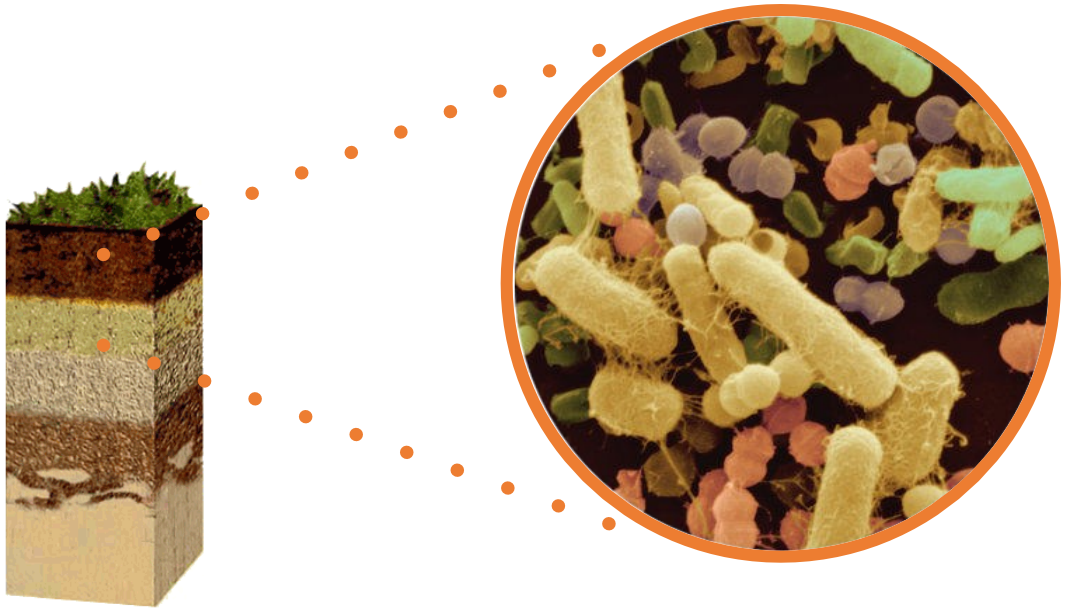
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
SUS-MIRRI.IT

DNA analysis, targeted and untargeted methods in metagenomics

Manuela Costanzo & Annamaria Bevivino
ENEA SSPT-BIOAG-SOQUAS

November 27, 2023
Hybrid course
c/o ENEA CR Casaccia
Sala Blu, Via Anguillarese 301, Roma

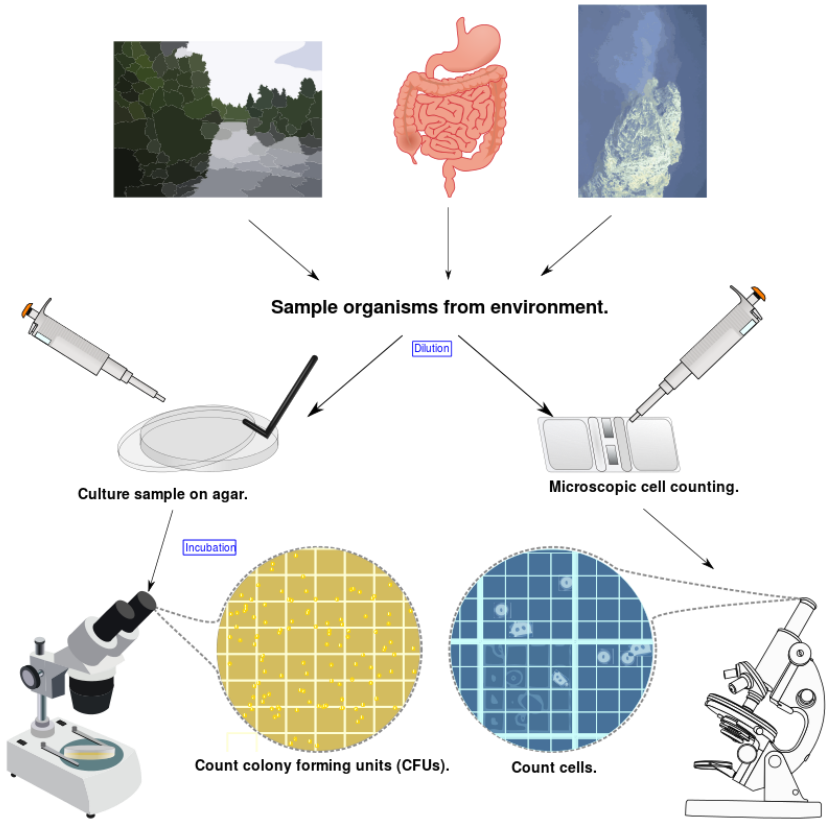
Historical perspective of microbial diversity



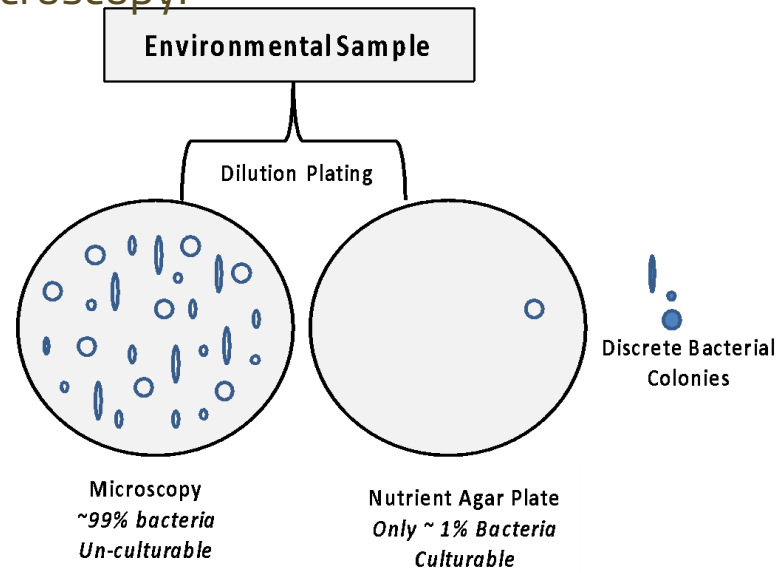
Only 1-10% of all soil bacteria are considered to be culturable

- species may be uncultivable
- some species require special conditions
- species in relatively low density

The great plate count anomaly



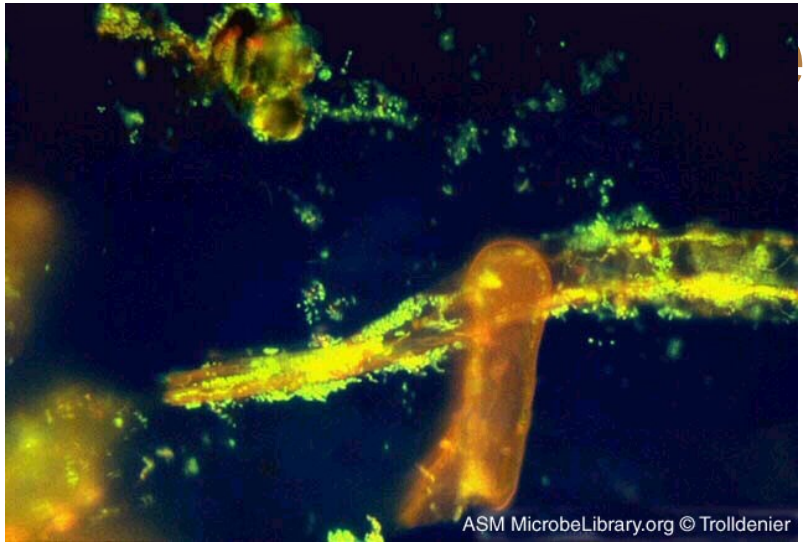
The term "the great plate count anomaly" was coined by Staley and Konopka (*Ann. Rev. Microbiol. 39:321-46, 1985)* to describe the difference between the numbers of cells from natural environments that form viable colonies on agar medium and the numbers obtained by microscopy.

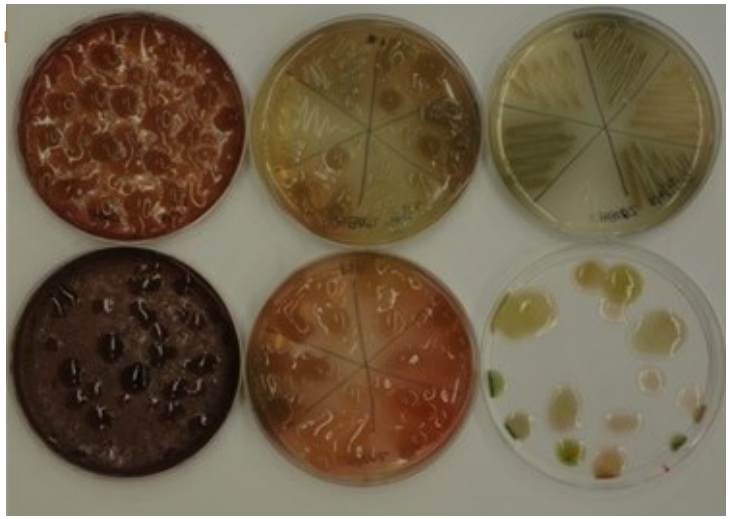


Staley and Konopka (Ann. Rev. Microbiol. 1985, 39:321-46)

How to measure belowground biodiversity?

