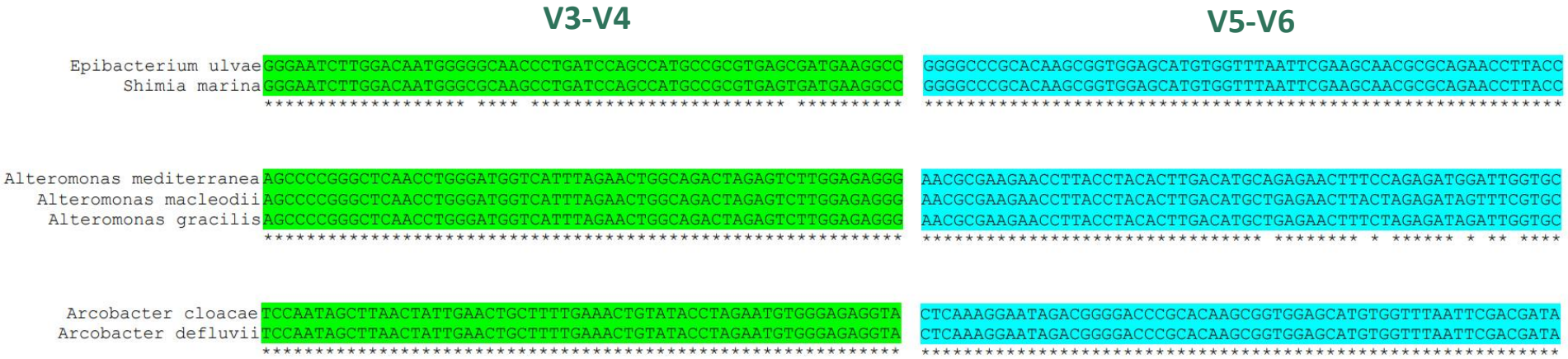
Use the **lowest number of cycles** to prevent a bias introduced by an over-amplification



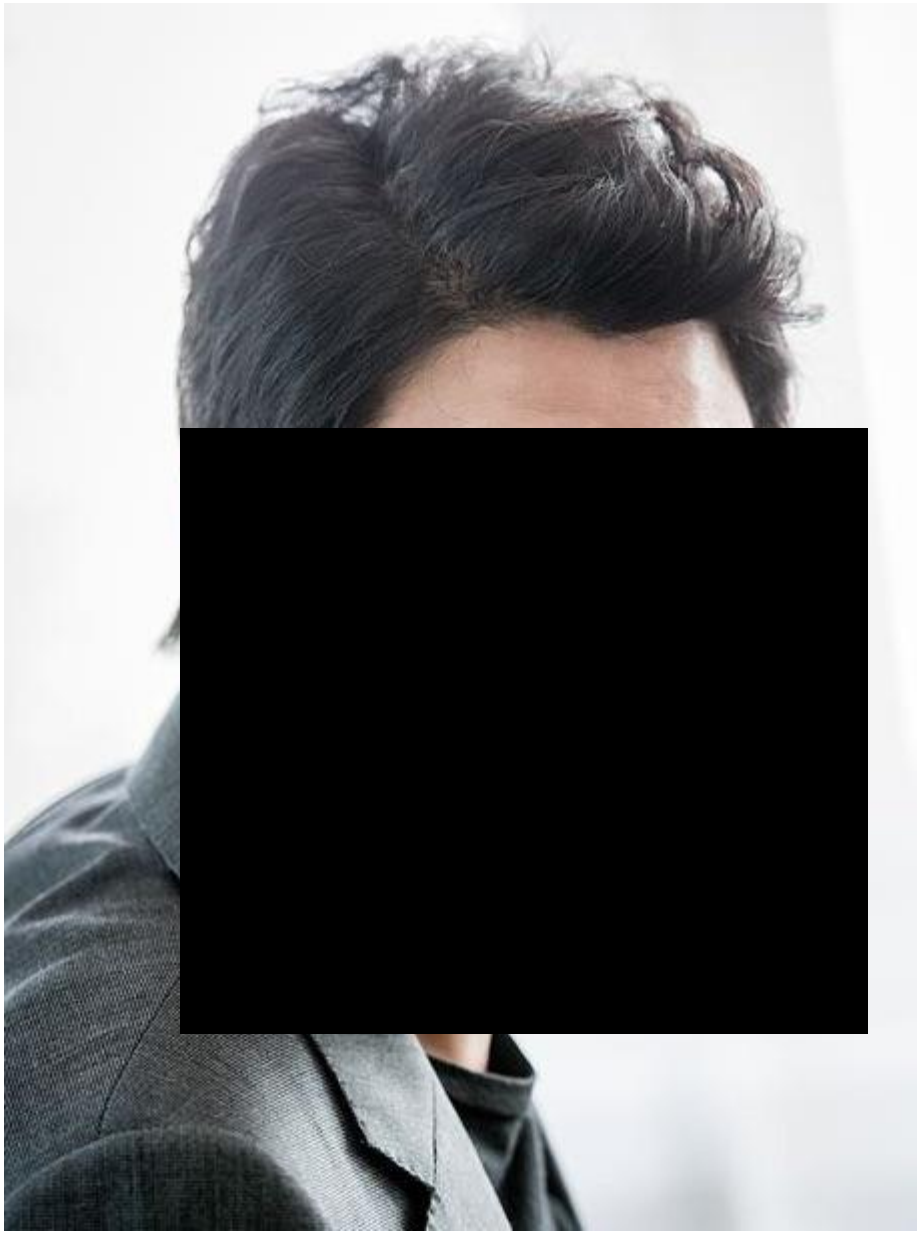
Types of Metagenome study: 1. Amplicon Sequencing

Comparison between short-read and long-read sequencing





Source: Pootakham, W., Mhuantong, W., Yoocha, T. *et al.* High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. *Sci Rep* **7**, 2774 (2017). https://doi.org/10.1038/s41598-017-03139-4

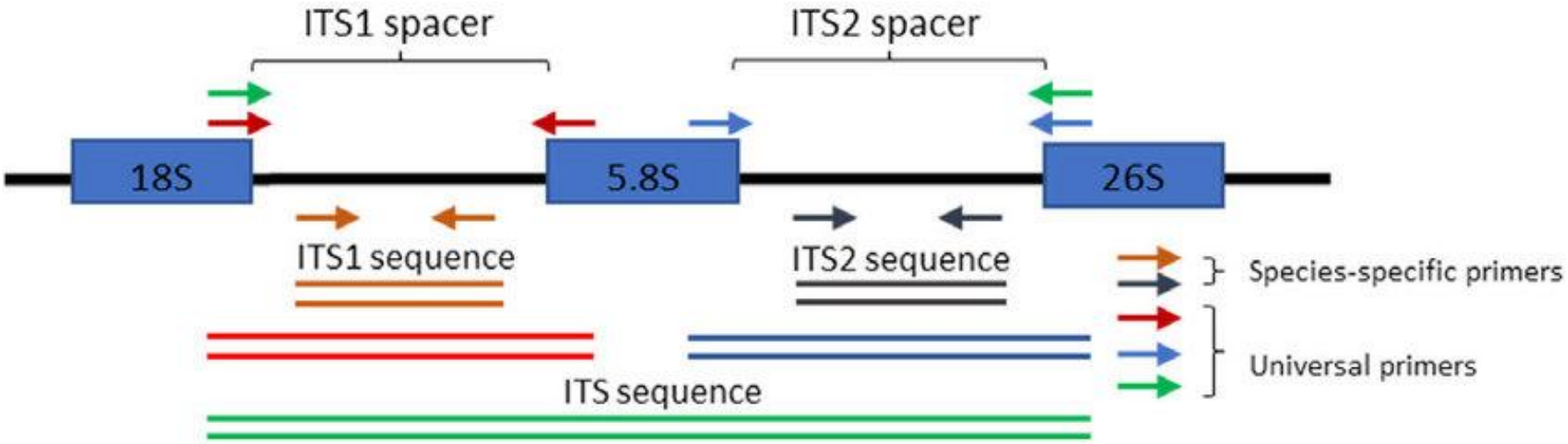


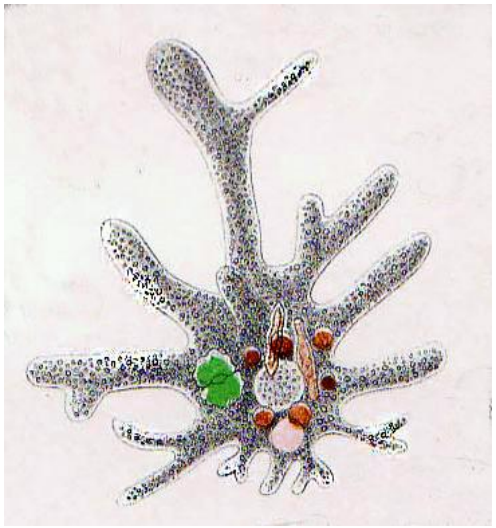
Song Kang-Ho

Song Kang

Types of Metagenome study: 1. Amplicon Sequencing

18S rRNA gene/ Internal transcribed spacers (ITS): marker genes for Eukaryote





Types of Metagenome study: 1. Amplicon Sequencing

Specific metabolic gene

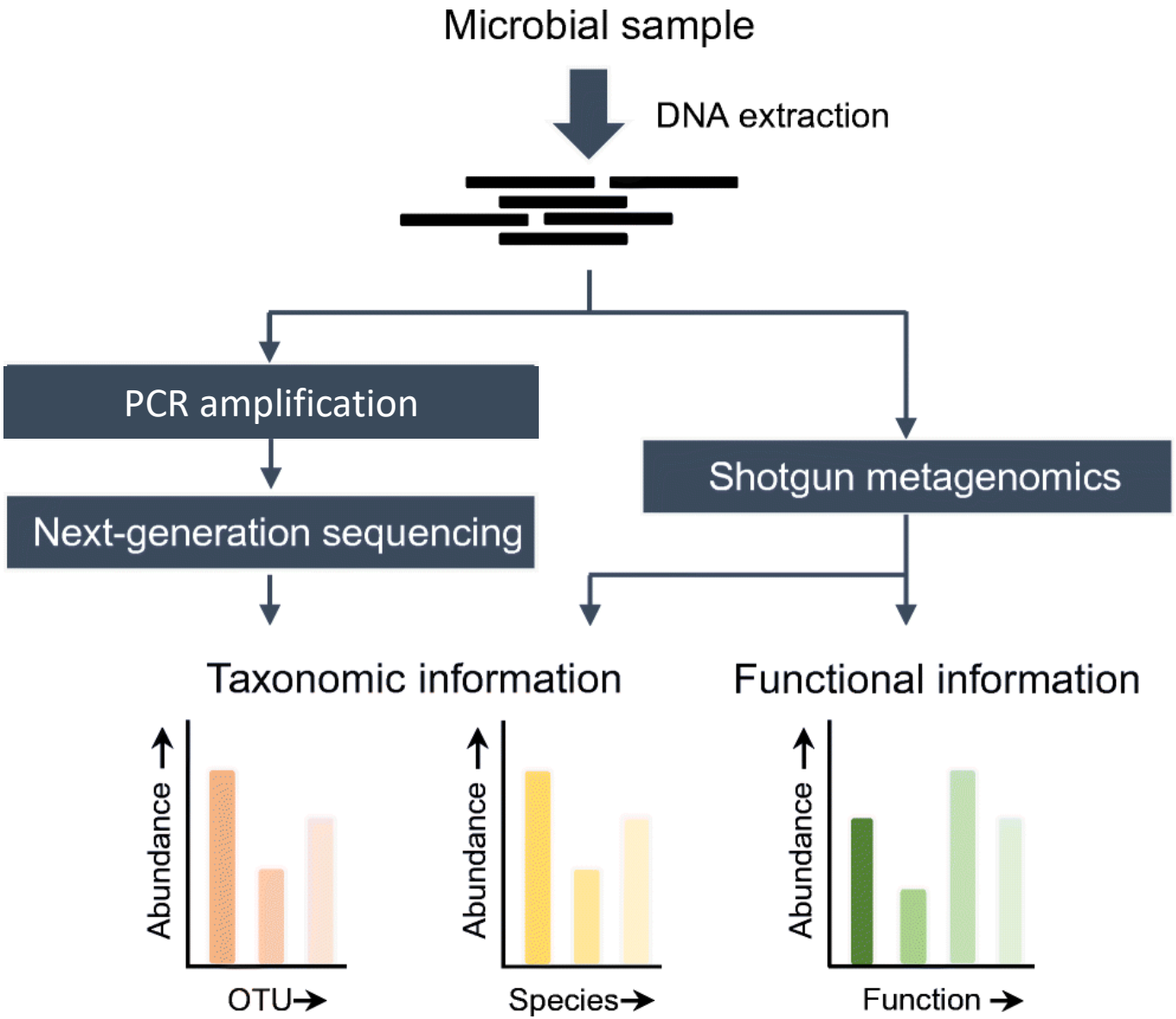
*mcr*A gene: methanogenic archaea
AOA: Ammonia-oxidizing archaea
AOB: Ammonia-oxidizing bacteria

Marker genes

กลุ่มสิ่งมีชีวิต	Marker Gene	ตัวอย่าง Primers
Bacteria & Archaea	16S rRNA	515F/806R (V4), 341F/785R (V3–V4)
Fungi	ITS (ITS1, ITS2)	ITS1/ITS2, ITS3/ITS4
Protists & Microbial Eukaryotes	18S rRNA	Euk1391f/EukBr (V9), TAReuk454FWD1/TAReukREV3 (V4)
Plants & Algae	rbcL, matK, trnL	rbcL-aF/rbcL-R, matK-3F/matK-1R
Metazoa (Animals)	COI (COX1)	LCO1490/HCO2198, mlCOIintF/jgHCO2198

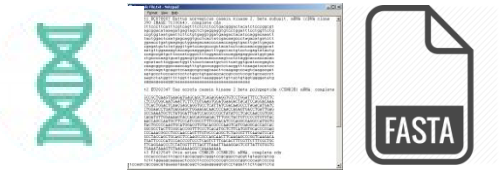
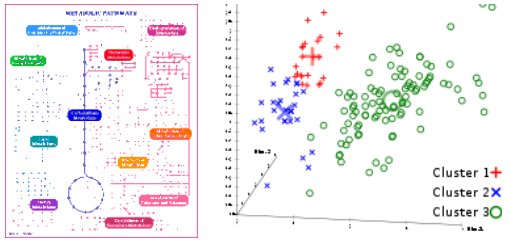
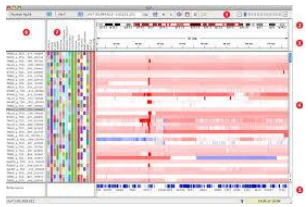
Source: https://www.nature.com/articles/nrmicro3330