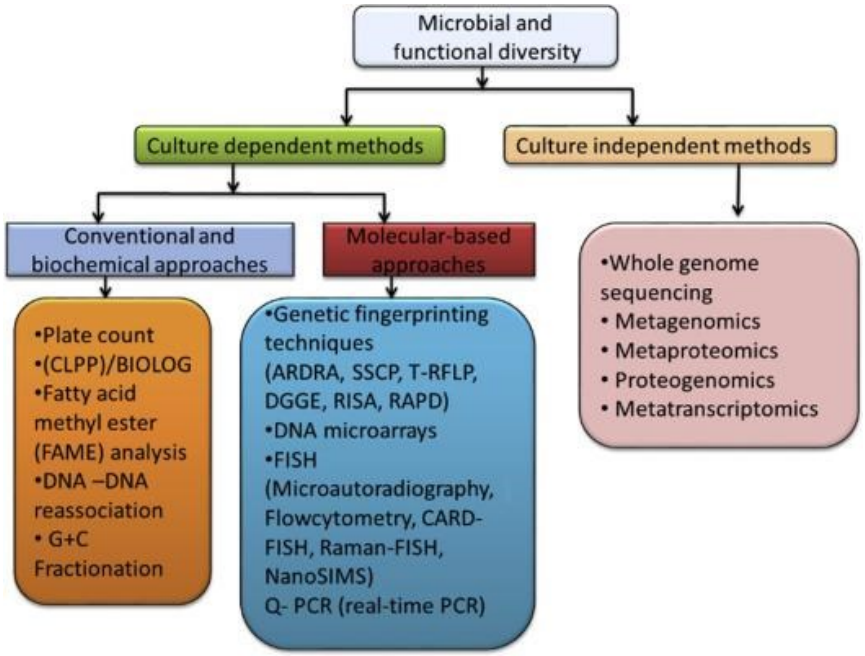
Does it matter if one species disappears?

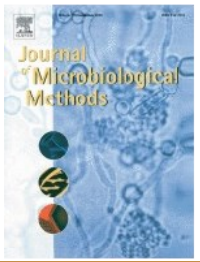




Methods for studying microorganisms

- Sample
  - Biochemical microbiological methods
    - Plate-count method
    - Phenotype Microarrays
    - MALDI-TOF
  - Molecular methods
    - GC content
    - MLST
    - PFGE
    - T-RFLP
    - DGGE
    - DNA sequence analysis
      - Genomics
      - Metagenomics





Kirk JL, Beaudette LA, Hart M, et al. Methods of studying soil microbial diversity. *J Microbiol Methods*. 2004;58(2):169-188.

Methods to study the belowground microbial communities

Mercado-Blanco J, Abrantes I, Barra Caracciolo A, Bevivino A*, Ciancio A, Grenni P, Hrynkiewicz K, Kredics L, Proença DN. Belowground Microbiota and the Health of Tree Crops. Front Microbiol. 2018 Jun 5;9:1006.

Molecular methods

Mercado-Blanco J, Abrantes I, Barra Caracciolo A, Bevivino A*, Ciancio A, Grenni P, Hrynkiewicz K, Kredics L, Proença DN. Belowground Microbiota and the Health of Tree Crops. Front Microbiol. 2018 Jun 5;9:1006.