Types of Metagenome study: 2. Shotgun Sequencing

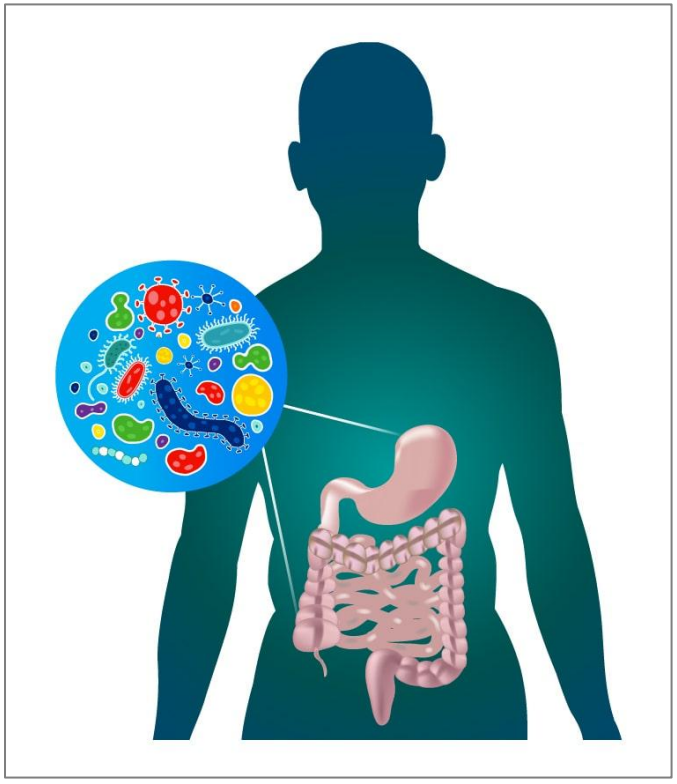


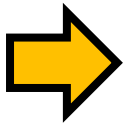
Metagenome Data Analysis

Data Analysis

	Cooking	Metagenome Analysis
Material		
Equipment/ Tools		
Process		
Output	  	  

Microbiome in computer





```
>b24aadee9a22976aeff52f5c5b8dd79f
TGGGGAATATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGAGTGATGAAGGCCTTAGGGTTGTAAAGCTCTTTTGTCCGGGAAGATAATGACT
GTACCGGAAGAATAAGCCCCGGCTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGGGGCTAGCGTTGCTCGGAATCACTGGGCGTAAAGGGCGCGTAGGC
GGACTCTTAAGTCGGGGGTGAAAGCCCAGGGCTCAACCCTGGAATTGCCTTCGATACTGAGAGTCTTGAGTTCGGAAGAGGTTGGTGGAACTGCGAGTGTAG
AGGTGAAATTCGTAGATATTCGCAAGAACACCAGTGGCGAAGGCGGCCAACTGGTCCGATACTGACGCTGAGGCGCGAAAGCGTGGGGAGCAAACA
>5b678dbbecb30309b675d6fd541a3fcb
TAAGGAATATTGCTCAATGGGGGAAACCCTGAAGCAGCAACGCCGCGTGAGGGATGAAGGGCTTTTGCTTGTAAACCTCTGTAGTGAGGGAAAAAATTTCCA
CTCTGTGGAATATGATGGTACCTCAAAAGTAAGCCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGGGCAAACGTTGTCCGGATTTACTGGGT
GTAAAGGGTGCTCAGGCGGTTTTGTAAGTCAGAAGTGAAATCCTAGGGCTTAACTCTGGAACTGCTTTTGATACTGCAAGGCTTGAATGTGGAAGAGGAGGA
TGGAATTTCTGGTGTAGCAGTGAAATGCGTAGATATCAGAAAGAACACCAGTGGCGAAGGCGGTCCTCTGGTCCACTATTGACGCTAAAGCACGAAAGCGTG
GGGAGCAAACA
>d894419825f4c6996e3c1b1738611875
TTTCGAATCATTCACAATGGGCGAAAGCCTGATGGTGCGACGCCGCGTGAGGGATGAAGGTCTTCGGATTGTAAACCTCTGTCACTGGGGAGAAAACGCTTT
GAGTTAATAGTTCAAAGCCTGATTTAACCCGGAGAGGAAGCAGTGGCTAACTCTGTGCCAGCAGCCGCGGTAATACAGAGACTGCAAGCGTTATTCGGATTC
ACTGGGCGTAAAGGGTGCGCAGGCGGCCGTGTGTGTTAAATGTGAAATCCCGGGGCTCAACCCCGGAACTGCGTCTAAAACTACATGGCTAGAGTACTGGAG
AGGGTAACGGAATTCACGGTGTAGCAGTGAAATGCGTAGATATCGTGAGGAACACCAGAGGCGAAGGCGGTTACCTGGACAGTTACTGACGCTCAGGCACGA
AAGCGTGGGGAGCAAAAG
>a241f37c6efa1ecc89756f5821edebab
TAGGGAATATTGGGCAATGGGCGGAAGCCTGACCCAGCCATGCCGCGTGCAGGATGAAGGCGTTATGCGTTGTAAACTGCTTTTTTATAGGAAGAAATAGTT
TTTGCGAAAACAGTTGACGGTACTATAAGAATAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCGAGCGTTGTCCGGATTTATTGG
GTTTAAAGGGTGCGTAGGCGGTATATTAAGTCAGTGGTGAAATACGGTCGCTCAACGATTGCCGTGCCATTGATACTGATAAACTTGAGTATGATAGAGGTA
TTTGGAATGGATAGTGTAGCGGTGAAATGCATAGATATTATCCAGAACGCCAATAGCGAAGGCAGAGTACTATGTCATTACTGACGCTGATGCACGAAAGTG
TGGGGATCAAACA
>f28b860f07e180c90d1d37cd32ee74f9
TGGGGAATATTGGACAATGGGGGCAACCCTGATCCAGCCATGCCGCGTGAGTGAAGAAGGCCTTCGGGTTGTAAAGCTCTTTCGACGGGGACGATGATGACG
GTACCCGTATAAGAAGCCCCGGCAAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGGGGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCGCGTAGGC
GGTGATCTTTGTCAGAGGTGAAAGCCTGGAGCTCAACTCCAGAATTGCCTTTGAAACGGGATCGCTAGAGTCCGAGAGAGGATGGCGGAATTCCTAGTGTAG
AGGTGAAATTCGTAGATATTAGGAAGAACACCGGTGGCGAAGGCGGCCATCTGGCTCGGTACTGACGCTCAGGCGCGAAAGCGTGGGGAGCAAACA
>44178992ec59d6e207140428b6125bac
TGGGGAATCTTAGACAATGGGCGCAAGCCTGATCTAGCCATGCCGCGTGAGCGATGAAGGCCTTAGGGTTGTAAAGCTCTTTCAGTGGGGAAGATAATGACT
GTACCCACAGAAGAAGCCCCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGGGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCACGTAGGC
GGACTGGAAAGTCAGAGGTGAAATCCCAGGGCTCAACCTTGGAACTGCCTTTGAAACTCCCGGTCTTGAGGTCGAGAGAGGTGAGTGGAATTCCGAGTGTAG
AGGTGAAATTCGTAGATATTCGGAGGAACACCAGTGGCGAAGGCGGCTCACTGGCTCGATACTGACGCTGAGGTGCGAAAGCGTGGGGAGCAAACA
>50ddd0e6e0d692fc17e352a3d4cebdeb
TAACGAATCTTCCACAATGTACGAAAGTATGATGGAGCGACGCCGCGTGTAGGATGAAGTCCTTCGGGATGTAAACTACTGTCAGGGGTAAGAAAGTTCTGA
TCTACCCCAGAGGAAGAGACGGCTAACTCTGTGCCAGCAGCCGCGGTAATACAGAGGTCTCGAGCGTTAGGCGGAATCACTGGGCTTAAAGCGTGTGTAGGT
GGATGGGTAAGTACCTTGTGAAATCCCACGGCTCAACCGTGGAACTGCTTGGTATACTGCCCGTCTTGAGTCACTCAGGGGCGAGCGGAACAGATGGTGGAG
CGGTGAAATGCGTAGATATCATCTGGAACGCCAATGGTGAAGACAGCTCGCTGGGGGTGTACTGACACTGAGACACGAAAGCCAGGGGAGCAAACG
>851302af3052393b183e7677ac9a2e4c
TGGGGAATATTGGGCAATGGGCGAAAGCCTGACCCAGCAACGCCGCGTGAAGGAAGAAGGTTTTCGGATTGTAAACTTCTGTCCTCAGGGAAGAAACAAATG
ACGGTACCTGAGGAGGAAGCCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGGGCGAGCGTTGTCCGGAATGACTGGGCGTAAAGGGCGCGTA
GACGGCTGTCTAAGTCTGGAGTGAAAGTCCTGCTTTCAAGGTGGGAATTGCTTTGGAGACTGGATGGCTTGAGTGCGGAAGAGGTAAGTGGAATTCCCAGTG
```

Sequencing File Format

DNA Extraction

Sequencing

SEQUENCING DATA

FASTQ

FASTA + QUAL

Demultiplex

FASTA

Analysis

BiOM Biological Observation Matrix

Assembly

Mapping

REFERENCE

Annotation

GFF

BAM

SNP Calling

VCF

Sequencing File Format

DNA Extraction

Sequencing

SEQUENCING DATA

FASTQ

FASTA + QUAL

Demultiplex

FASTA

OTU picking

Biological Observation Matrix

Assembly

Mapping

REFERENCE

Annotation

GFF

BAM

SNP Calling

VCF

Sequencing File Format

DNA Extraction

Sequencing

SEQUENCING DATA

FASTQ

FASTA + QUAL

FASTA

FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes.

```
>AY919613.1 Rhodothermus marinus heme-copper oxygen reductase (coxA3) gene, partial cds
ATGGCAGAGCACATTGCAGCCTCGACGCCGGTCGTGGCCGATCCGACCCCGTTTACGCTGCCCGAAACGC
AGCGTCGCCTGCTGCGCTGGACGCTGTACGTGGGCTATGCGGCGCTGACGGCCGGCATCTTTCATGGACT
GGCACAGGCCCTCTCGTACGCGGGTATCGACATTCTGGGCTATTTCCCGGCCCTGCGCAGCTACTACCAG
GGCCTGACGGCCCACGGCGTGGCCAACGCCATCATCTTTACGTTCTCTTTTGCAAACGCCTTTCTGCCGC
TCATGGTGGCGCGGGCGCTCTCGCGTCGCCTGGACGAGCGGCTGCTTTGGGCCAGCTTCGGCACGCTGGT
GCTGGGCAACCTGCTCGTGATCTACGCGGTGGTCACGAACAAGGCCAGCGTGCTTTACACCTCGTACGCA
CCGCTGCAGGCGCACTGGACCTATTACGTGGGGCTGGTTTTCGTCGTGATCAGCACCTGGCTGGCCCTGC
TGAACATGTTGCTGACCTGGCGGGGCTGGAAGCGGGAAAATCCAGGTGTGCGCATGCCGCTGCTGGCCCA
CATCTCGATCGTCTCCTACGTGATGTGGTTTCTGGCCTCGCTGCCCATTGCCGTCGAGTTTCTGTTTTTC
CTGATTCCCTGGTCGTTCGGCTGGGTGGAGCGGACCGATCCGCTGCTGACGCGTACGCTCTTCTGGTTCA
CCGGCCACGCCATCGTGTACGCCTGGCTGTTGCCGGCCTACGTGTCGTGGTACGCGCTGGTGCCGCGCCA
GGCGGGCGGTAAAATTGTGAGCGACTCGCTCACGCGGCTGGTGTTCATTCTCTTTCTGCTGCTGTCGATC
CCGACGGGCTTTCACCACCAGTACACGGACCCCGGCATTCACGAAGGGTTCAAGTTCGTCCACGCCATCC
TGACCTTCGGCGTGTTCTTCCCGAGCCTGATCACGGCCTTCAGCGTGATGGCCTCGCTGGAGATGGGCGG
CCGGGCGCACGGCGGCCGGGGCCTGCTGGGCTGGATTCCGAAGCTCCCCTGGGGCGATCCGTCGCTCTCG
GCCCAGTTGCTGGCTATGATCACGTTCGTTTTTGGCGGCATCACGGGCCTGATCAACGCCTCGTTCACGA
```

1

2

> ID/ACC description

Sequencing File Format

FASTQ

FASTQ format is a text-based format for storing both a nucleotide sequence and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity.

```
1 @m170714_190721_42258_c101041032550000001823267712031612_s1_X0/404/ccs 1 22
2 TGGAAACAGCTATGACCATGTACGGTTACCTTGTTACGACTTCACCCCAGTCATGAATCACTCCGTGGTAATCGTCCTCCCGAGGGTTA
3 +
4 KKK4KKKCKKCJKKKIKKIKIKKKKEKKFJ<KKAKKKJJHKKKG8KKKFJFF@883JHKKKJKKKIJKKKKEJK;5KC25GKIGKK:KK
```

Row 1: Starts with a '@' character and is followed by a sequence identifier and an optional description.

Row 2 : The raw sequence letters.

Row 3: Starts with a '+' character and is *optionally* followed by the same sequence identifier.

Row 4: Encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

Sequence Quality Control

DNA Extraction

Sequencing

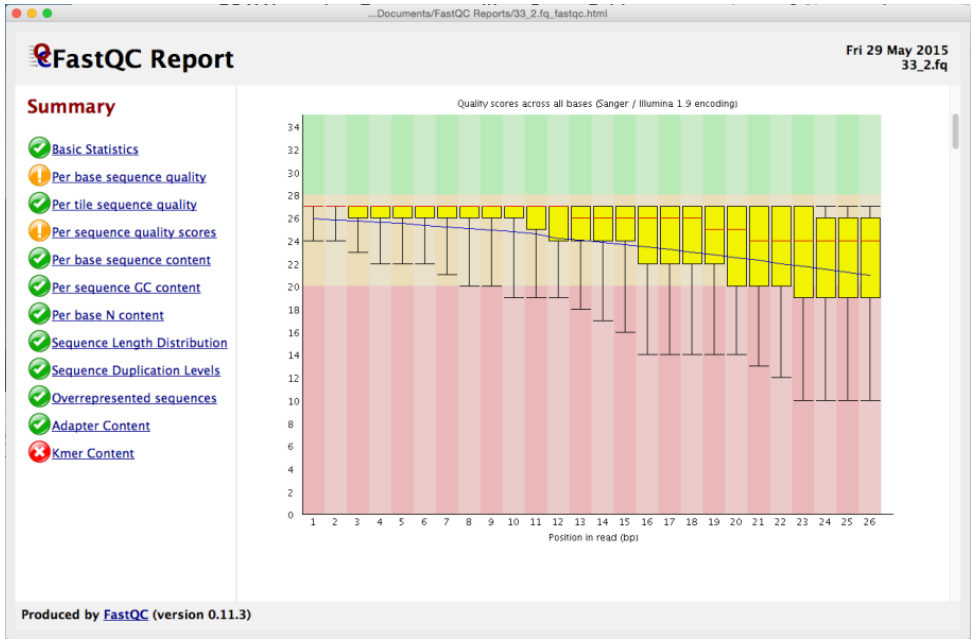
SEQUENCING DATA

READS+QUALITY

Why Quality Control and Preprocess raw reads?

- **Is the quality of my sequence data, OK?**
- Remove sequencing errors
- Reduce PCR amplification bias
- Screen out vectors/adaptors/primers

FASTQC

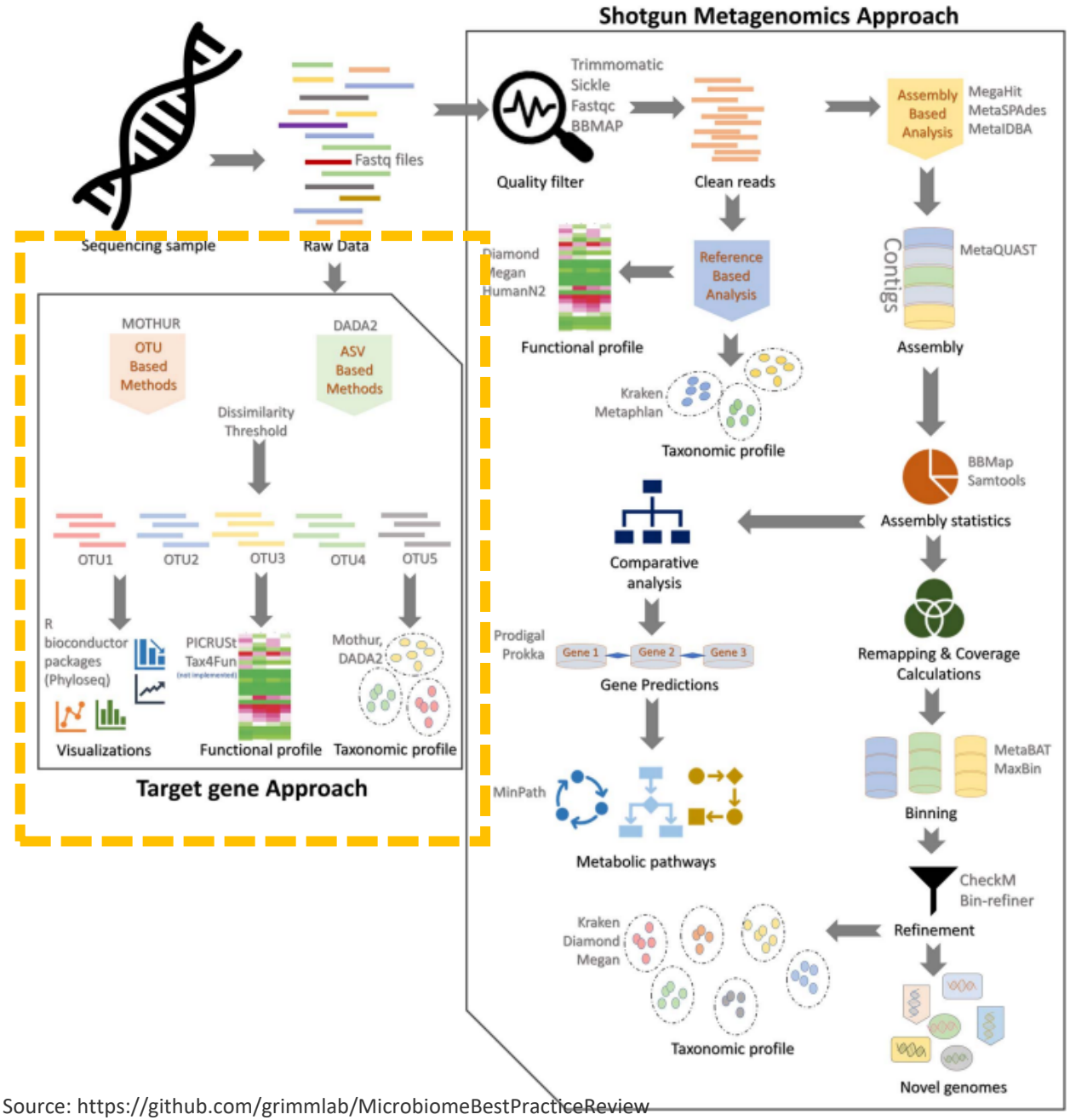


- **FASTQC** is a tool for evaluating the quality of raw sequence data.