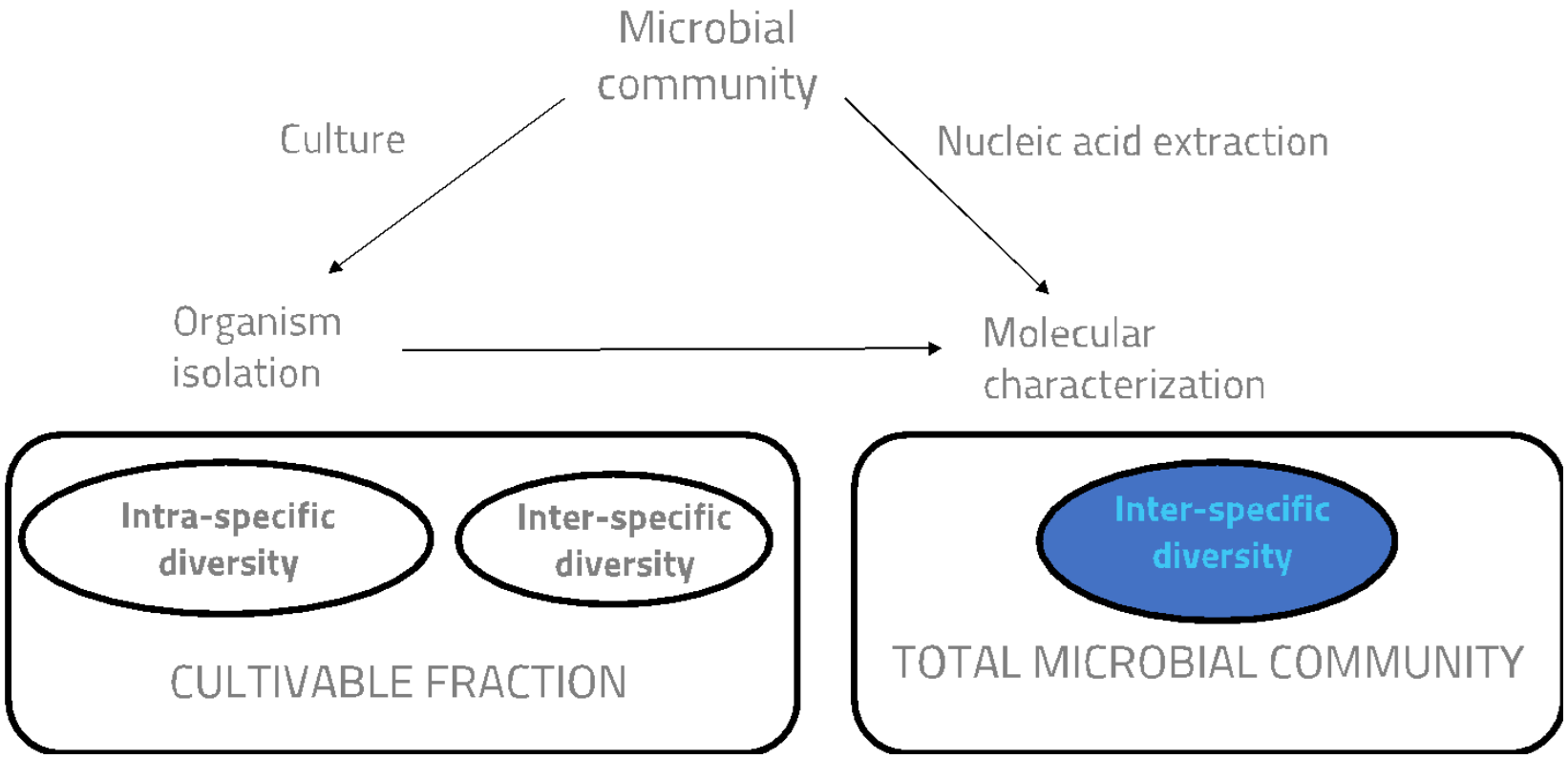
Molecular methods

TABLE 1 | Continued

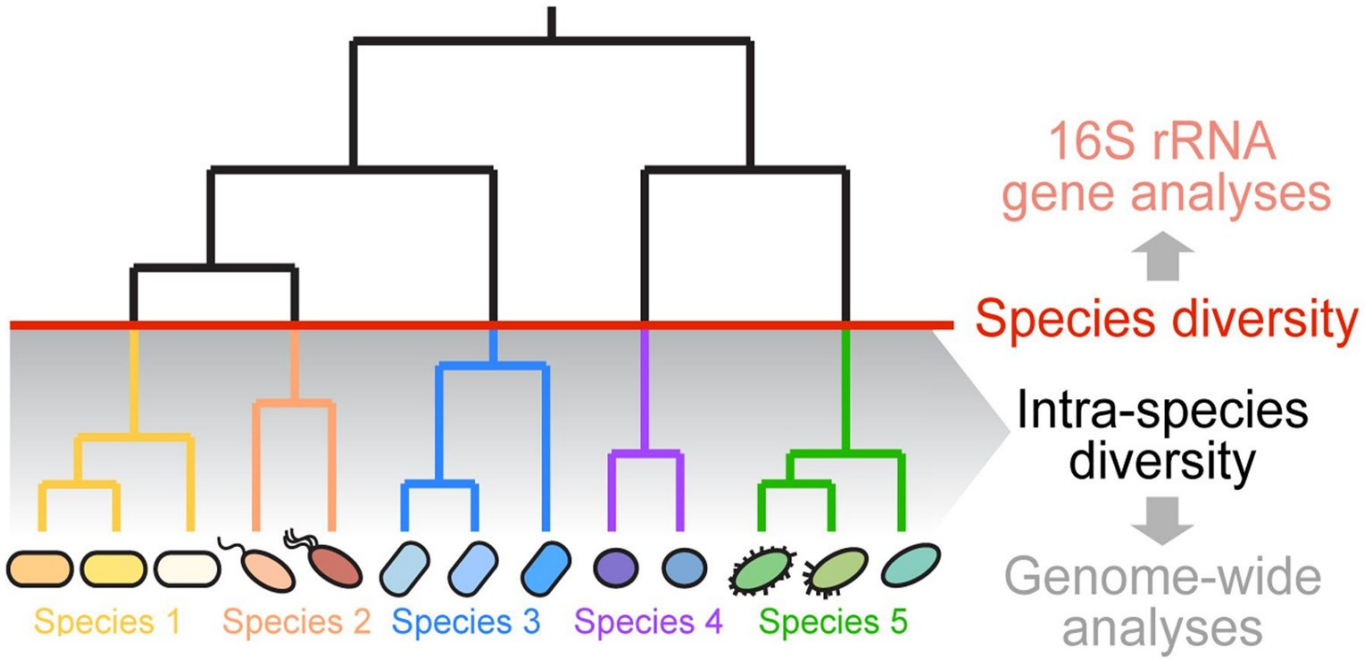
Methods	Advantages	Disadvantages	Crop examples
Next-generation sequencing (NGS)	Rapid to assess biodiversity and abundance of many species/organizational taxonomic units simultaneously	Massive amount of sequencing data of DNA (genomic or PCR amplified fragments) or RNA error distribution within reads of a library; insertions or substitution errors; relatively expensive; replication and statistical analysis are essential; computational intensive; challenging in terms of data analysis	Maritime pine (*Pinus pinaster*)
DNA sequence analysis of the internal transcribed spacer (ITS) region for mycorrhizal studies	Fast and accurate for the identification of mycorrhizal fungi and characterization of their distribution.	Relatively expensive, especially in case of metagenomic analyses	Ectomycorrhizas of poplar (*Populus nigra* x *maximowiczii*) and willow clone (*Salix viminalis*) cultivated as SRF, mycorrhizal fungi of willow (*Salix* spp. L.) from hydrocarbon-contaminated soils, AMF of *Acacia gerrardii* under salt stress

A few illustrative examples of herbaceous plants are also given.