- It provides a modular set of analyses which you can use to explore your data.

Source: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Metagenome Data Analysis: Amplicon vs. Shotgun



Protocol for the acquisition and analysis of targeted amplicon and shotgun metagenomics data from sequencing to functional annotation.

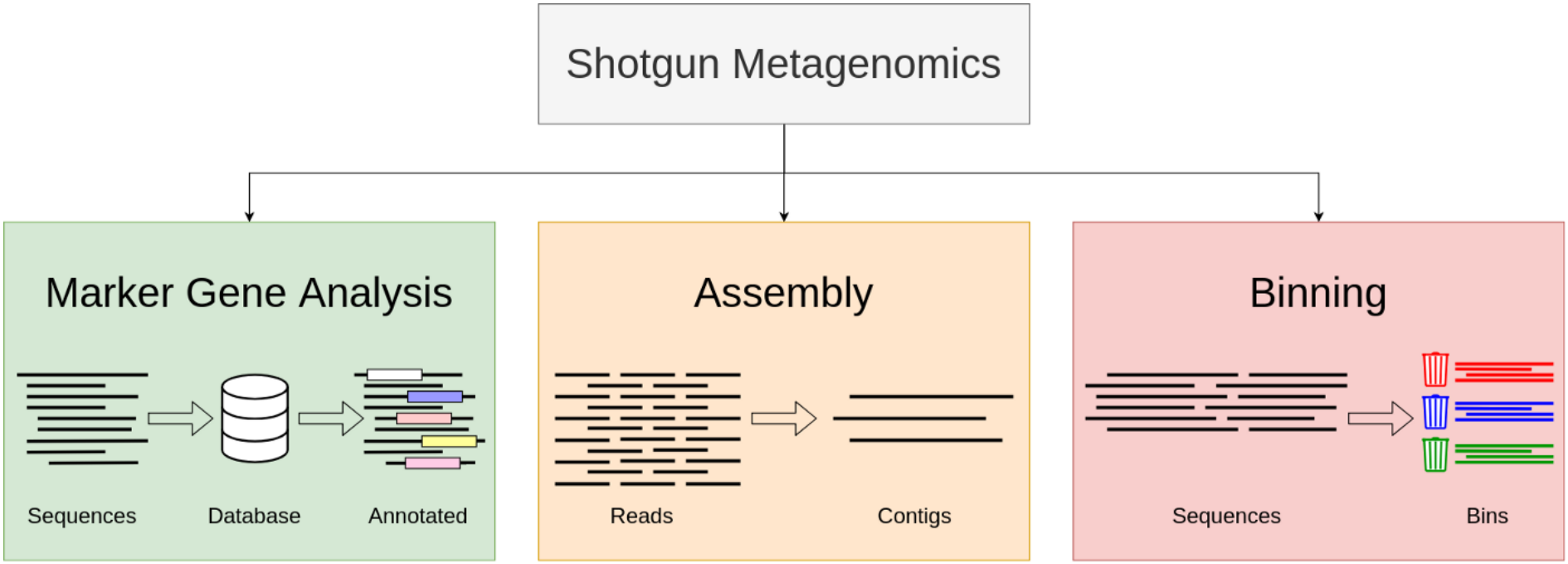
Source: https://github.com/grimmlab/MicrobiomeBestPracticeReview

Metagenome Data Analysis: 1. Shotgun Sequencing

Challenges:

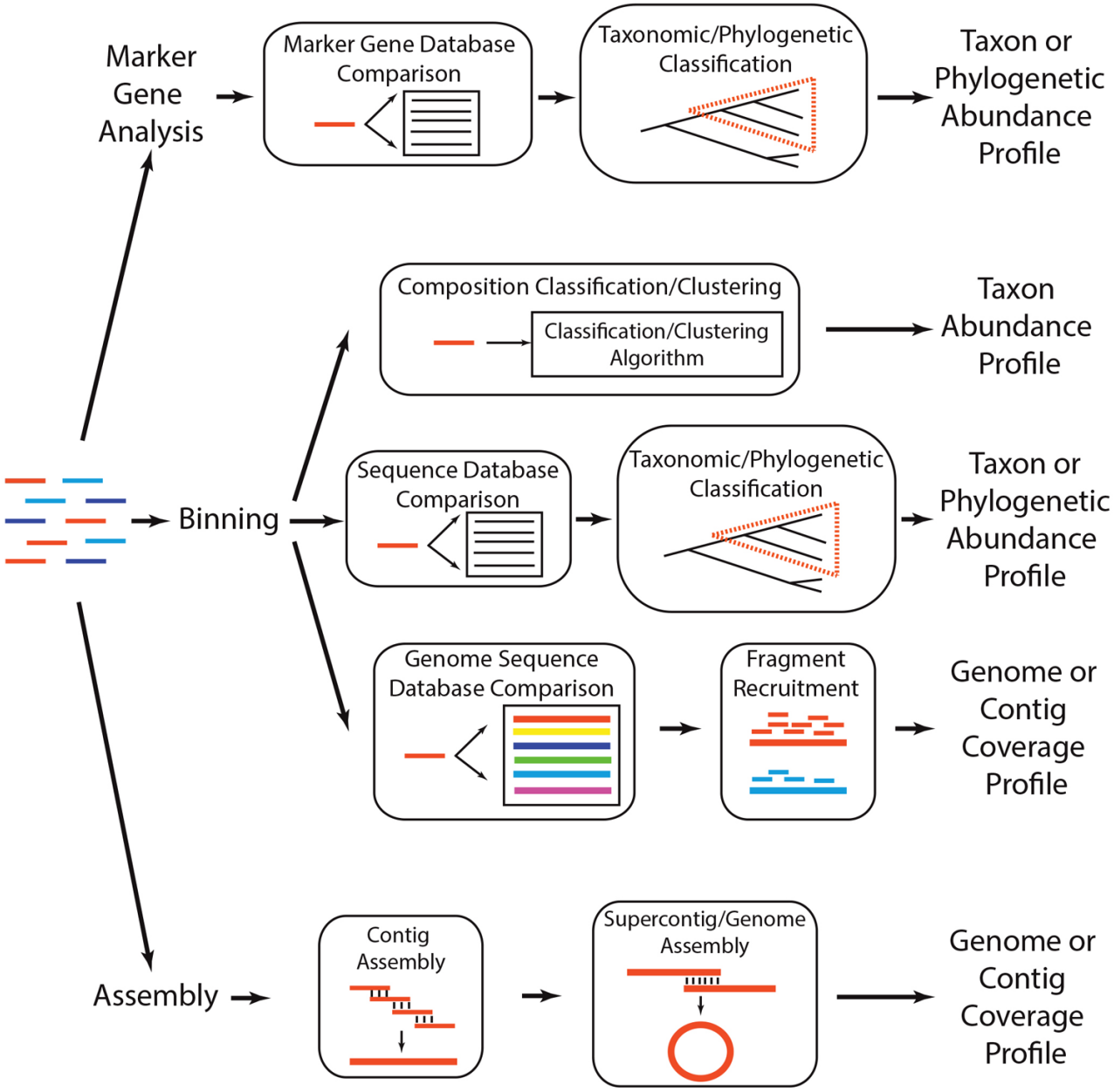
- Difficult to determine from which genome a read is derived
- Most communities are so diverse that most genomes are not completely represented by reads
- Sometimes it is not clear whether overlapping reads are from the same genome
- Contamination from the host genome can complicate the analysis
- **Expensive** to generate data and analyze

Metagenome Data Analysis: 1. Shotgun Sequencing



Source: Sharpton. (2014) An introduction to the analysis of shotgun metagenomic data.

Metagenome Data Analysis: 1. Shotgun Sequencing



Source: Sharpton. (2014) An introduction to the analysis of shotgun metagenomic data.

Metagenome Data Analysis: 1. Shotgun Sequencing

Binning

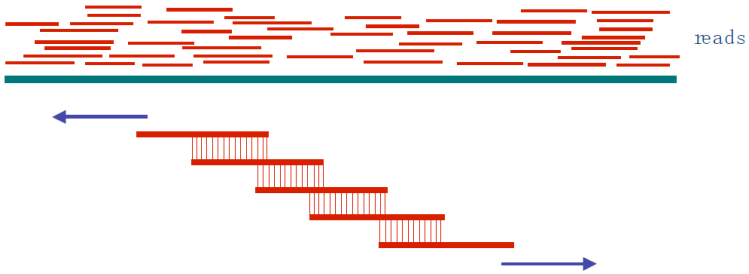
Binning refers to the process of sorting DNA sequences into groups that might represent an individual genome or genomes from closely related organisms.

- G/C content, gene content
- Codon usage
- Particular abundance distribution of K-mer

Metagenome Data Analysis: 1. Shotgun Sequencing

Genome assembly

Reference-based assembly



- Works well if metagenomic dataset contains sequences where closely related reference genomes are available.

De novo assembly

From scratch. Without any reference sequences.

- Requires larger computational resources (hundreds of GB's of memory)
- Longer run time (days)
- No prior knowledge of the organisms is required

Metagenome Data Analysis: 1. Shotgun Sequencing

Annotation

- Process of classifying predicted genes into known and well-characterized gene families
- Based on similarity to known sequences

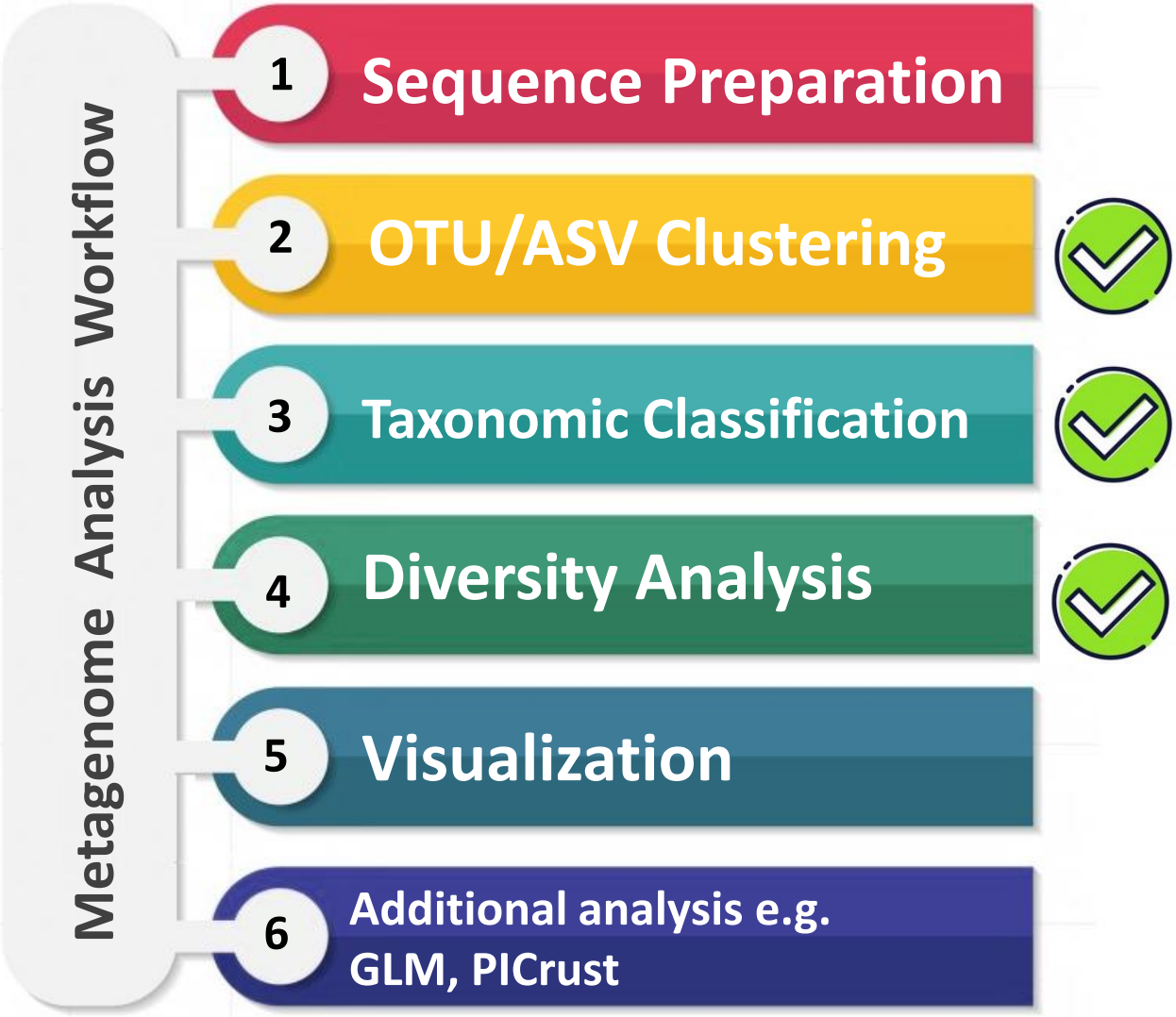
Gene Ontology covers three domains:

I. **cellular component** - the parts of a cell or its extracellular environment

II. **molecular function** - activities of a gene product at the molecular level, such as binding or catalysis

III. **biological process** - operations or sets of molecular events

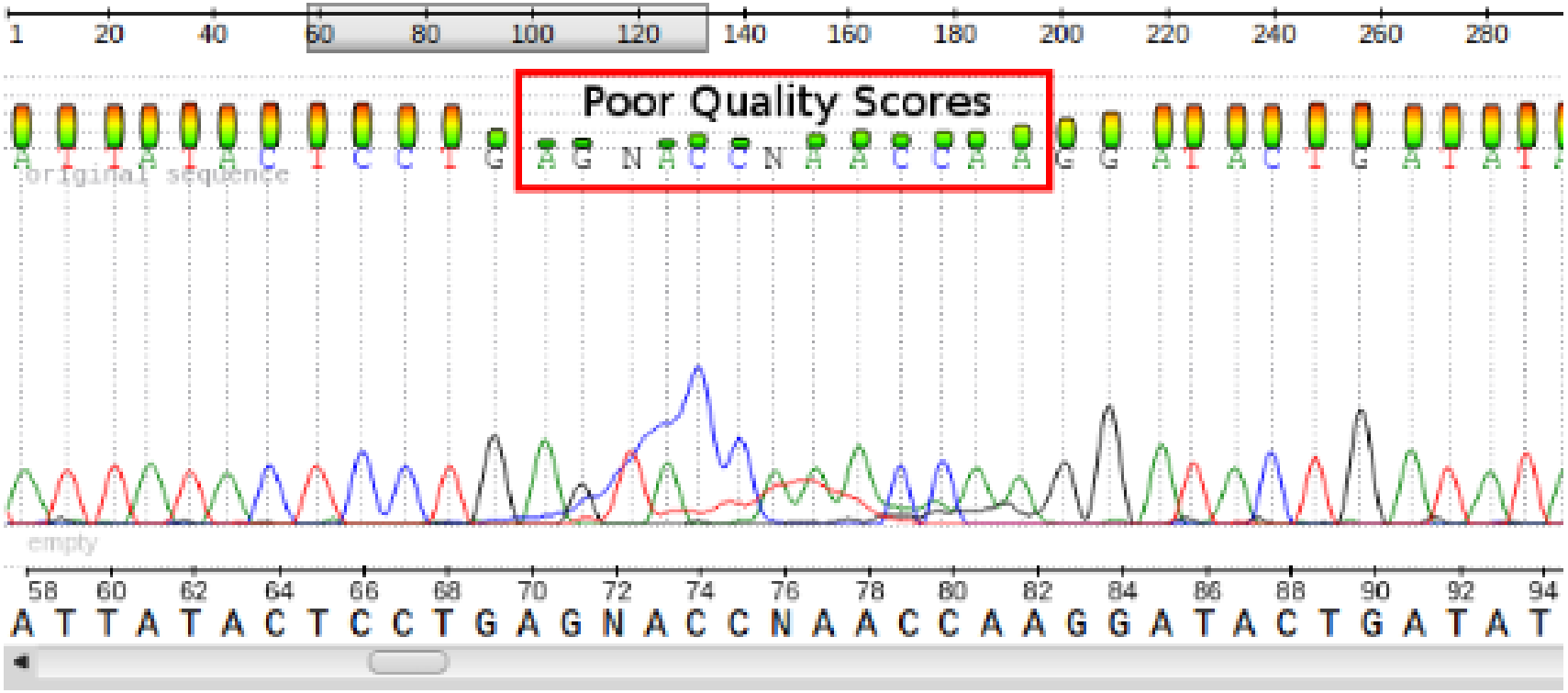
Metagenome Data Analysis: 2. Amplicon Sequencing