Approaches for assessing diversity



DNA-based methods for biodiversity assessment

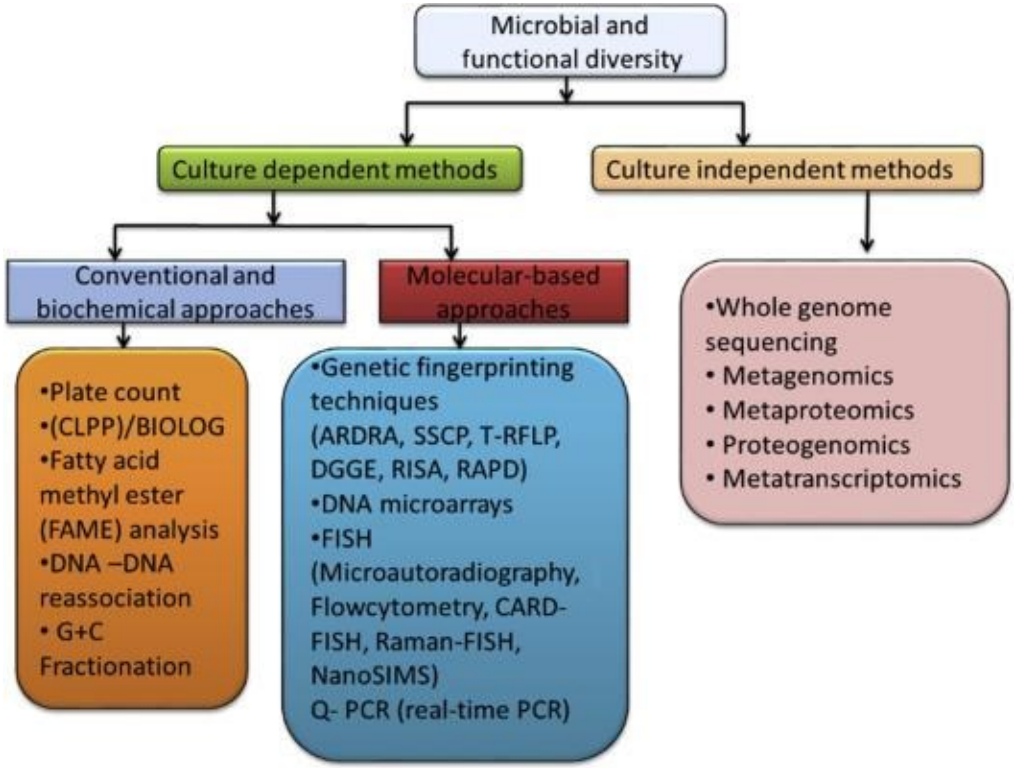


16S rRNA gene analyses

Species diversity

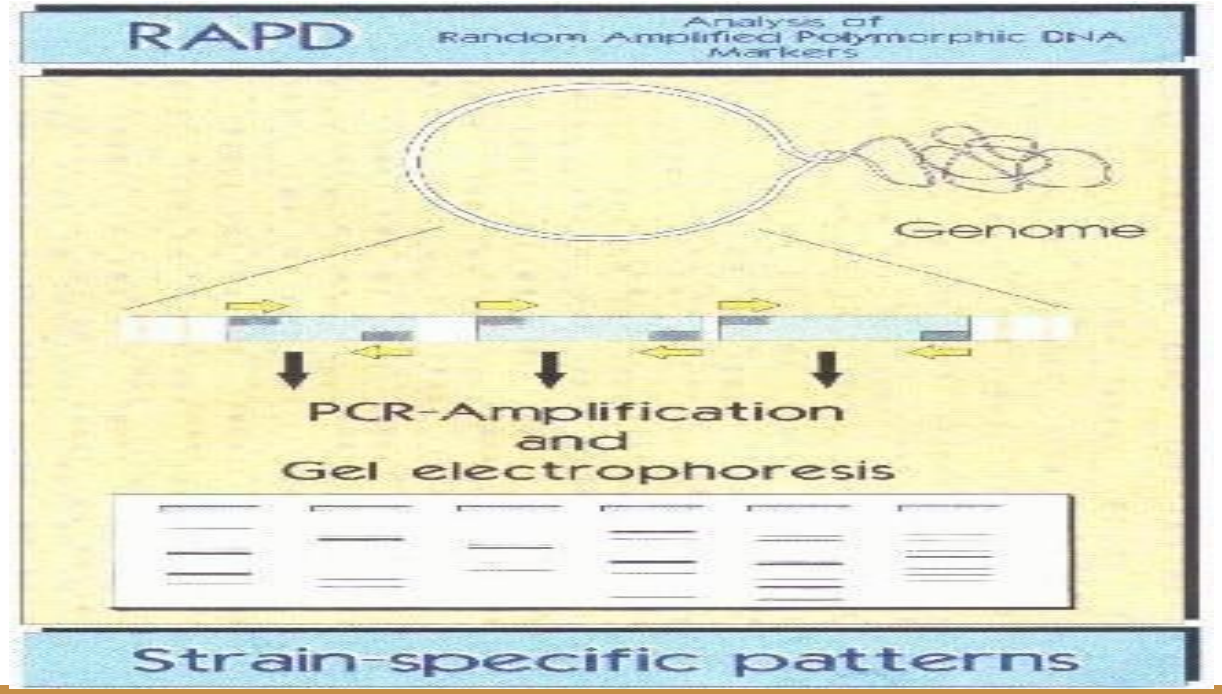
Intra-species diversity

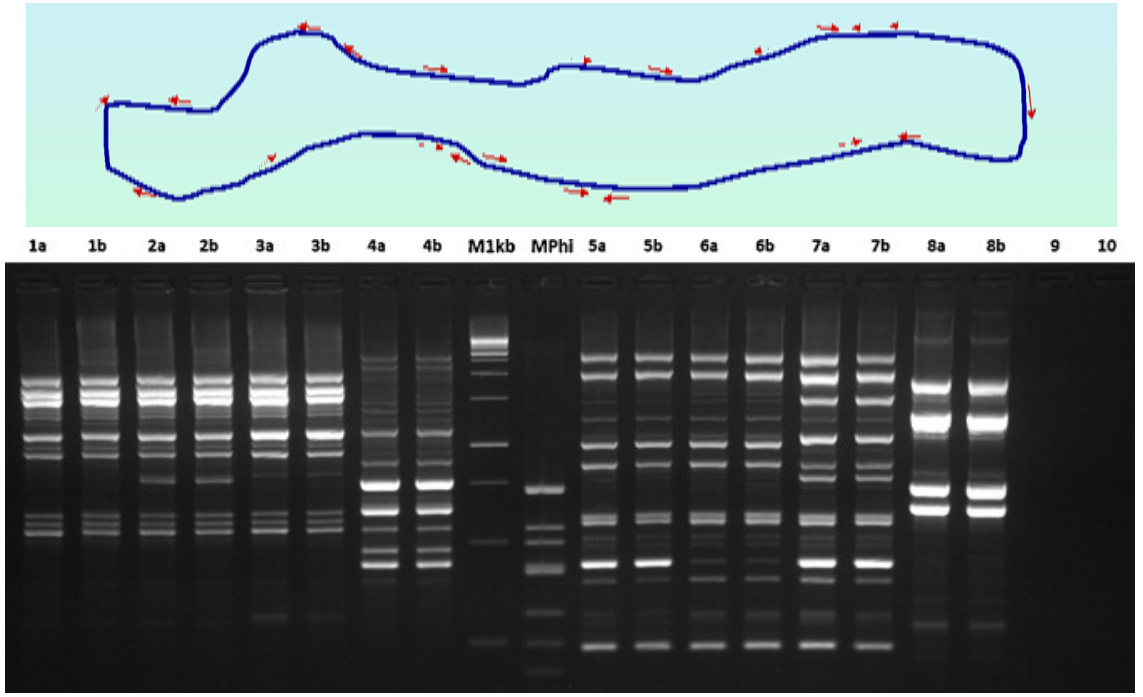
Genome-wide analyses



Typing methods: Intra-species diversity

- Random Amplified Polymorphic DNA (RAPD) is a PCR-based technology for identification genetic variation
- It is particularly suitable for study of genetic variation at population level
- This procedure detects nucleotide sequence polymorphism in DNA

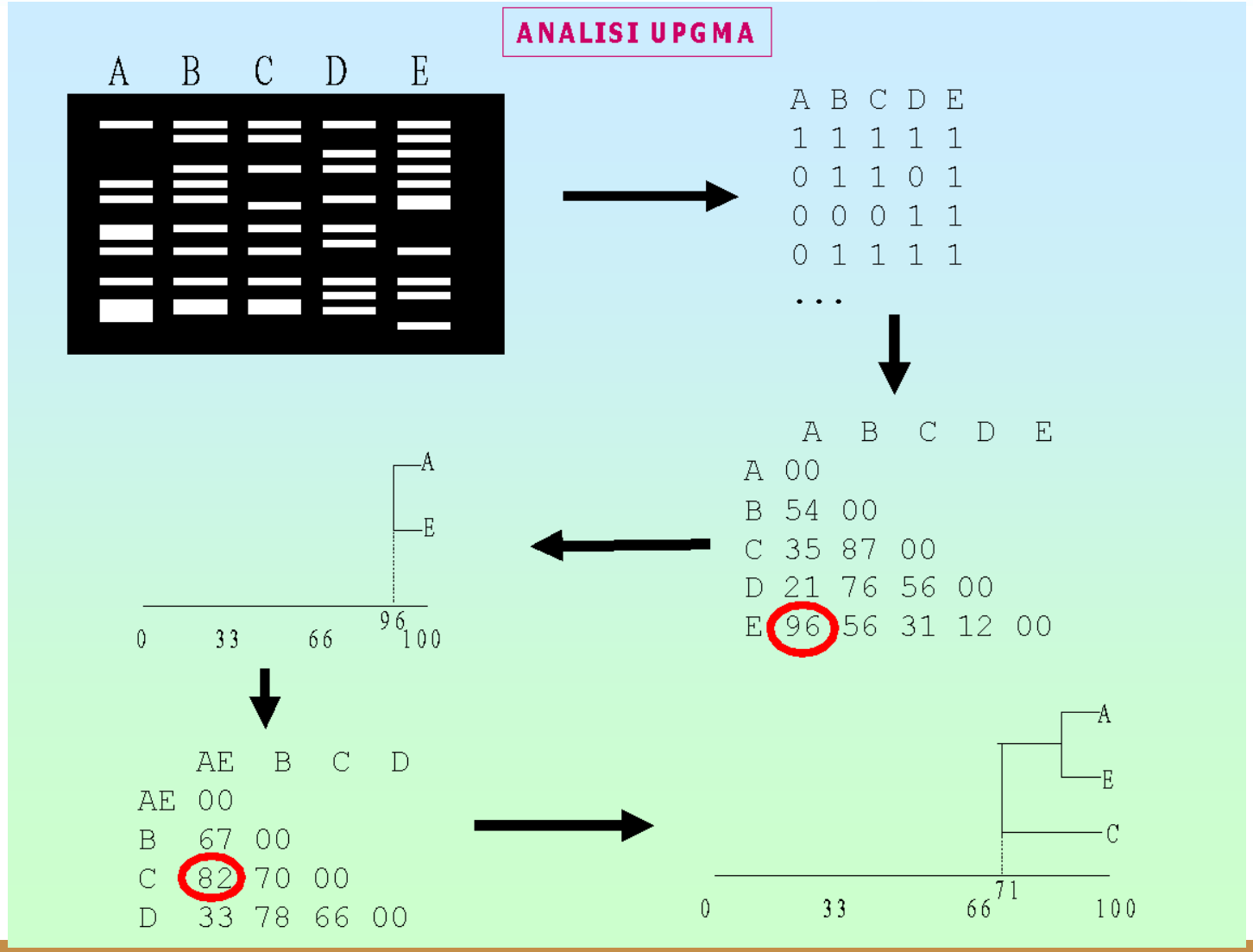




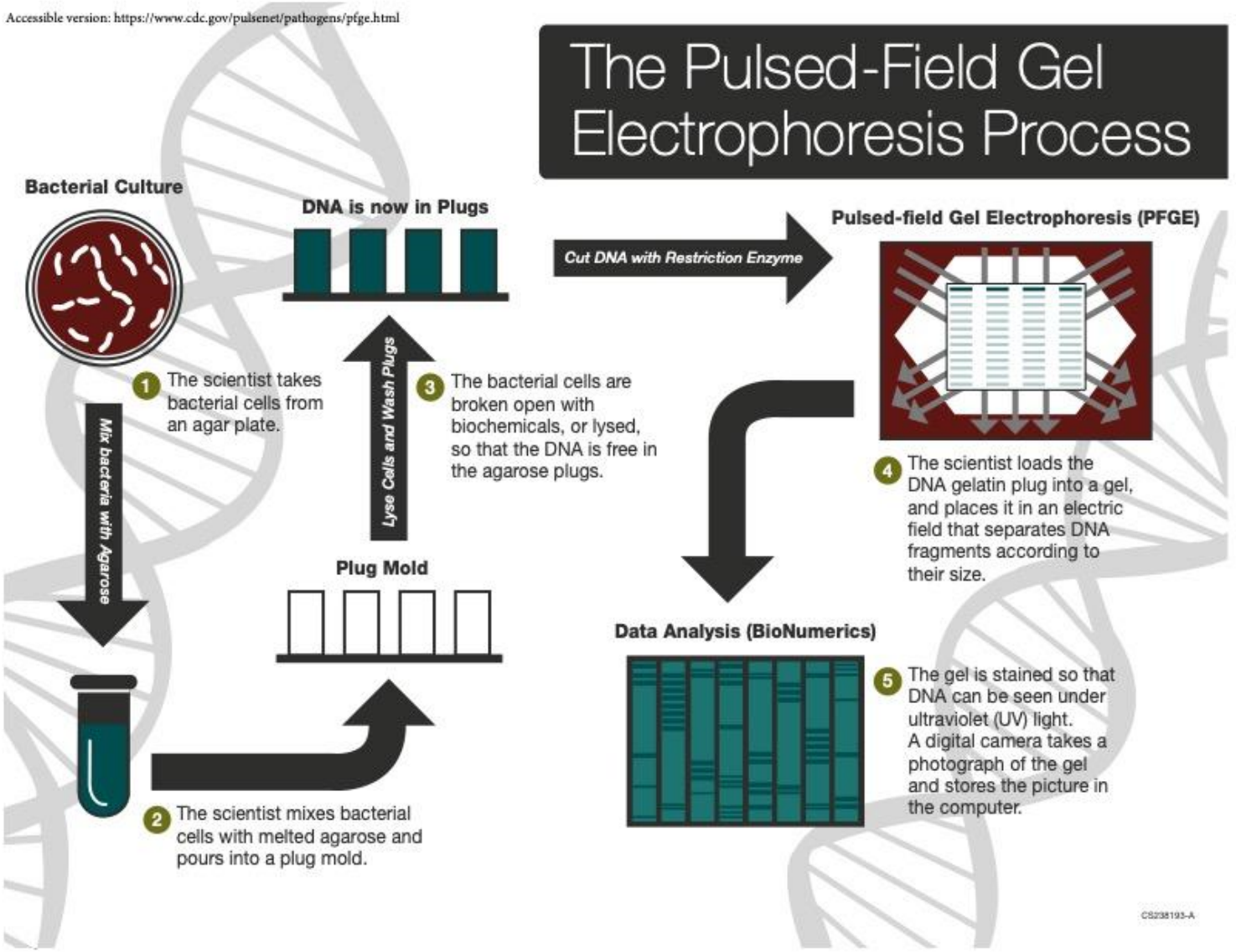
The Similarity index (SI) values between the RAPD profile of any two individuals were calculated using the Nei genetic similarity index, on the basis of the equation: SI=2Nij/(Ni+NJ)

Nij is the number of common bonds shared between 2 samples I and j
Ni and NJ are the total number of DNA bands for genotypes I and j, respectively.

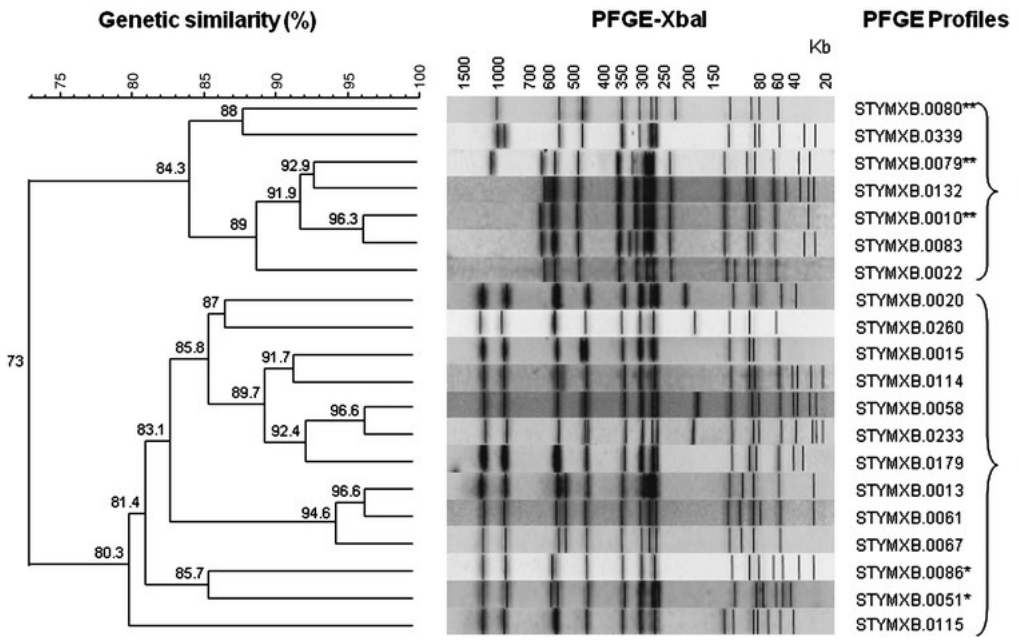
The similarity matrix data were subjected to an unweighted pair group method for arithmetic average (UPGMA) cluster analysis to generate a dendrogram. The results were analyzed based on the principle that a band is considered to be 'polymorphic' if it is present in some individuals and absent in others, and 'Monomorphic' if present in all the individuals or accessions.

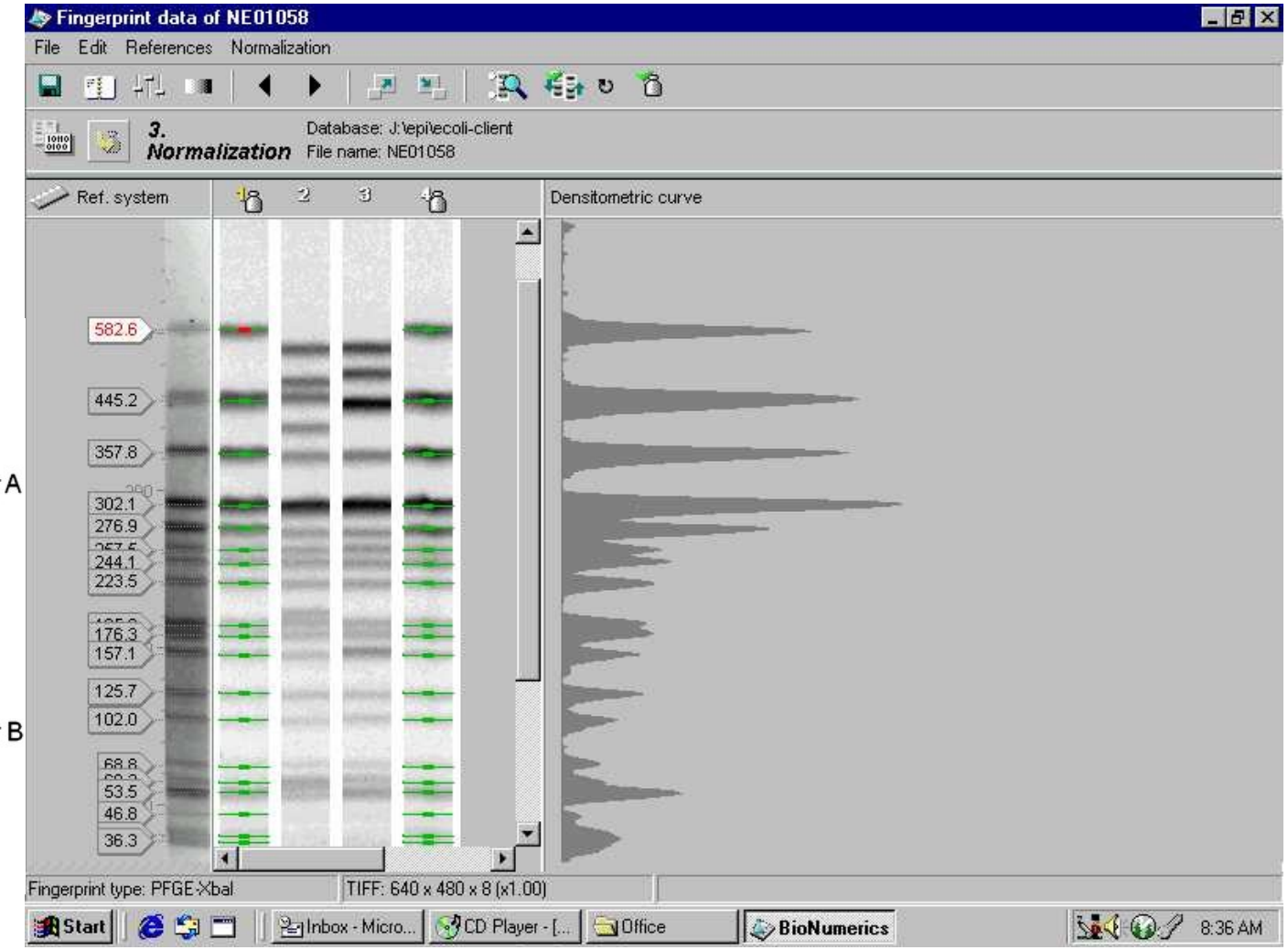


Pulsed-field Gel Electrophoresis (PFGE)



The software used to analyze PFGE patterns normalizes differences in gel conditions to a global reference. This allows us to compare patterns between gels.



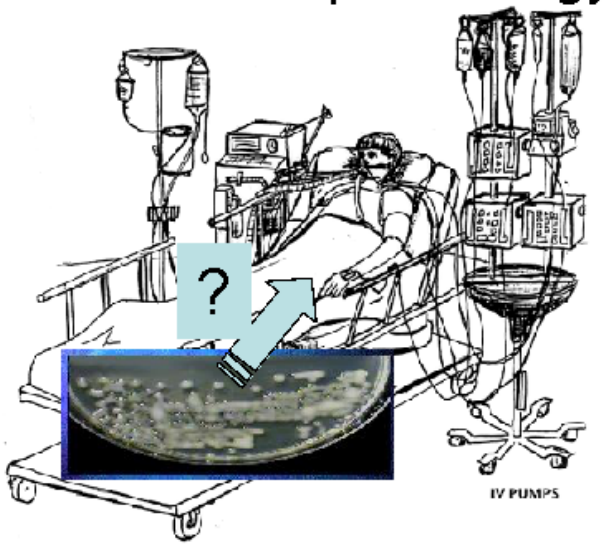


Analysis of strain diversity within species

SUS-MIRRI.IT

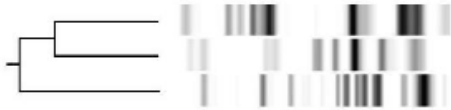
Typing

Aim: Local epidemiology



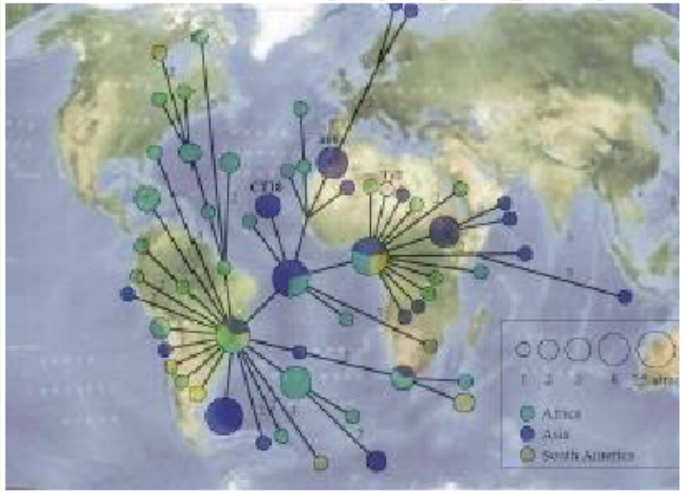
Fingerprinting methods:

- PFGE
- RAPD
- rep-PCR



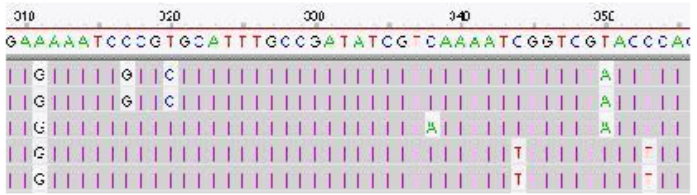
Population genetics

Aim: Diversity, phylogeny, evolution



Multilocus methods:

- MLST
- MLVA-VNTR
- SNPs



Microbial Typing: the pursuit of a common language

MLST
Multi Locus Sequence Typing

Policy Document Saturday 23rd September 2017

DATA ANALYSIS

DATABASES

SUBMISSIONS

NEWS

LINKS

NEW MLST SCHEMES IN DEVELOPMENT

Site requirements

Welcome to the Multi Locus Sequence Typing home page

MLST is a nucleotide sequence based approach for the unambiguous characterisation of isolates of bacteria and other organisms via the internet.