Metagenome Data Analysis: 2. Amplicon Sequencing

1. Sequence Preparation

- Filtering low-quality sequences
- Trimming barcode and primer sequences



Metagenome Data Analysis: 2. Amplicon Sequencing

1. Sequence Preparation

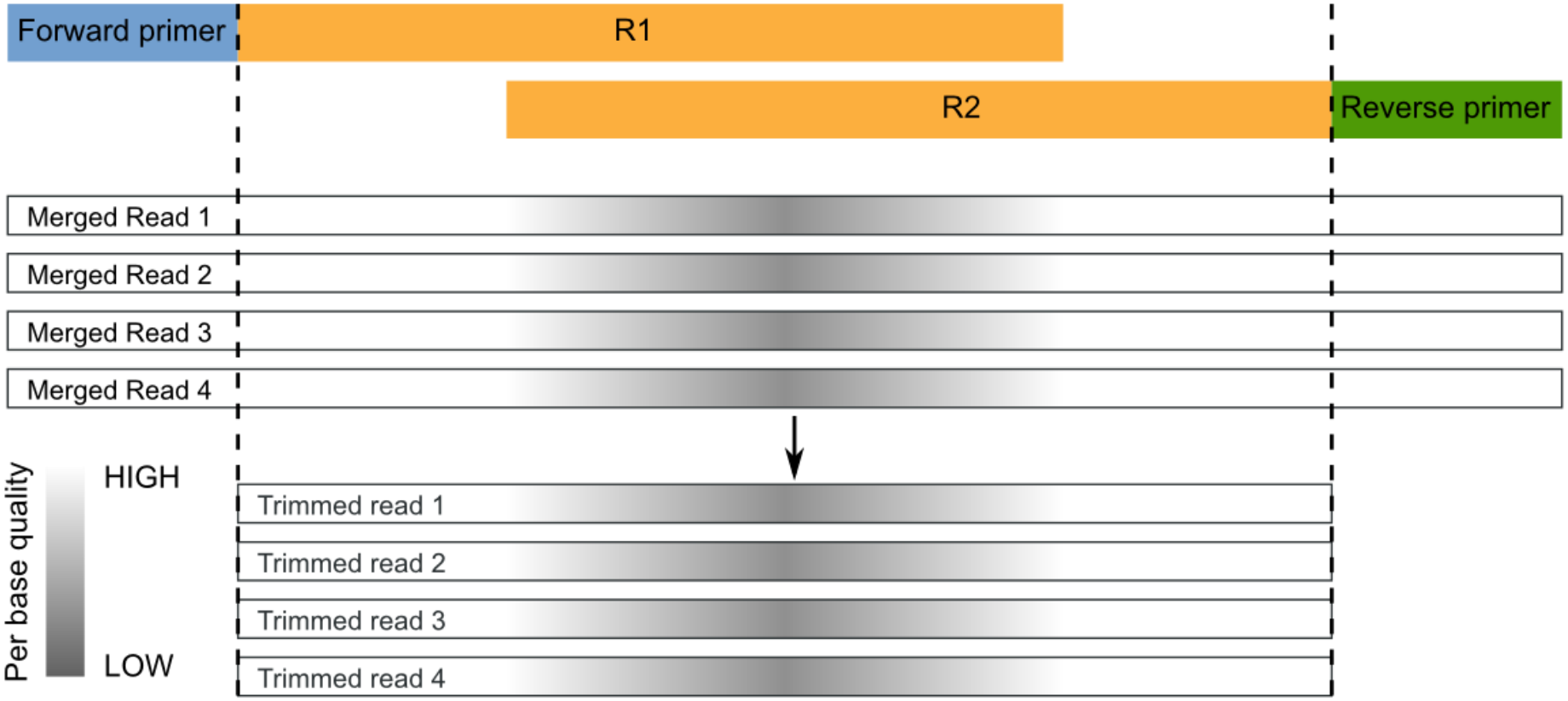
- Filtering low-quality sequences
- Trimming barcode and primer sequences

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Metagenome Data Analysis: 2. Amplicon Sequencing

1. Sequence Preparation

- Join Pair-end sequences



2. OTU/ASV Clustering

What is the Operational taxonomic unit (OTU)

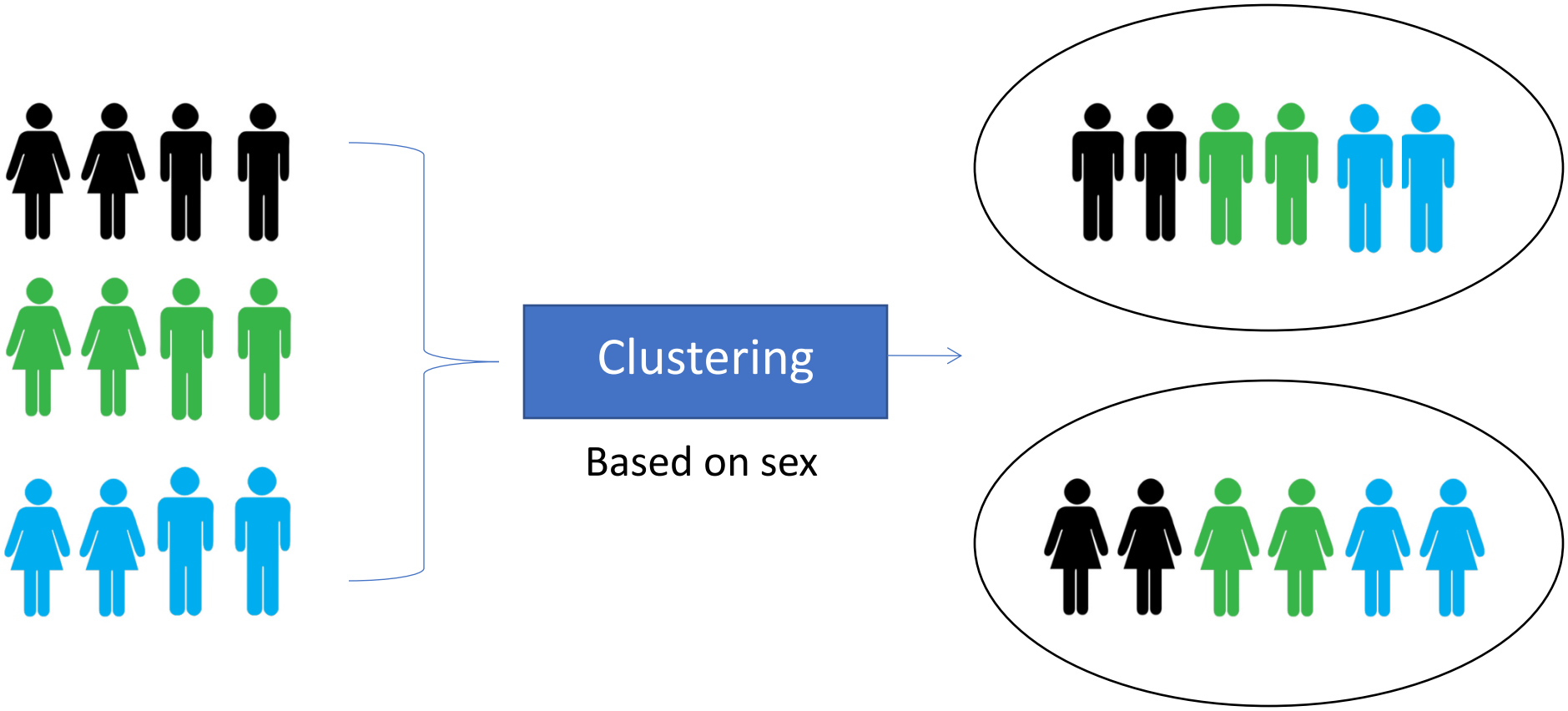
- **"The thing(s) being studied** [1]" ➔ individual organism, a named taxonomic group, or a group with undetermined evolutionary relationships

- "**OTU**" is generally used in a different context and refers to **clusters of organisms**, grouped by DNA sequence similarity of a specific taxonomic marker gene.[2]

- **Units of microbial diversity**, especially when analyzing small subunit **16S** or **18S** rRNA marker gene sequence datasets.

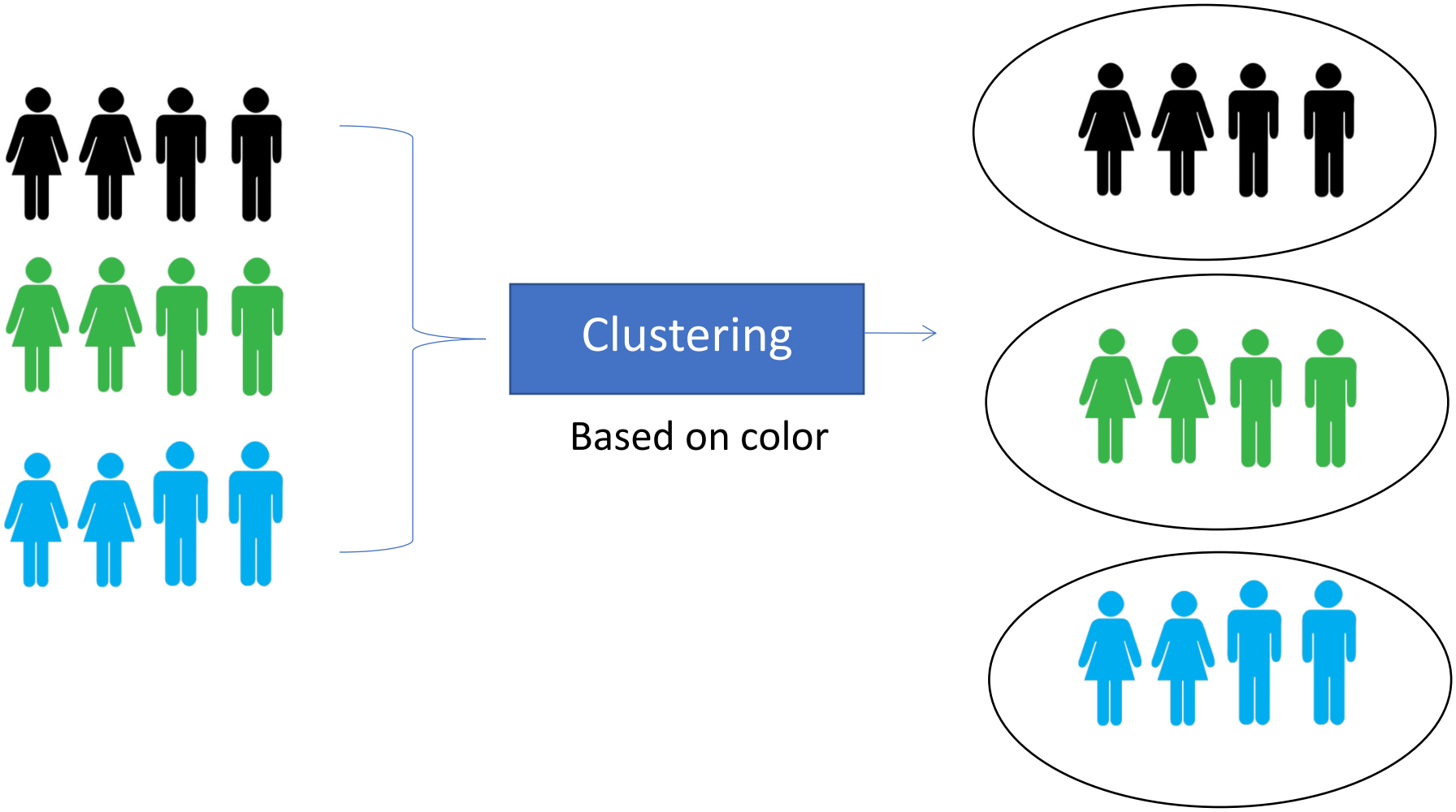
Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering



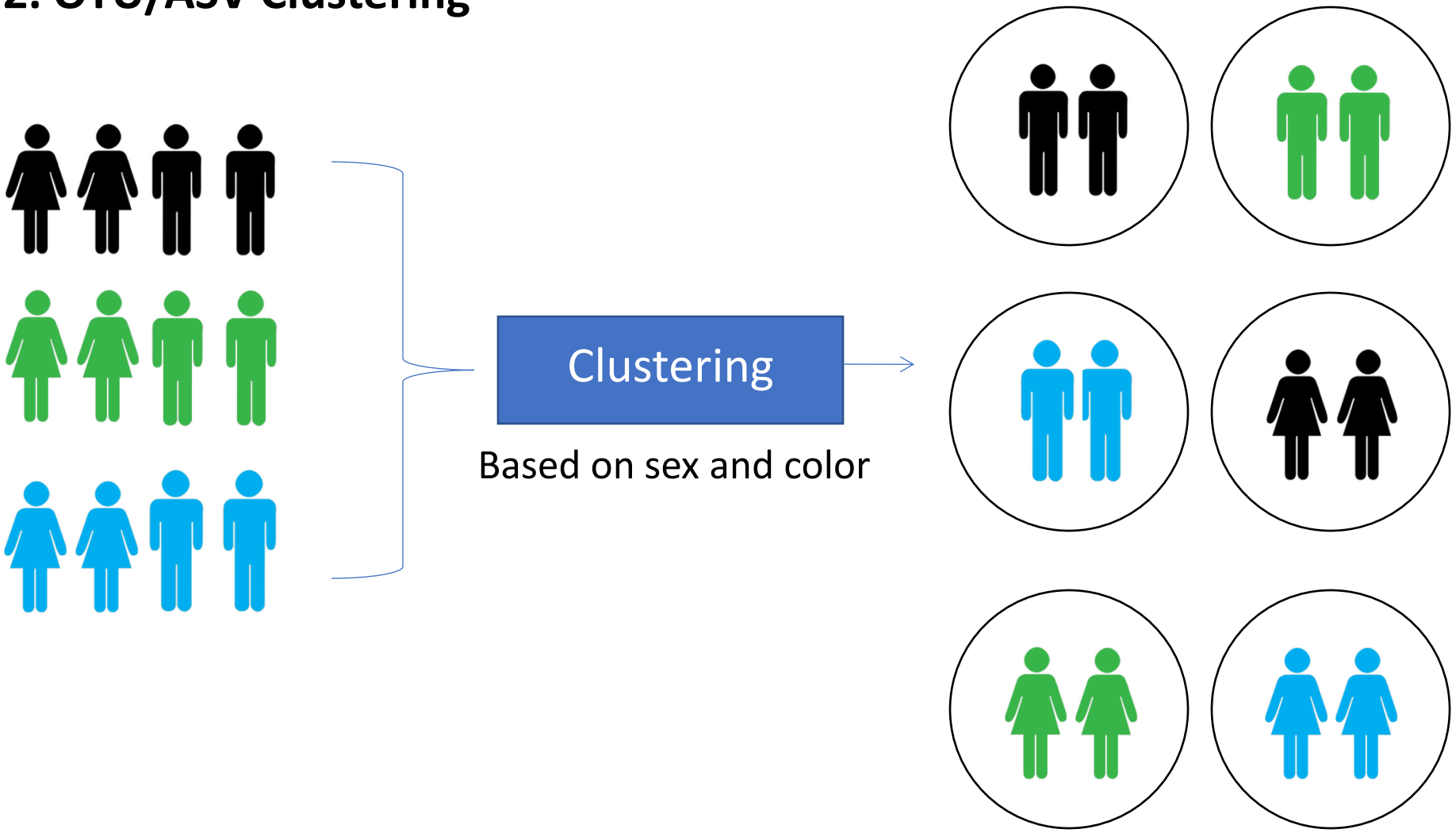
Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering



Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering



Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering

>unknown sequence 1
ATGCGTACTGATGCATGGTCAGTATGC……
>unknown sequence 2
ATGCGTACTGATGCAGTATGCTGGTCA…….
>unknown sequence 3
CGTACTGATGATGCATGGTCAGTATGC…….
……
>unknown sequence 20,000
CGTACTGATGATGCATGGTCAGTATGC…….