The aim of MLST is to provide a portable, accurate, and highly discriminating typing system that can be used for most bacteria and some other organisms. It is envisaged that this approach will be particularly helpful for the typing of bacterial pathogens.

To achieve this aim we have taken the proven concepts of multilocus enzyme electrophoresis (MLEE) and have adapted them so that alleles at each locus are defined directly, by nucleotide sequencing, rather than indirectly from the electrophoretic moblity of their gene products.

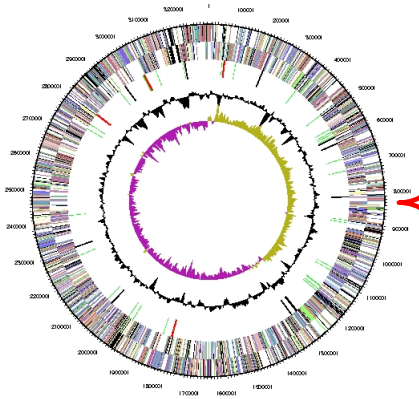
MLST was developed in the laboratories of Martin Maiden, Dominique Caugant, Ian Feavers, Mark Achtman and Brian Spratt.

This site is hosted at **Imperial College** with funding from the **Wellcome Trust**.
The location of the subsites for the individual species are shown on their respective front pages.

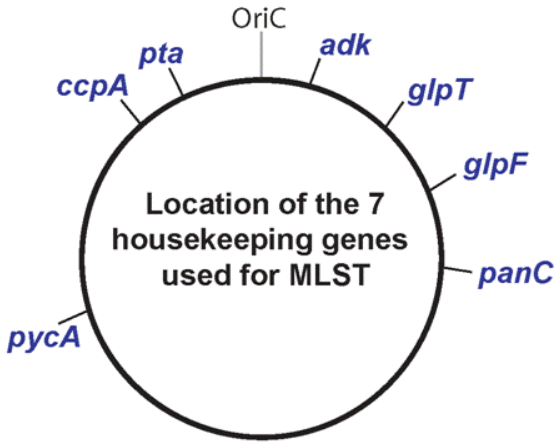
For general information please **Click here** or to register feedback or interest **Click here**

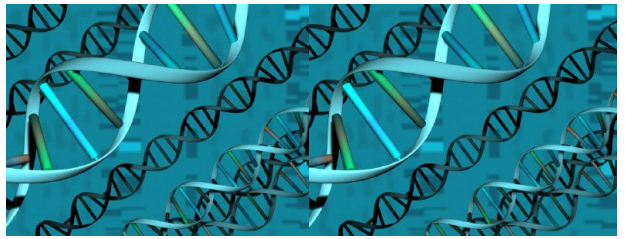
http://www.mlst.net

Multilocus sequence typing (MLST) is a molecular typing technique whereby a number of well chosen housekeeping genes (loci) are sequenced.



genomic DNA

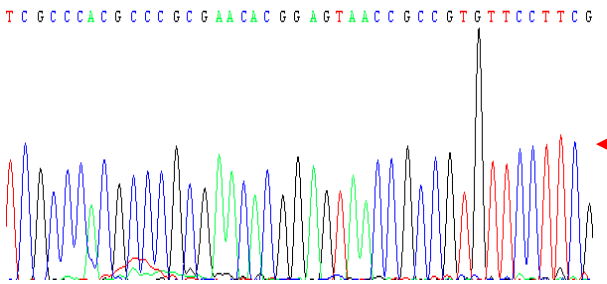




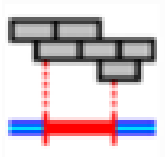
PCR amplification

(718 – 1721 bp)

Purification of PCR product



Sequences (301 – 454 bp)

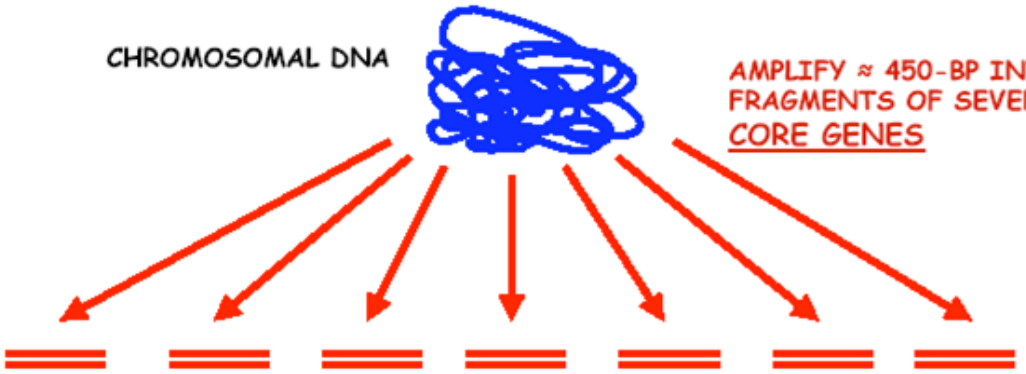


SeqMan

Sequencing analysis and database

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
SUS-MIRRI.IT

Multilocus Sequence Typing (MLST)

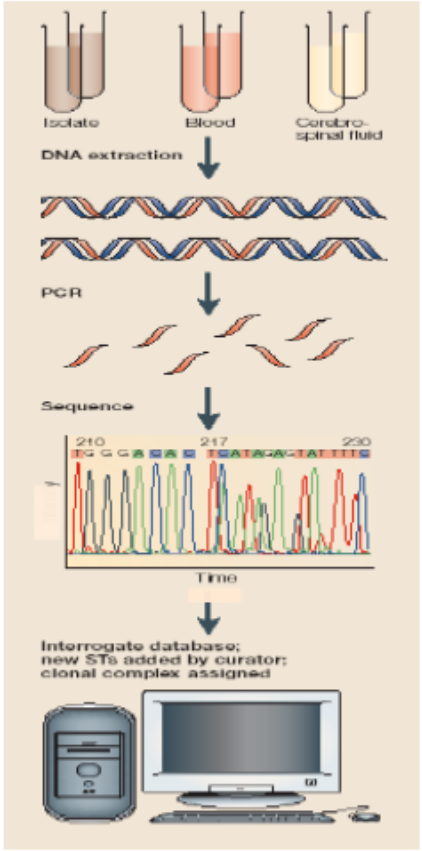


Sequence the seven gene fragments on both strands

Compare the sequences of each gene fragment with the known alleles at the locus held on a database on the Internet

Assign alleles at seven loci to give the Sequence Type (ST)

Compare the ST to those of previously typed isolates

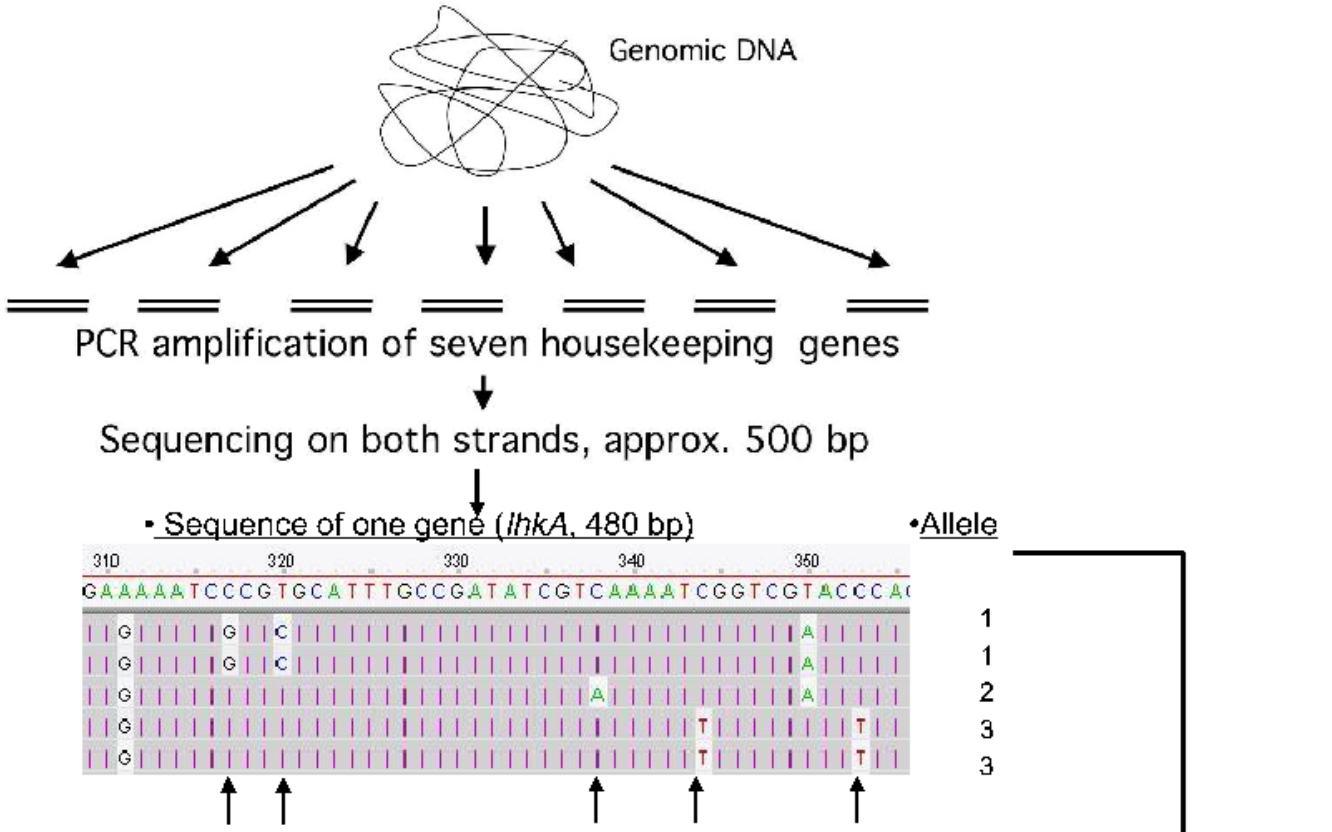


Multi-Locus Sequence Typing (MLST)

	ST	Loci: A	B	C	D	E	F	G	
Isolate 1	1	1	3	1	1	2	4	1	i
Isolate 2	1	1	3	1	1	2	4	1	
Isolate 3	1	1	3	1	1	2	4	1	
Isolate 4	2	1	3	1	1	3	4	1	
Isolate 5	3	2	1	2	3	1	1	1	ii
Isolate 6	4	2	1	2	3	6	1	1	
Isolate 7	3	2	1	2	3	1	1	1	
Isolate 8	3	2	1	2	3	1	1	1	
Isolate 9	5	3	2	1	2	2	2	2	iii
Isolate 10	6	3	4	1	2	2	2	2	

Feil, E.J. *Nat. Rev. Mic.* 2(6):483-95

MLST Multilocus Sequence Typing



Alleles:

abcZ	*bglA*	*cat*	*dapE*	*dat*	*ldh*	*lhkA*
3	1	1	1	3	1	3

Sequence type (ST)	Alleles
1	3 1 1 1 3 1 3
2	1 2 1 7 22 39 7
3	3 1 1 1 3 1 2

Single Locus Variants (SLV)

MLST databases: global epidemiology and population biology

US-MIRRI.IT

pubmlst.org

PubMLST

www.mlst.net

MLST

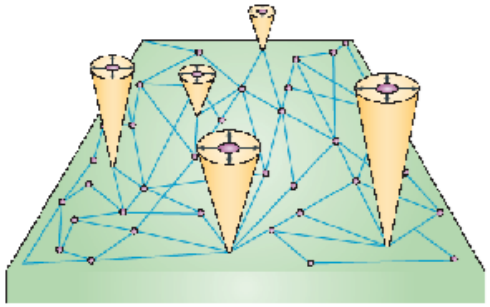
Multi Locus Sequence Typing

Warwick databases

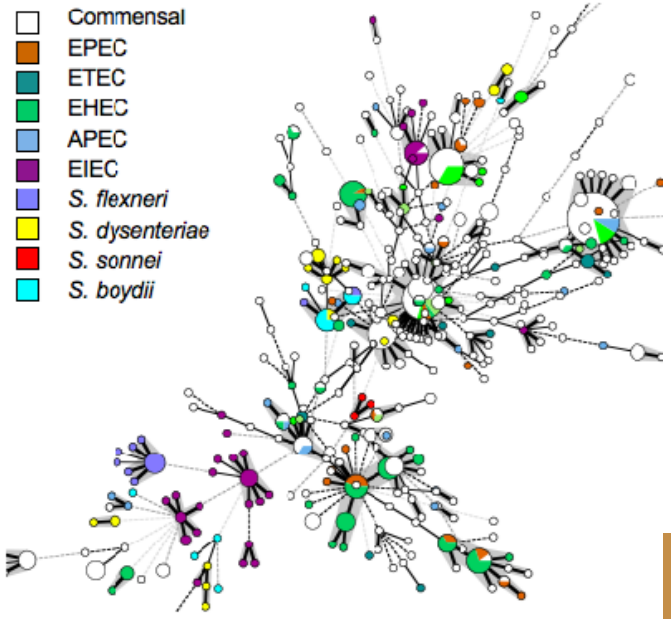
- *Escherichia coli*
- *Moraxella catarrhalis*
- *Salmonella enterica*
- *Yersinia pseudotuberculosis*

www.pasteur.fr/mlst

Institut Pasteur MLST Databases



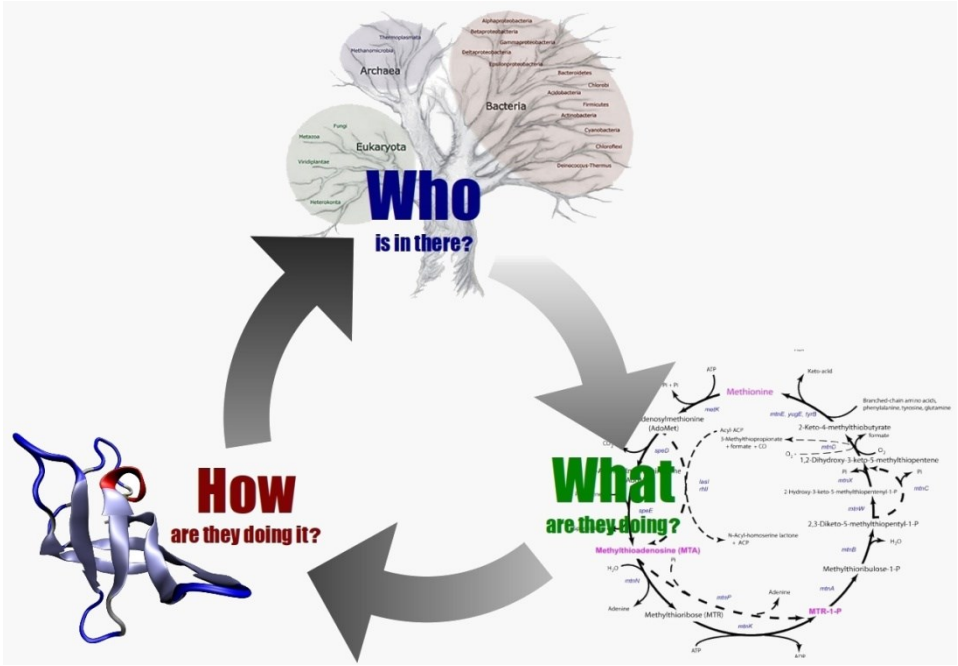
ST	dinB	icdA	pabB	polB	putP	trpA	trpB	uidA
1	1	1	2	1	1	2	3	1
2	8	2	7	3	7	1	4	2
3	3	8	5	11	8	3	5	3
4	2	4	6	4	1	6	1	1
5	5	3	3	10	5	8	2	5
6	1	7	1	9	2	20	1	6
7	6	6	4	2	6	7	2	4
8	23	9	8	12	9	11	7	13
9	9	20	15	7	4	9	6	9