CLUSTERING

(based on sequence similarity)

Cluster of sequences

Cluster 1
unknown cow sequence 1
unknown cow sequence 2
unknown cow sequence 5

Cluster 2
unknown cow sequence 26
unknown cow sequence 34
unknown cow sequence 50

……

……

Cluster n
unknown cow sequence x
unknown cow sequence y
unknown cow sequence z

Similarity 99% (0.01) ➔ strain
Similarity 97% (0.03) ➔ species
Similarity 95% (0.05) ➔ genus
Similarity 85% (0.15) ➔ class

- Distance
- Dissimilarity
- OTU definition
- OTU cutoff

Operational taxonomic unit (OTU)

Number of OTU == number of cluster

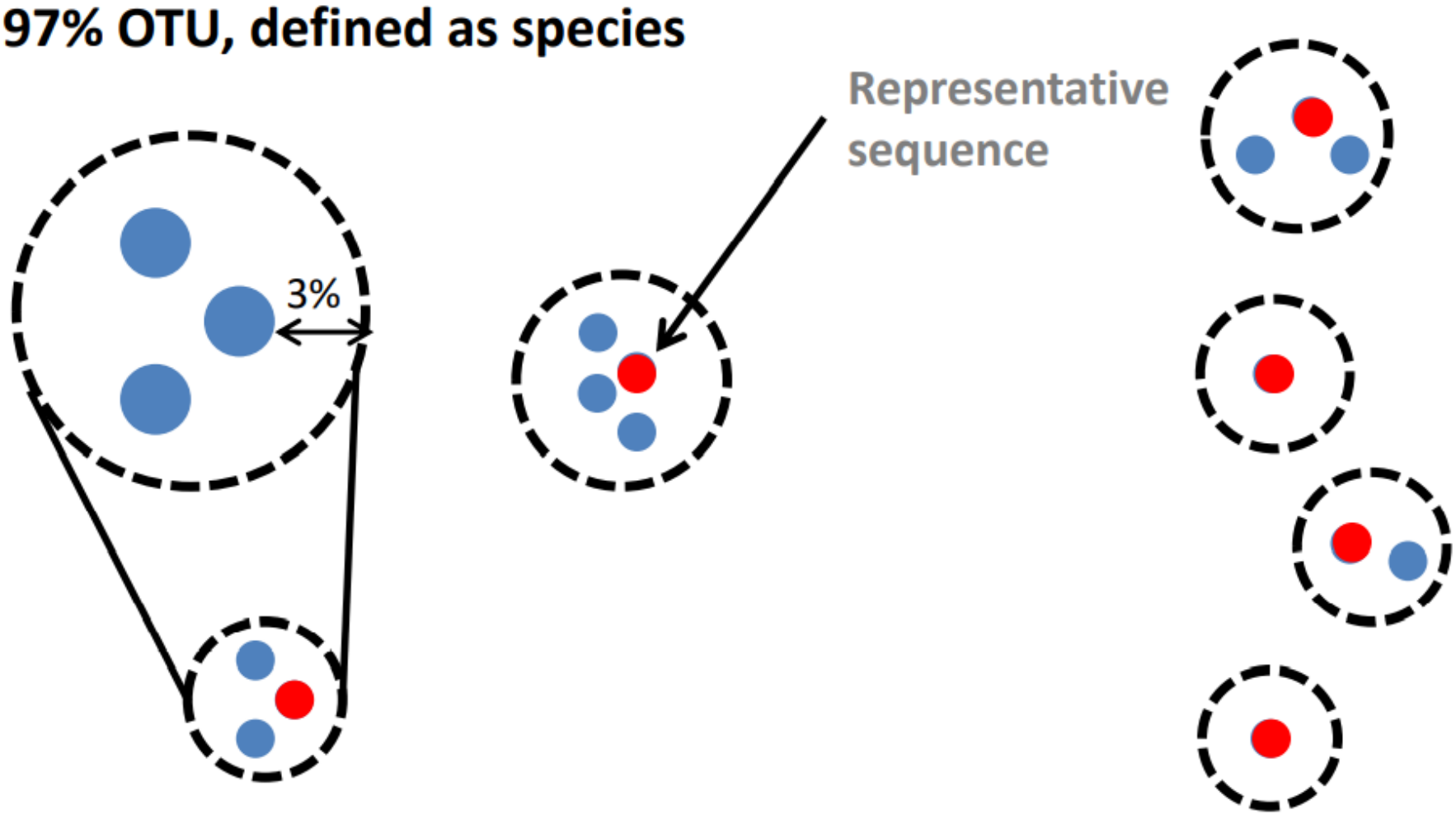
Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering: OTU picking method

QIIME provides 3 methods to define OTU

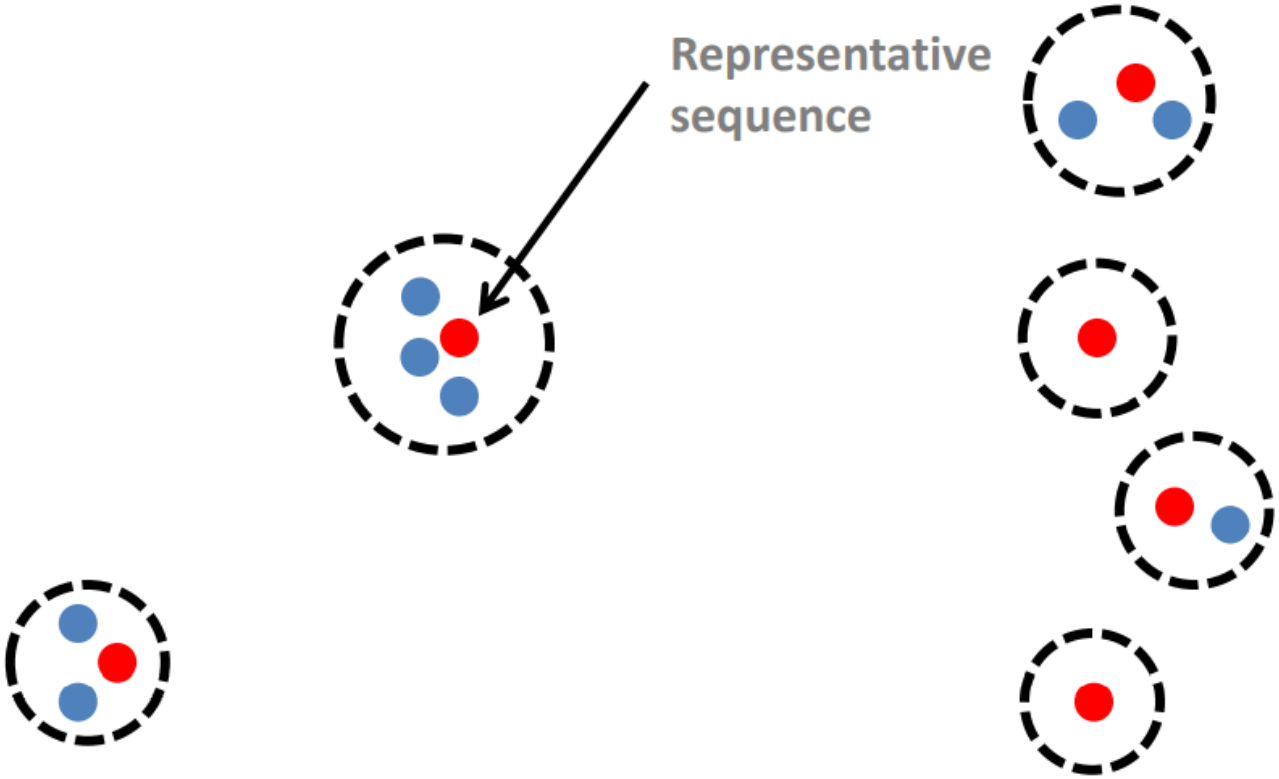
Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering: OTU picking method



Metagenome Data Analysis: 2. Amplicon Sequencing

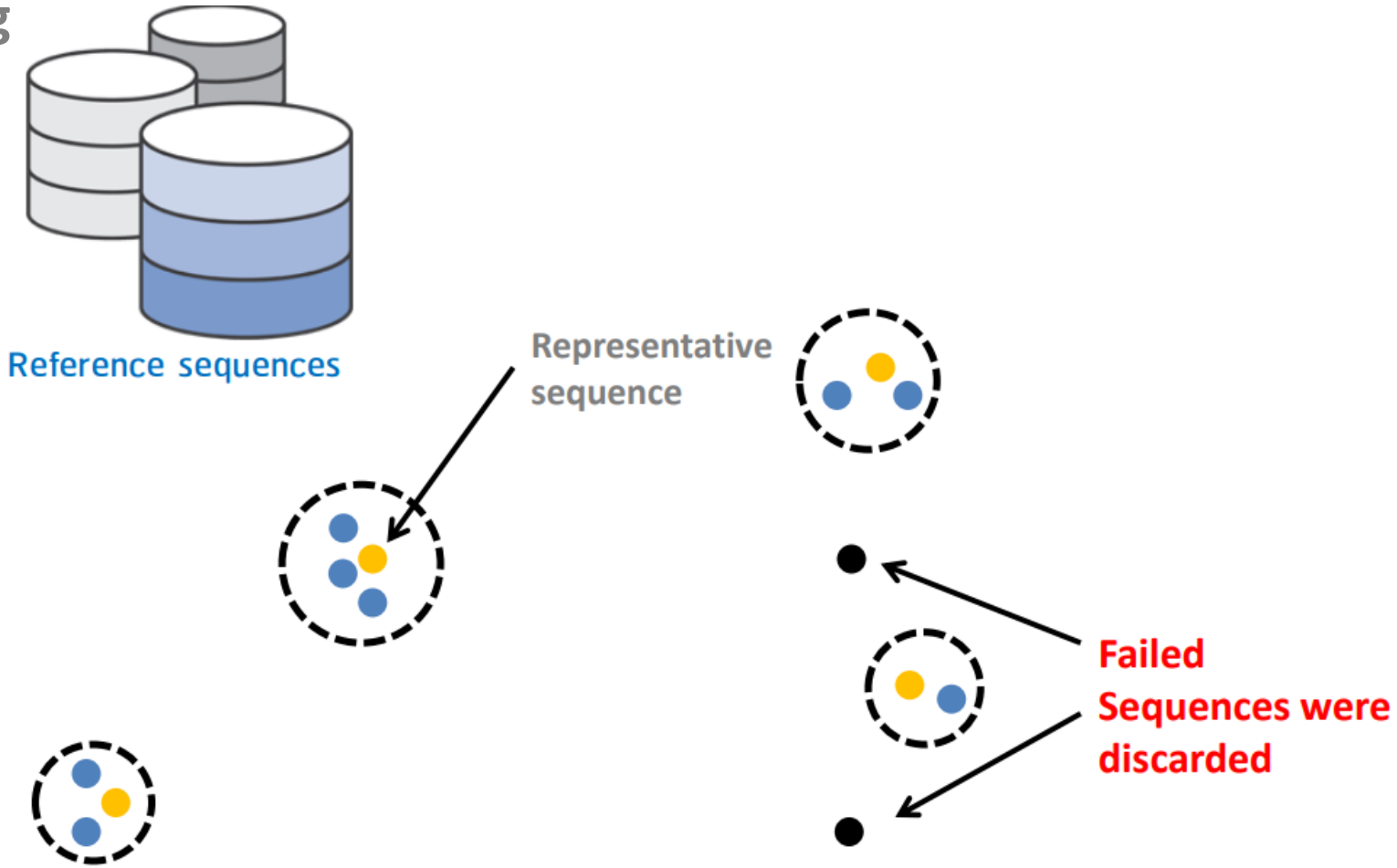
2. OTU/ASV Clustering: OTU picking method : 1. *De novo* OTU picking



De novo
OTU picking

Metagenome Data Analysis: 2. Amplicon Sequencing

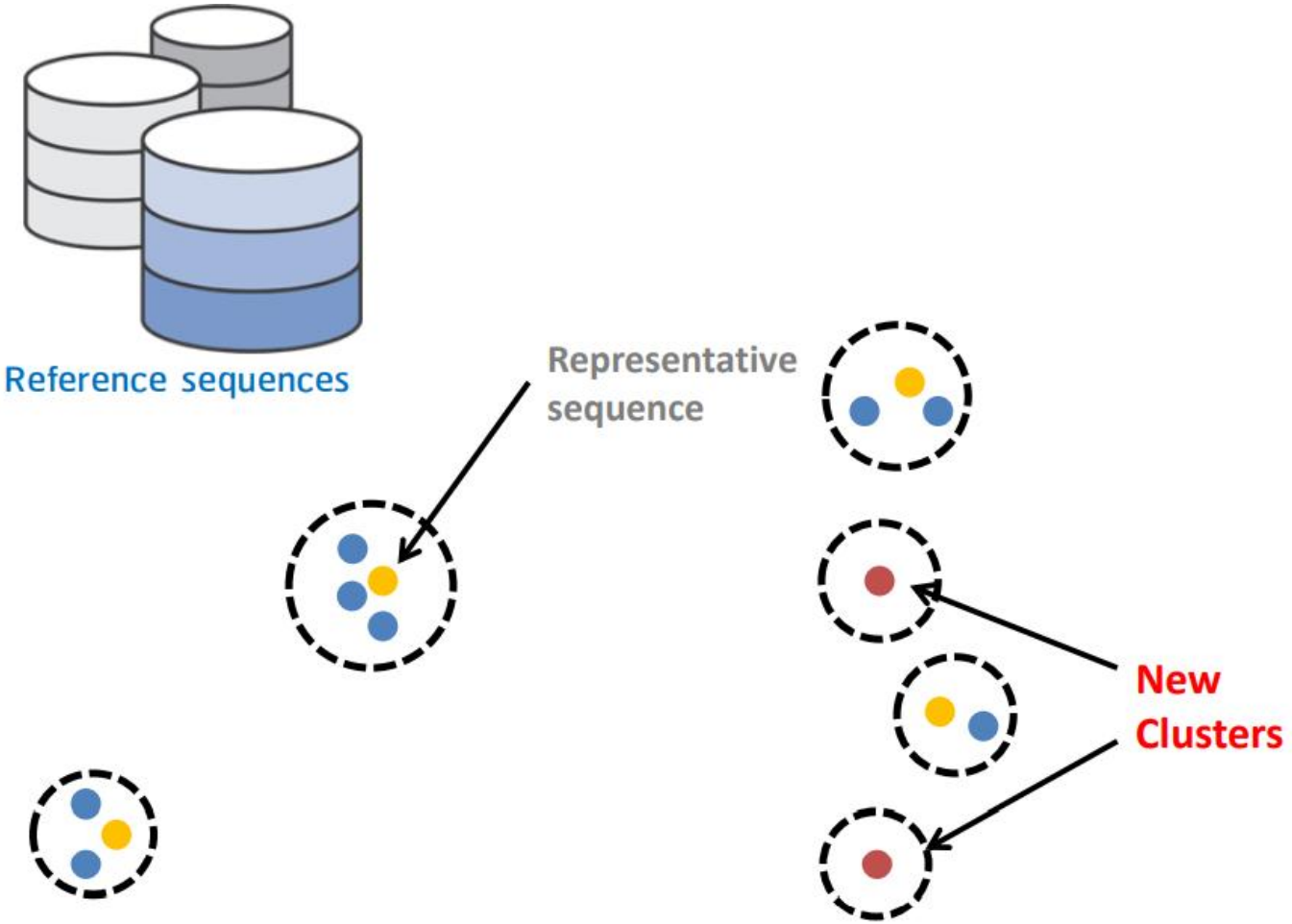
2. OTU/ASV Clustering: OTU picking method : 2. Close referenced OTU picking



Close Reference OTU picking

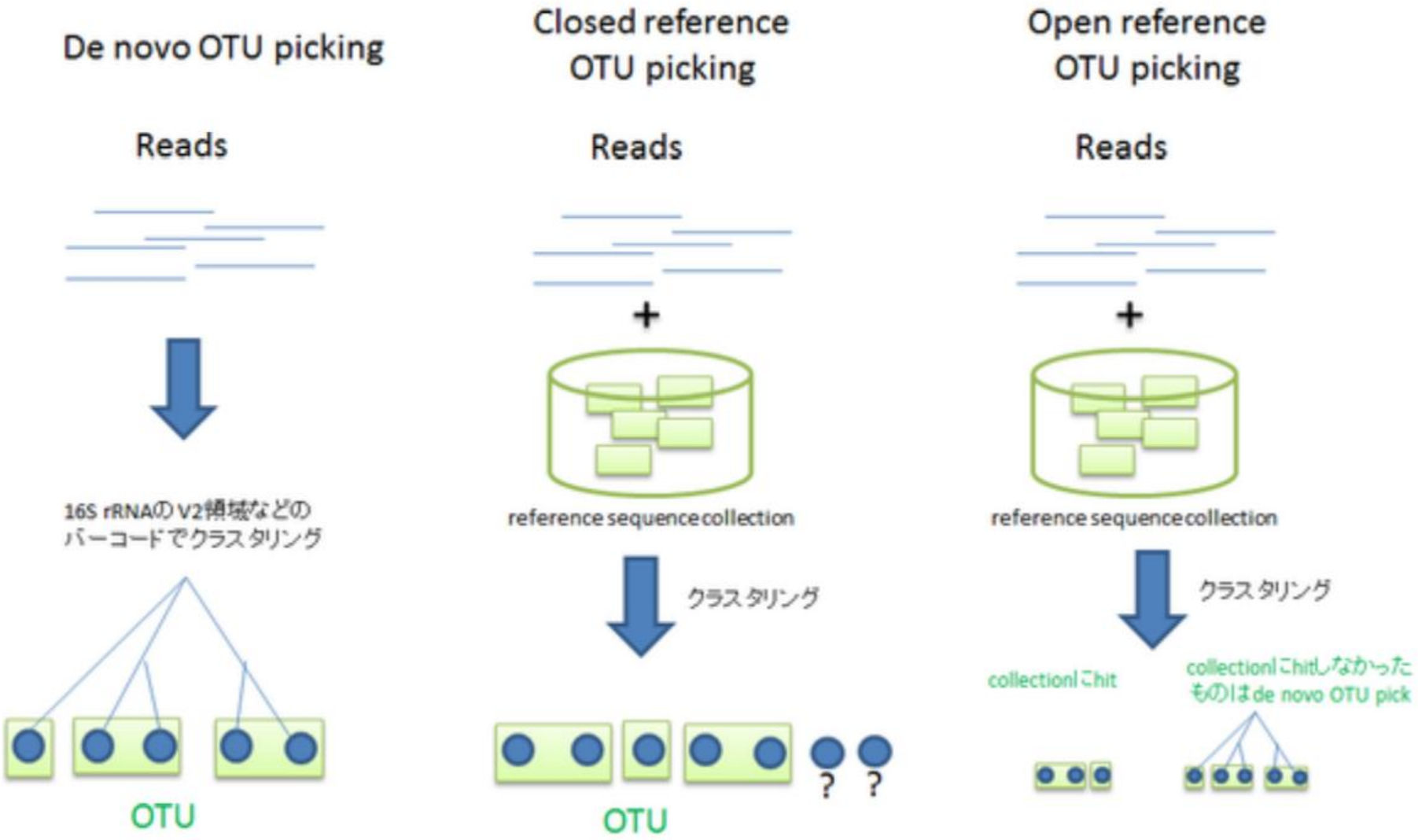
Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering: OTU picking method : 3. Open referenced OTU picking



Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering



Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering

	Advantages	Disadvantages
De novo	- All reads are clustered	- Slow - OTUs may be defined by erroneous reads
Close Reference	- Built-in quality filter - Fast - OTUs are defined by high-quality, trusted sequences	- Reads that don't hit reference dataset are excluded, so you can never observe new OTUs
Open Reference	- All reads are clustered	- Moderate speed - Mix of high quality sequences defining OTUs (i.e., the database sequences) and possible low quality sequences defining OTUs (i.e., the sequencing reads)

Each of these protocols are briefly described in this document; for a more detailed discussion of these OTU picking protocols, please see Rideout et al. (2014).

Source: http://qiime.org/tutorials/otu_picking.html; https://peerj.com/articles/545/

Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering

OTU table

Count table

	S1	S2	S3
OTU1	100	0	0
OTU2	100	40	600
OTU3	0	10	0

Relative abundance table

	S1	S2	S3
OTU1	.5	0	0
OTU2	.5	.8	1.0
OTU3	0	.2	0

After OTU picking step: we got the OTU table

#OTU ID	F3D0	F3D141	F3D142	F3D143	F3D144	F3D145	F3D146	F3D147
OTU_6	749	535	313	372	607	849	493	2025
OTU_25	29	57	14	2	14	22	16	127
OTU_1	613	497	312	247	472	719	349	1720
OTU_8	426	378	255	237	382	627	330	1417
OTU_31	149	38	10	19	25	21	43	31
OTU_2	366	392	327	185	313	542	248	1367
OTU_7	196	370	92	107	48	155	74	105
OTU_10	46	169	87	109	171	209	120	864
OTU_80	26	6	0	1	4	8	18	11

Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering

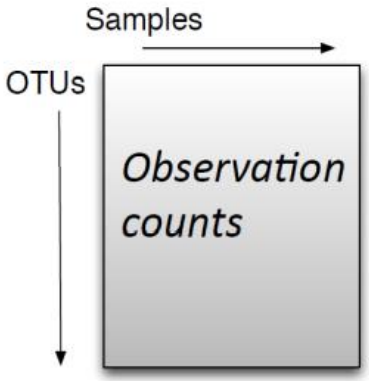
OTU tables are in Biological Observation Matrix (.biom) format

The Biological Observation Matrix (BIOM) format

The **BIOM file format** (canonically pronounced *biome*) is designed to be a general-use format for representing biological sample by observation contingency tables. BIOM is a recognized standard for the **Earth Microbiome Project** and is a **Genomics Standards Consortium** supported project.

The **BIOM format** is designed for general use in broad areas of comparative -omics. For example, in marker-gene surveys, the primary use of this format is to represent OTU tables: the observations in this case are OTUs and the matrix contains counts corresponding to the number of times each OTU is observed in each sample. With respect to metagenome data, this format would be used to represent metagenome tables: the observations in this case might correspond to SEED subsystems, and the matrix would contain counts corresponding to the number of times each subsystem is observed in each metagenome. Similarly, with respect to genome data, this format may be used to represent a set of genomes: the observations in this case again might correspond to SEED subsystems, and the counts would correspond to the number of times each subsystem is observed in each genome.

sample x observation contingency matrix



Biological Observation Matrix

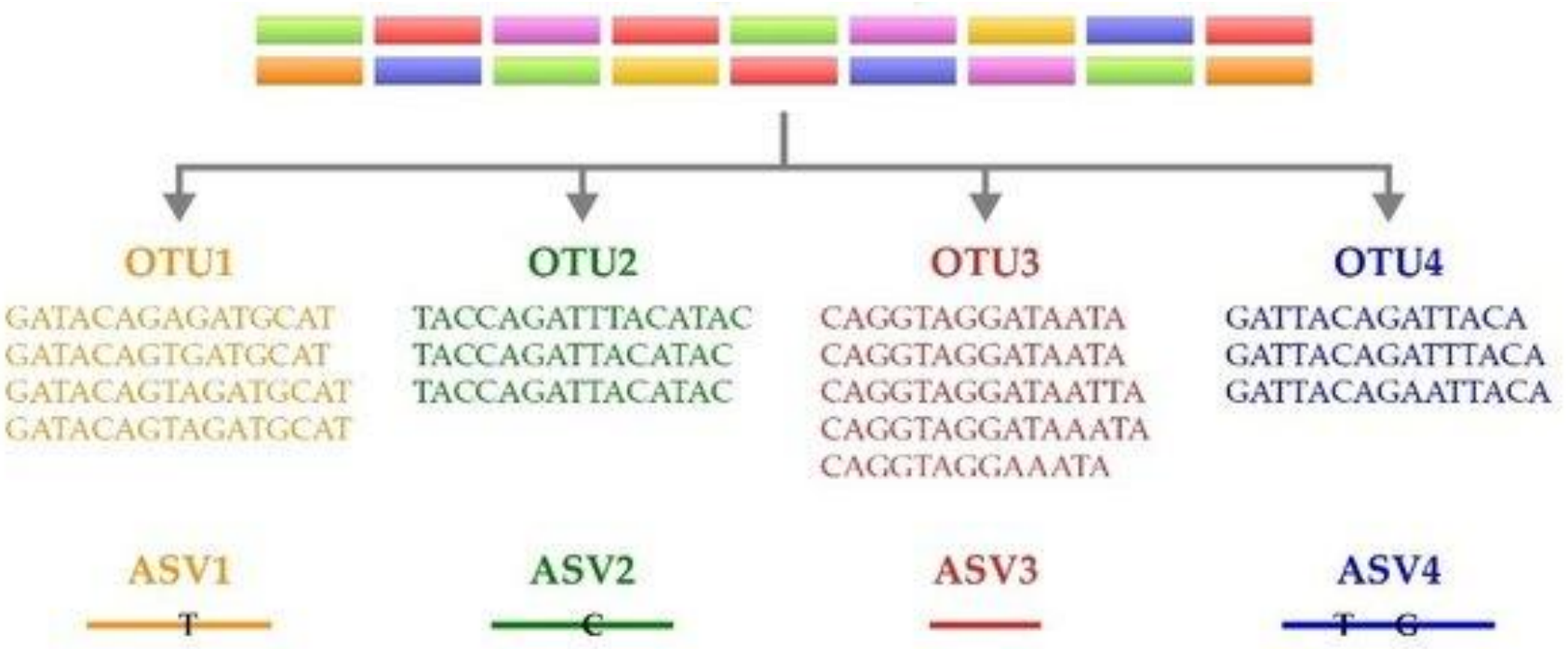
http://biom-format.org/

Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering

Amplicon Sequence Variant (ASV)

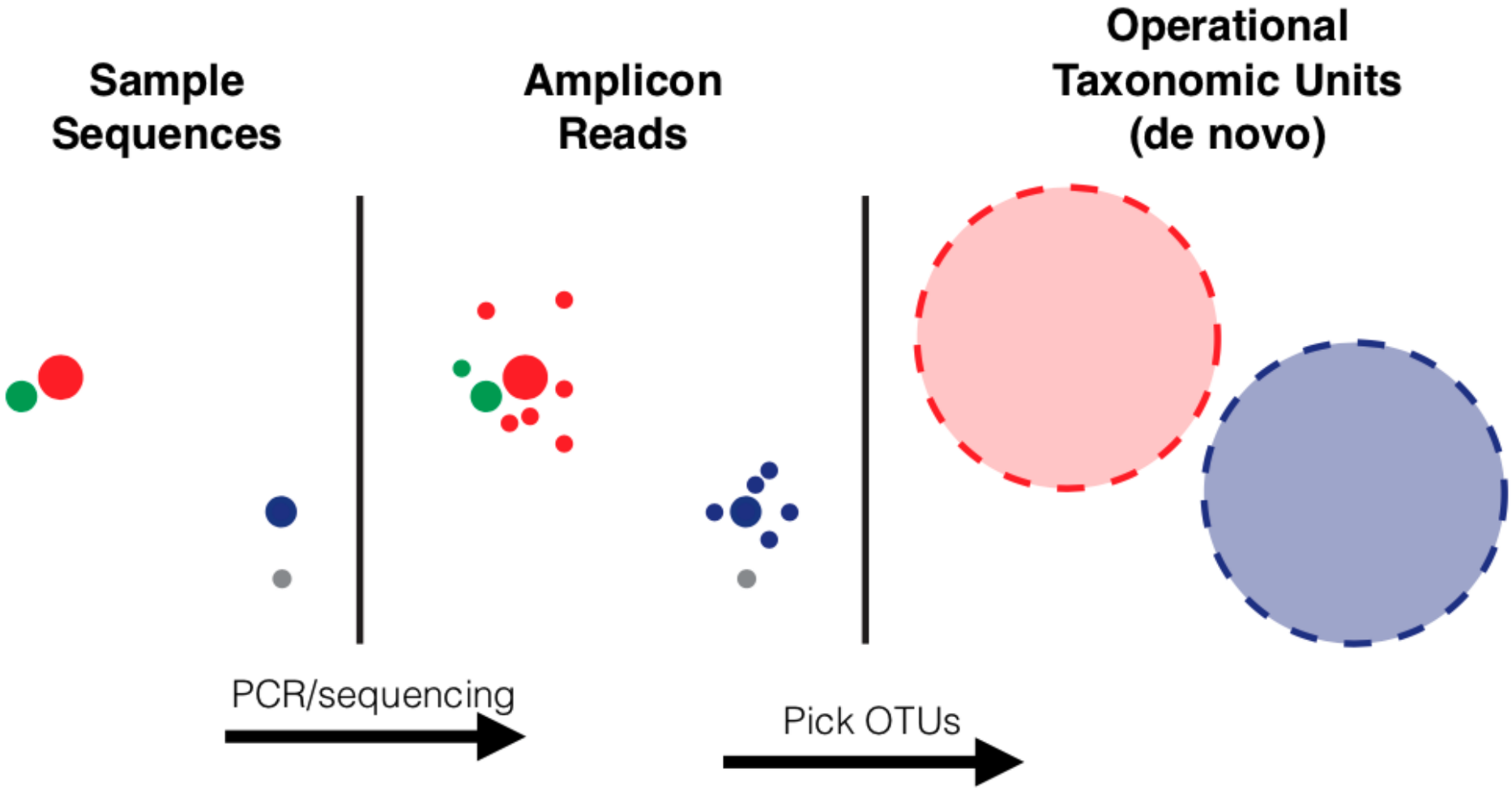
- Method to infer single DNA sequences recovered from marker genes.
- Remove an erroneous sequences generated during PCR and sequencing
- Distinguish sequence variation by a single nucleotide change.
- also referred to as exact sequence variants (ESVs), zero-radius OTUs (zOTUs), sub-OTUs (sOTUs)



Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering

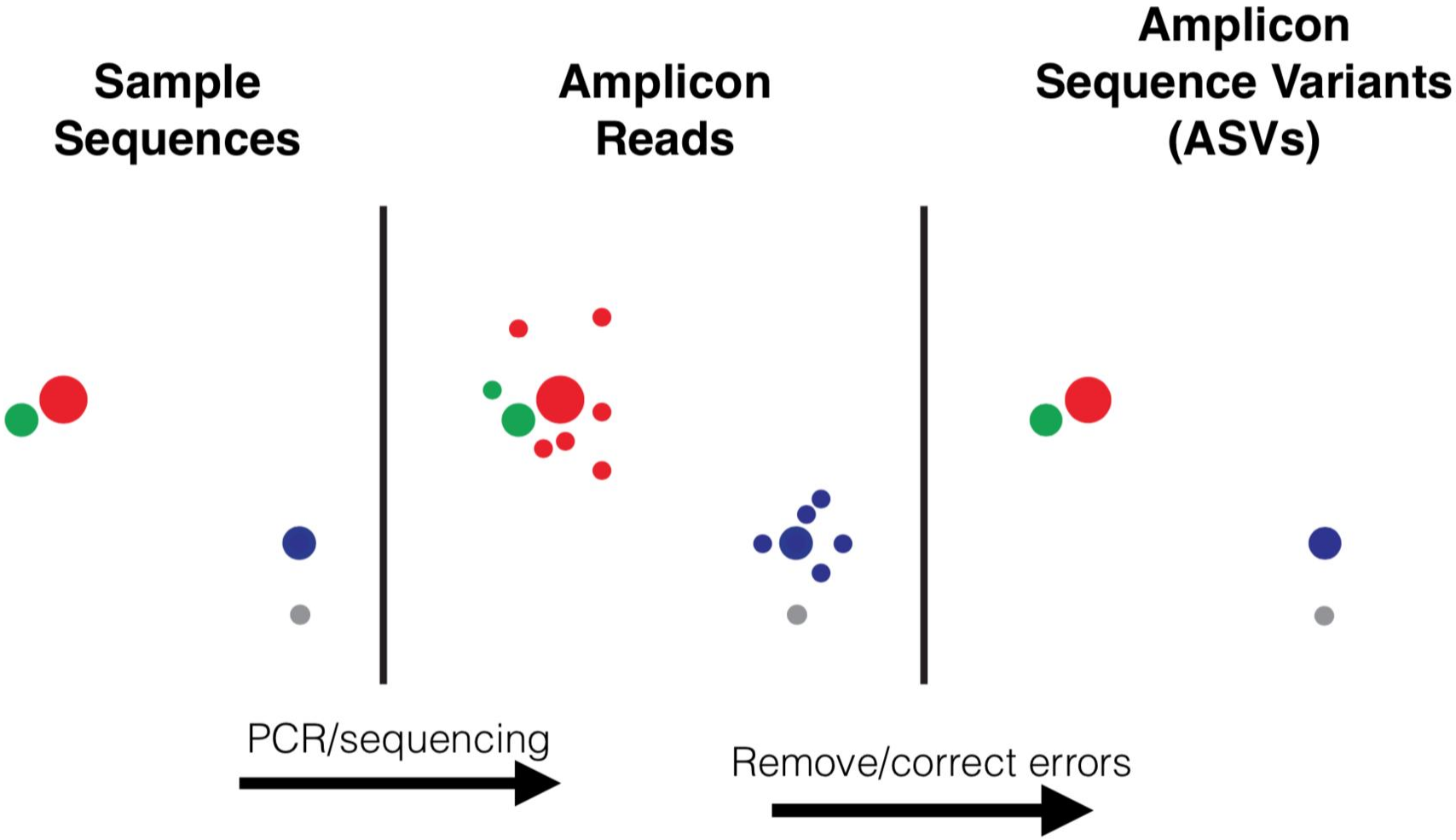
Amplicon Sequence Variant (ASV)



Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering

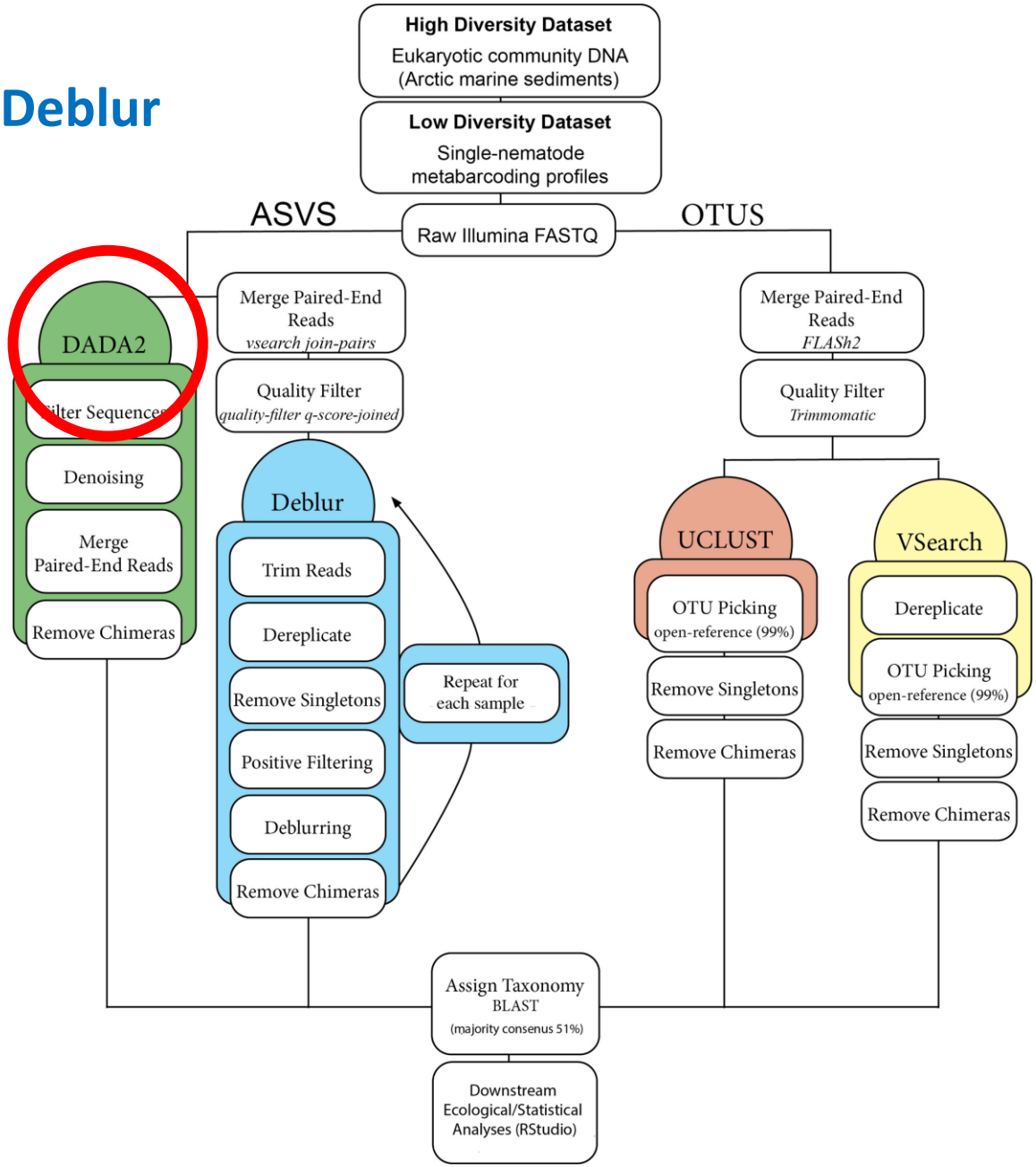
Amplicon Sequence Variant (ASV)



Metagenome Data Analysis: 2. Amplicon Sequencing

2. OTU/ASV Clustering

Tools for ASV: DADA2, Deblur



Metagenome Data Analysis: 2. Amplicon Sequencing

3. Taxonomic classification

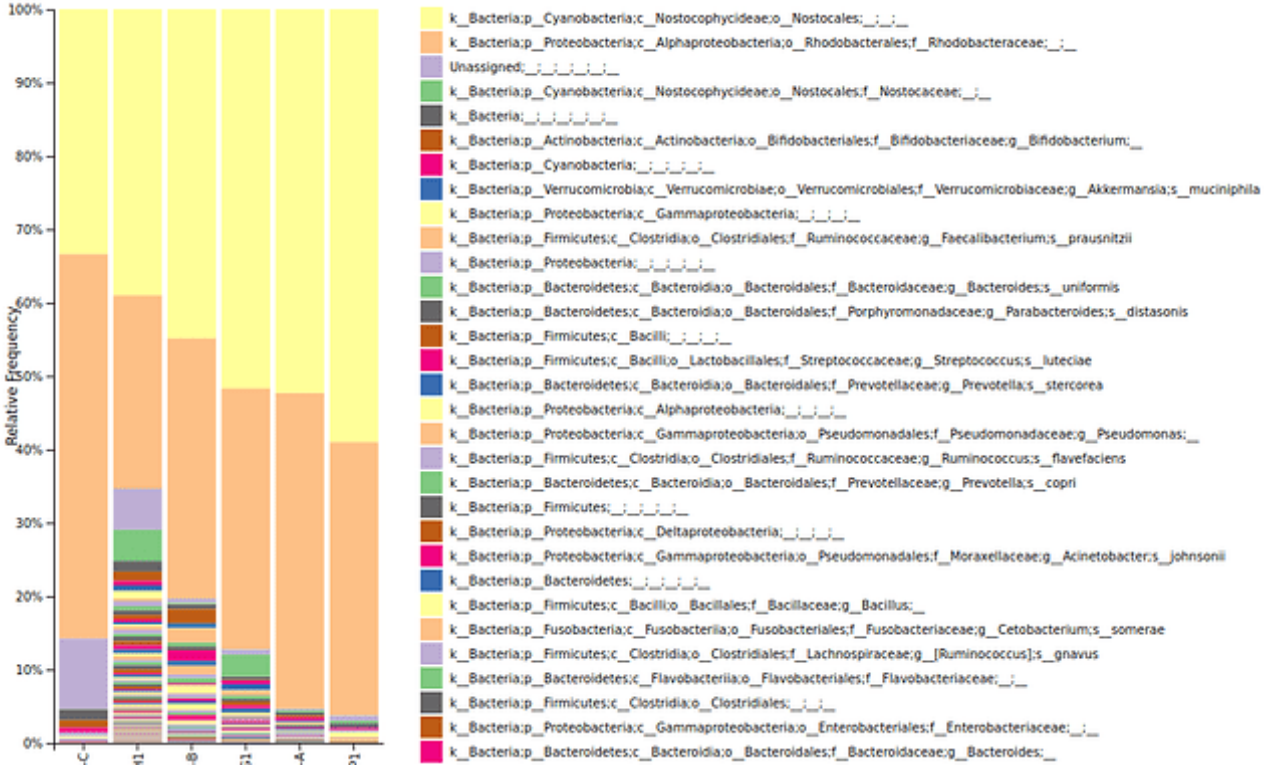
A. Homology consensus

BLAST+ consensus taxonomy classifier

VSEARCH-based consensus taxonomy classifier

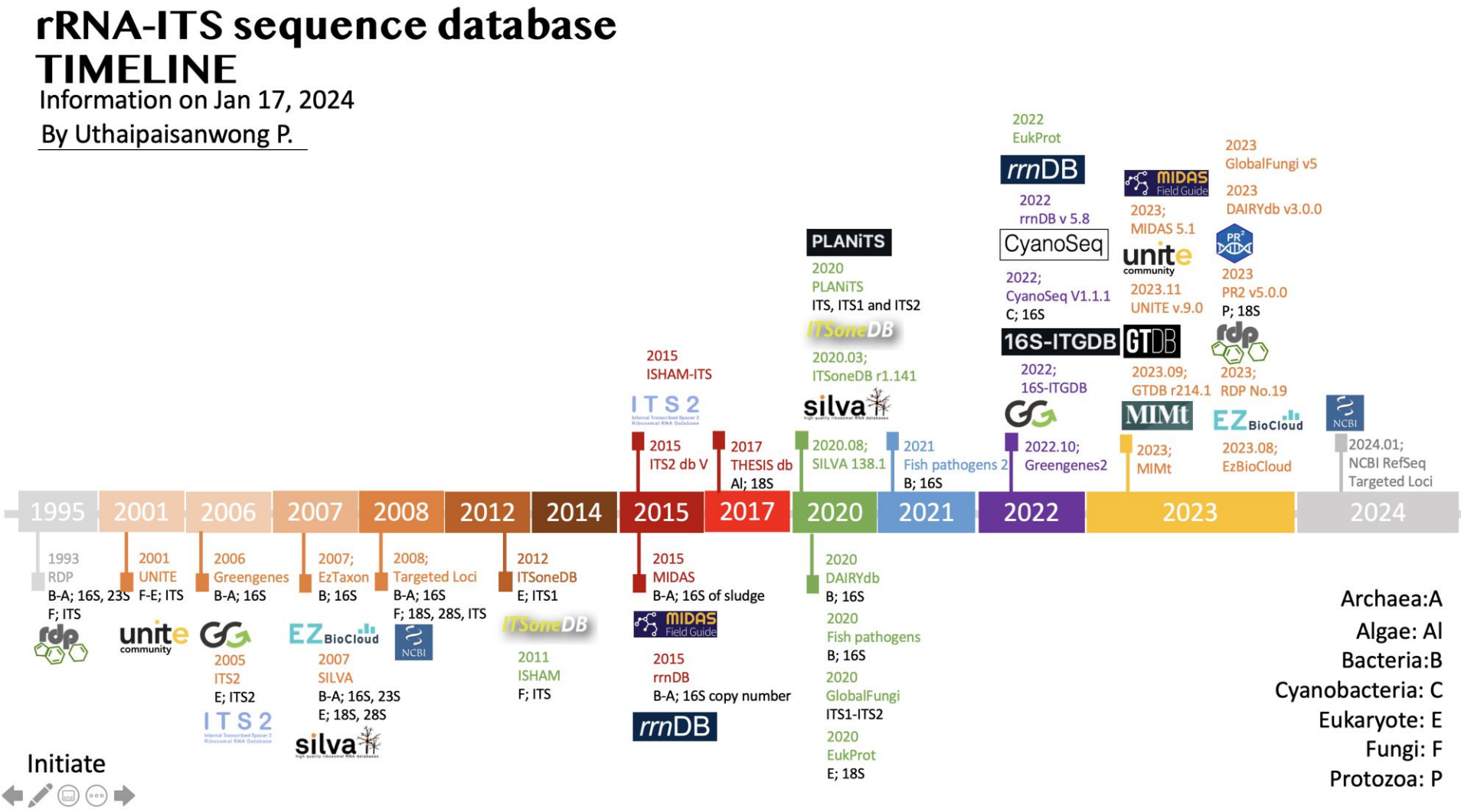
B. Machine Learning Classifier

classify-sklearn



Metagenome Data Analysis: 2. Amplicon Sequencing

3. Taxonomic classification



Metagenome Data Analysis: 2. Amplicon Sequencing

4. Diversity Analysis

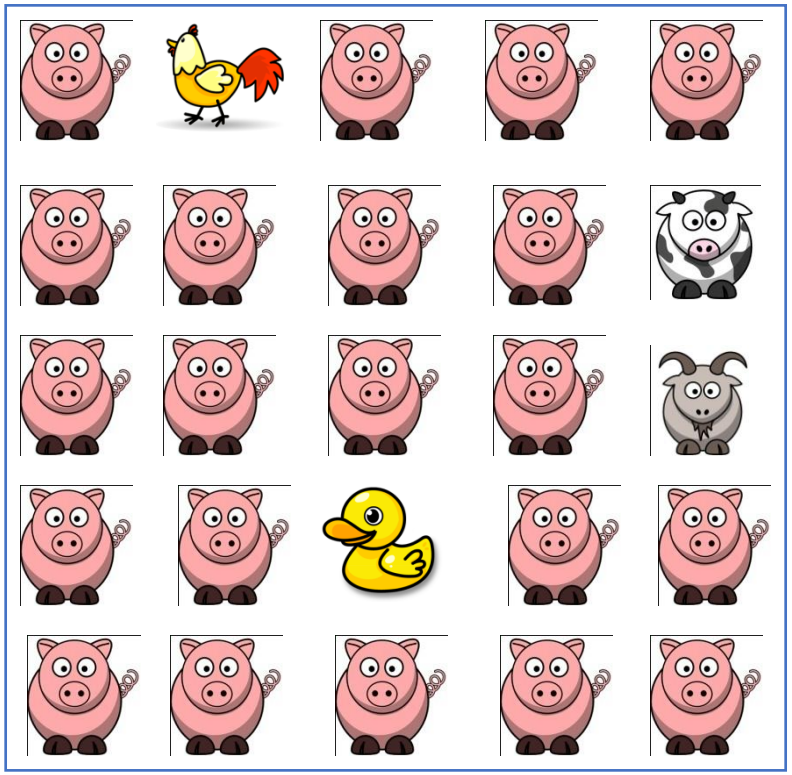
Biodiversity == **Richness** + **Evenness**

ความหลากหลายทางชีวภาพ == ความหลากชนิด + ความสม่ำเสมอ

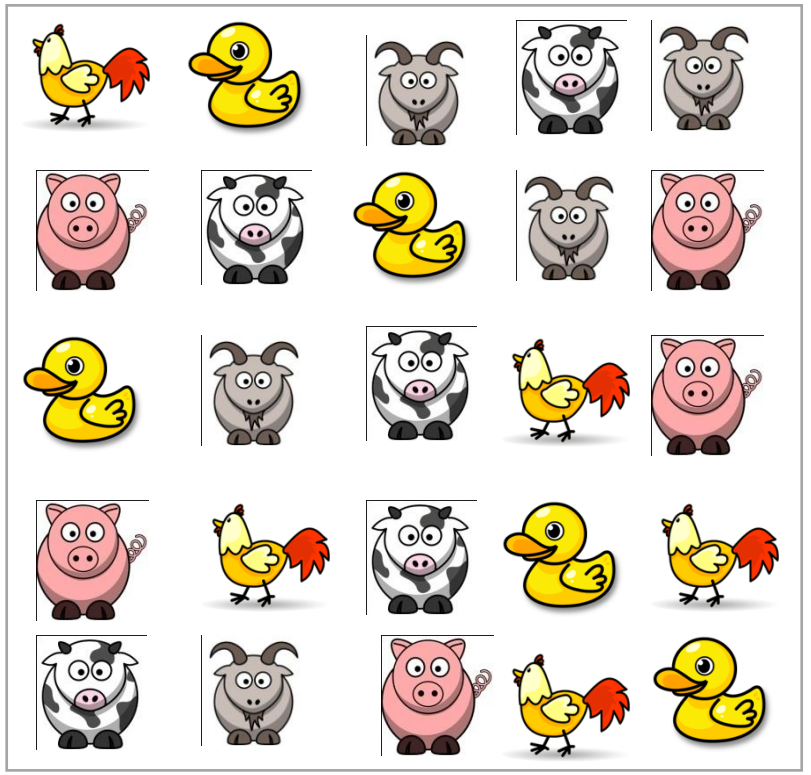
21 1 1
1 1 Total : 25

5 5 5
5 5 Total : 25

A



B



Metagenome Data Analysis: 2. Amplicon Sequencing

4. Diversity Analysis

Biodiversity == Richness + Evenness

Biodiversity index

Shannon-Wiener 's diversity index

$$H' = -\sum_{i=1}^{S} p_i \ln p_i$$

Evenness index

Shannon-Wiener 's evenness index

$$J' = \frac{H'}{H'_{\max}} \qquad H_{\max} = -\sum_{i=1}^{S} \frac{1}{S} \ln \frac{1}{S} = \ln S.$$

Claude Elwood Shannon (1916-2001)

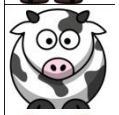
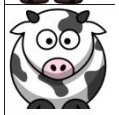
Simpson's diversity index

Simpson's evenness index

Edward Hugh Simpson

Metagenome Data Analysis: 2. Amplicon Sequencing

4. Diversity Analysis

	A	B
	 21  1  1  1  1 Total : 25	 5  5  5  5  5 Total : 25
Richness	5	5
Shannon-Wiener 's diversity index	0.662	1.610
Shannon-Wiener 's evenness index*	0.411	1.000

Metagenome Data Analysis: 2. Amplicon Sequencing

4. Diversity Analysis

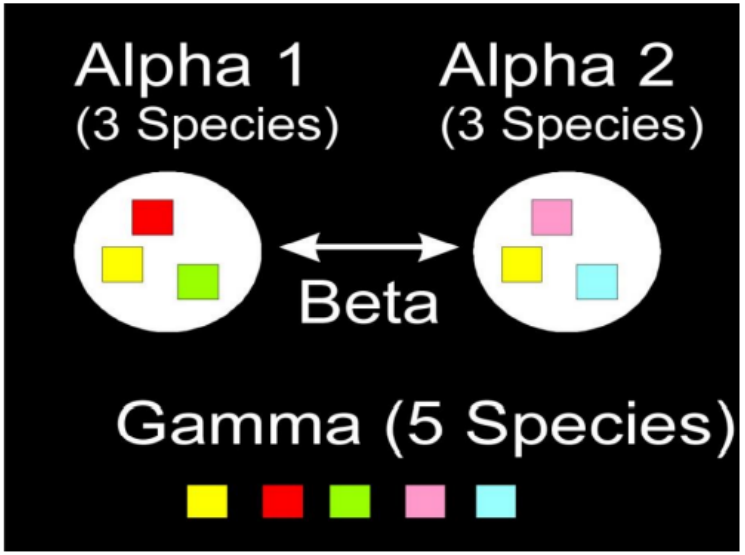
Alpha/Beta/Gamma Diversity

- Alpha Diversity (within a sample)
  - Species diversity in sites or habitats at a local scale.
  - e.g., how many species are in a sample

- Beta Diversity (between samples)
  - The differentiation in species composition between two sites or communities
  - e.g., how similar are two samples

- Gamma Diversity (total)
  - Total species diversity in a landscape

- There are several ways to calculate these diversities

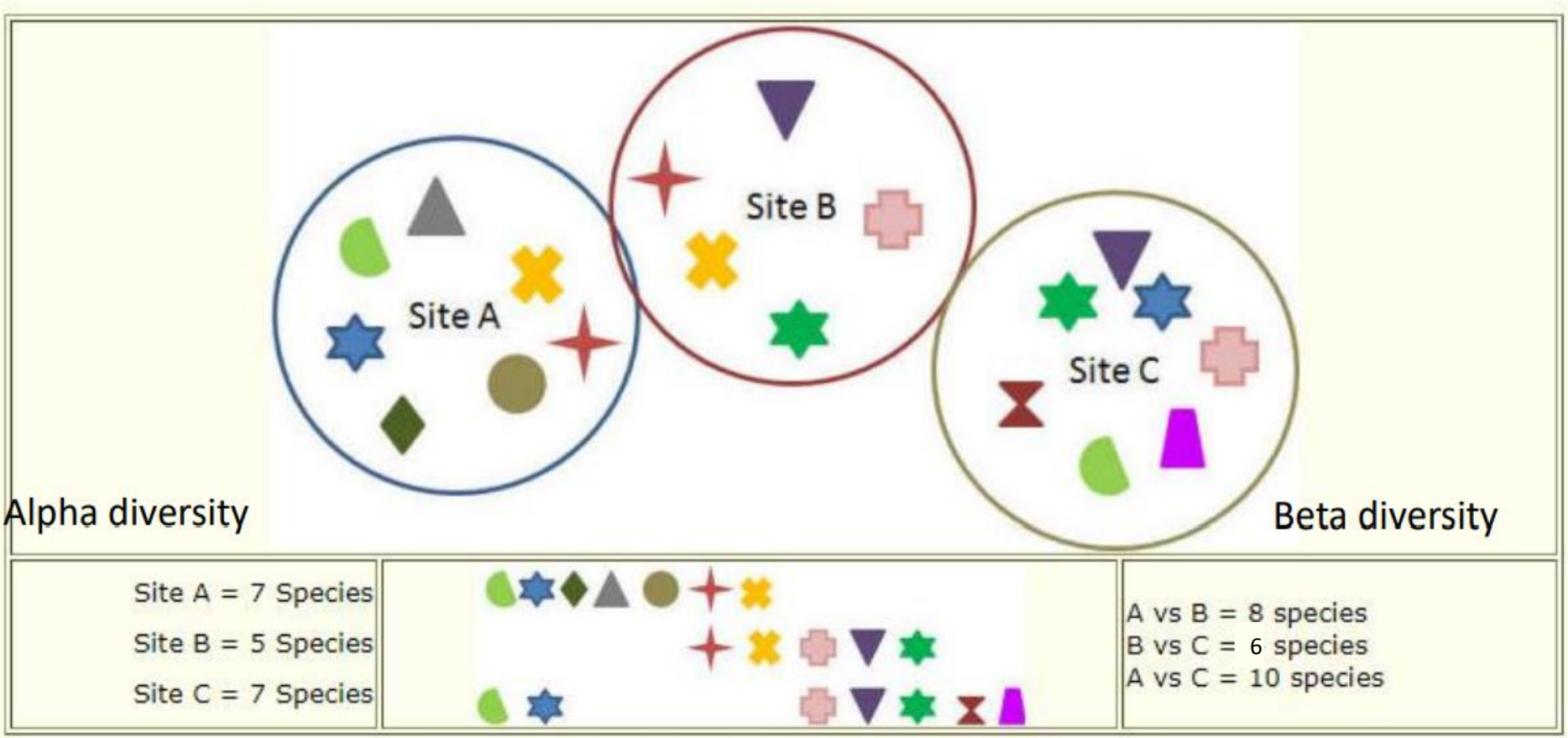


Point diversity (Alpha) = within community diversity
Between community diversity (Beta)
Total diversity (Gamma) = Area, region or globe

Metagenome Data Analysis: 2. Amplicon Sequencing

4. Diversity Analysis

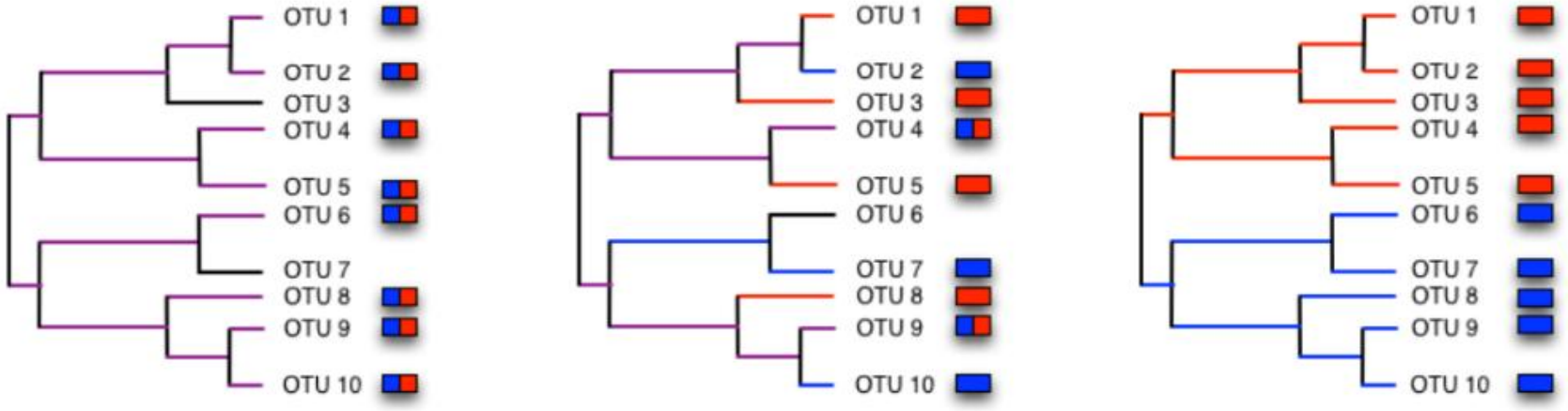
Alpha/Beta/Gamma Diversity



4. Diversity Analysis

Beta diversity comparison

Unweighted UniFrac: a phylogenetic measure of the dissimilarity of microbial communities



U = 0.0
Identical communities

U ≈ 0.5
Related communities

U = 1.0
Unrelated communities

Percent of observed branch length that is unique to either sample

Metagenome Data Analysis: 2. Amplicon Sequencing

4. Diversity Analysis

Beta diversity comparison

Unweighted UniFrac: a phylogeneMc measure of the dissimilarity of microbial communities

visually with ordination plots (e.g., PCoA, NMDS)

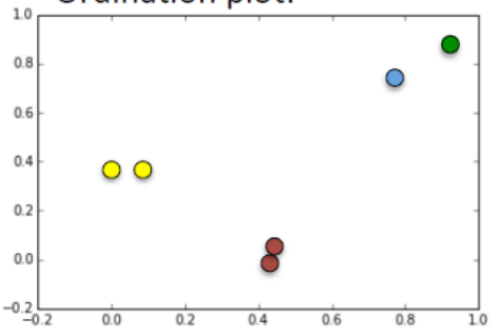
statistically (e.g., PERMANOVA, ANOSIM*)

Sample ID	Sample Type
A	Plant (yellow)
B	Plant (yellow)
C	Turtle (red)
D	Turtle (red)
E	Human (green)
F	Dog (blue)

Unweighted UniFrac distance matrix:

	A	B	C	D	E	F
A	0.00	0.35	0.83	0.83	0.90	0.90
B	0.35	0.00	0.86	0.85	0.92	0.91
C	0.83	0.86	0.00	0.25	0.88	0.87
D	0.83	0.85	0.25	0.00	0.88	0.88
E	0.90	0.92	0.88	0.88	0.00	0.50
F	0.90	0.91	0.87	0.88	0.50	0.00

Ordination plot:



List of all diversity index provided by QIIME2

https://forum.qiime2.org/t/alpha-and-beta-diversity-explanations-and-commands/2282

4. Diversity Analysis

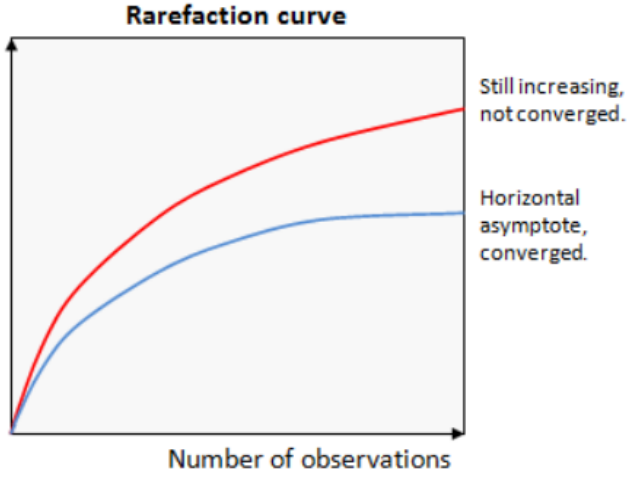
Rarefaction Analysis

- Numerical ecology technique that is often applied to OTU analysis.

- The goal is determine whether sufficient observations have been made to get a reasonable estimate of a quantity (species richness) that has been measured by sampling.

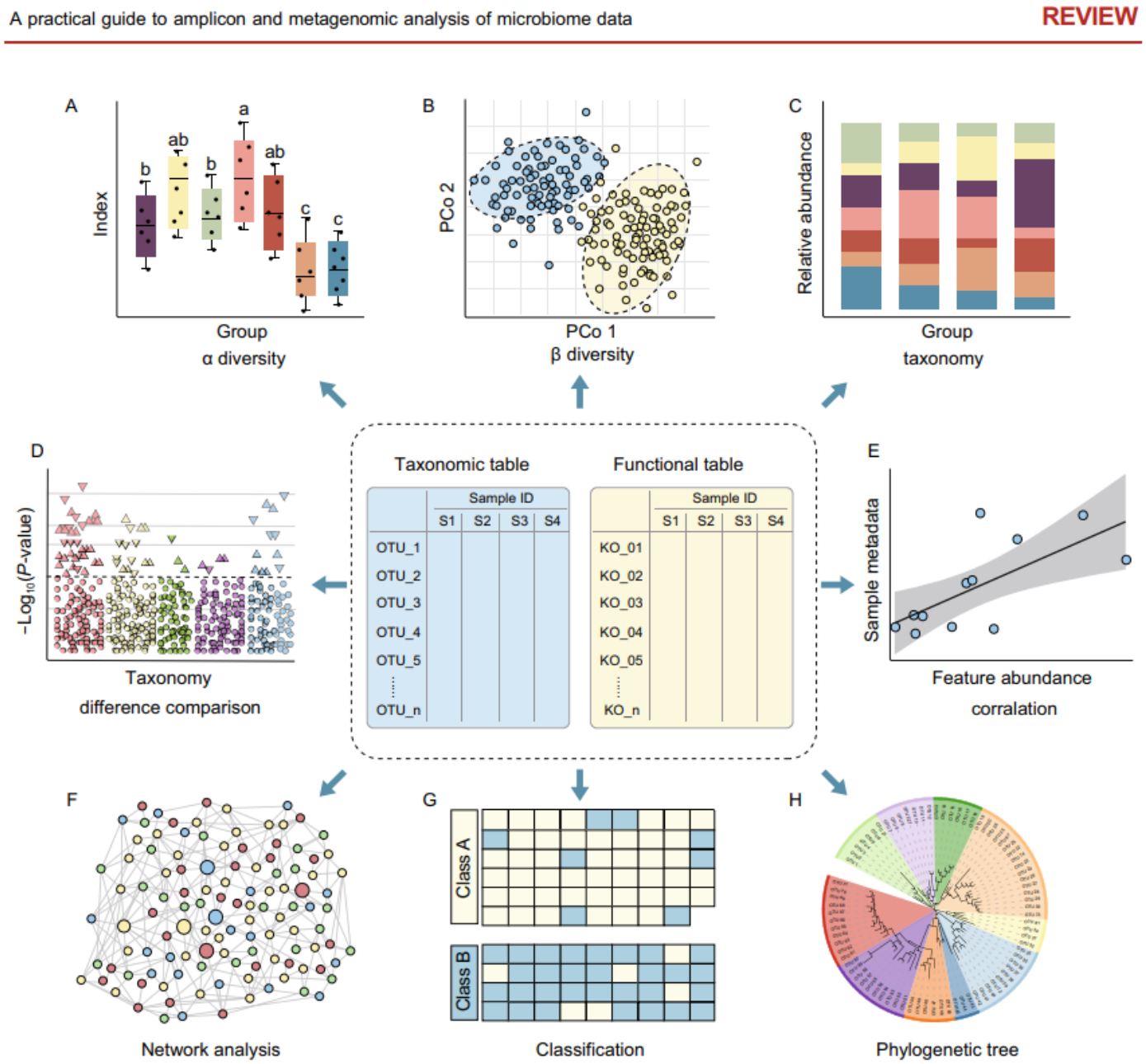
- Rarefaction curves plot the value of a species richness against the number of sampling

- Rarefaction curves generally grow rapidly at first, as the most common species are found, but the curves as only the rarest species remain to be sampled.



Rarefaction ➔ data from alpha diversity

Metagenome Data Analysis: 2. Amplicon Sequencing



End