Culture-independent methods

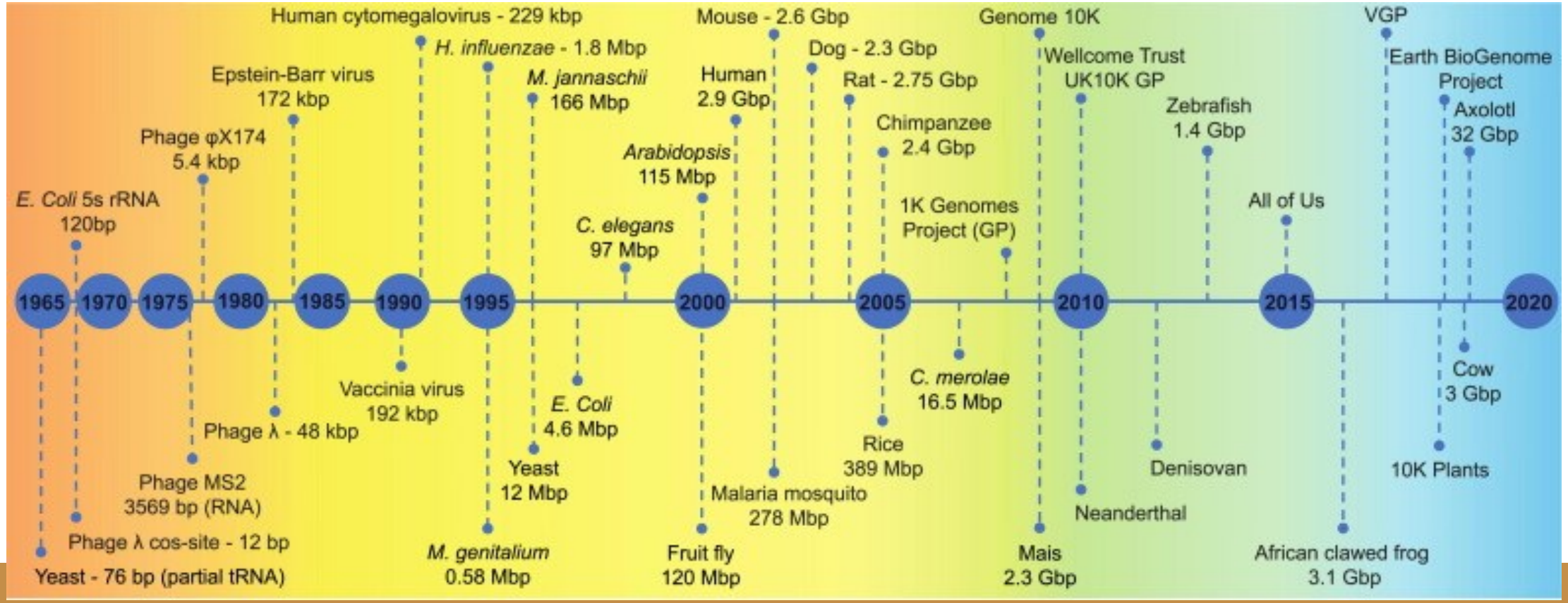
- Access to a much larger reservoir of genomic and metabolic information
- Link community structure and diversity to function

In a few years, the study of uncultured microorganisms has expanded beyond asking "Who is there?" to include the difficult question "What are they doing?"



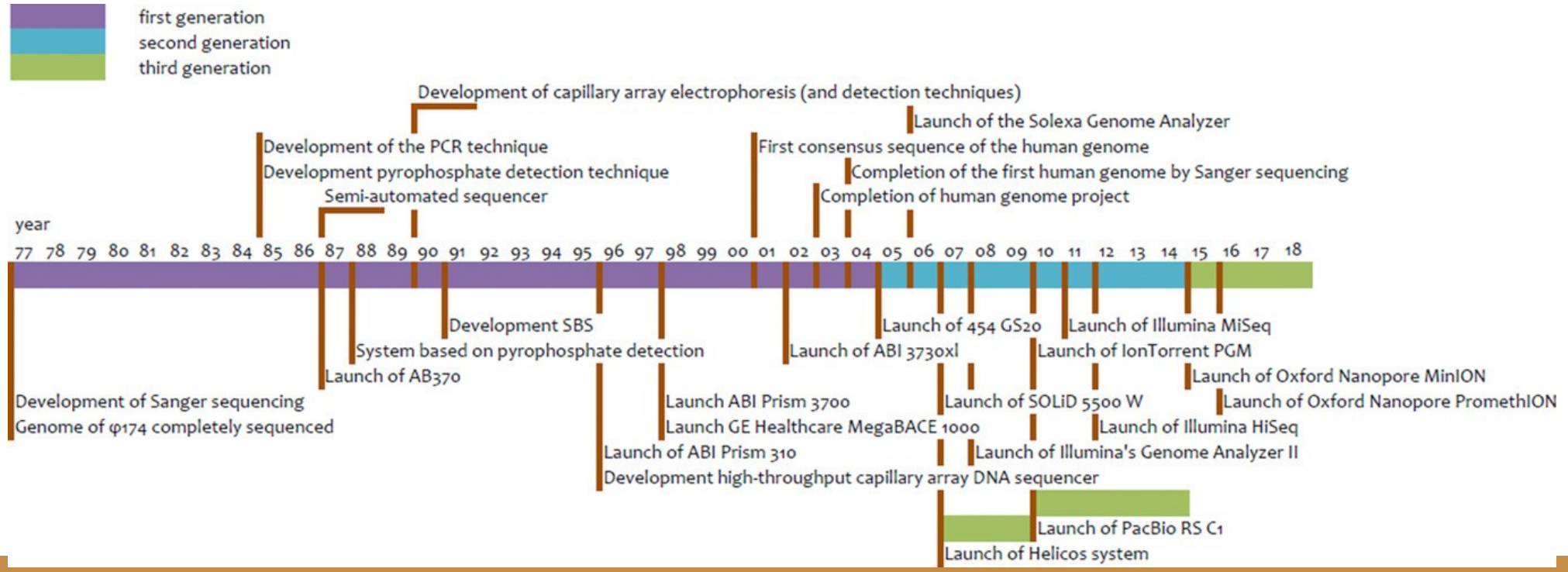
The long walk to genomics

- Increasingly "challenging" projects
- Increasingly difficult and larger targets



The demand for the sequencing market is growing exponentially

- Technologies in this field are evolving very rapidly
- Today's DNA sequencing market was estimated to be approximately $6 BILLION in 2017 and is expected to expand to $25 BILLION by 2025



Cost of services vs cost of technologies

- the cost of sequencing has been reduced in a way that no one could have predicted in the space of a few years, to the point that the goal of the $1000 genome has already been achieved and that of the $100 genome no longer seems like a mirage

- But how much technology is behind a new generation sequencer?

Illumina is currently the leader company in the sector

A NovaSeq sequencer that can generate a human genome in less than 48 hours for less than $1000 has a market cost of around $800,000!

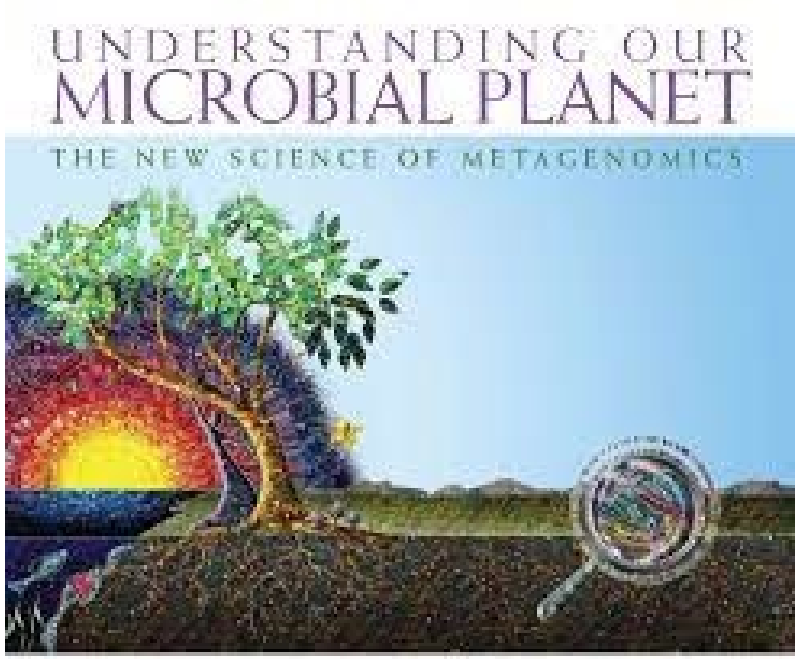
And a large sequencing center looks something like this…

To this we add the costs relating to reagents necessary for sequencing, those necessary for hardware and data storage resources (hundreds of Tb of sequence data) and those of technical personnel...

What is metagenomics?

Metagenomics is the study of microbial communities in their original habitats, which can give a comprehensive insight into the interactions within these communities.

Metagenomics can also help identify individual species within microbial habitats.



The branches of metagenomics

biotechnological field (how can some biochemical functions be exploited in research and industrial fields?)

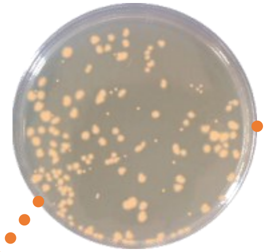
biomedical field (how are alterations in the microbiome linked to the development of diseases?)

ecological context (how does the diversity of microbial communities contribute to the support of food chains?)

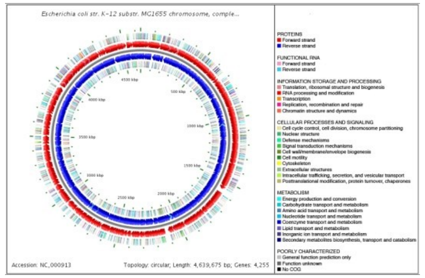
Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA | SUS-MIRRI.IT

Analysis strategies

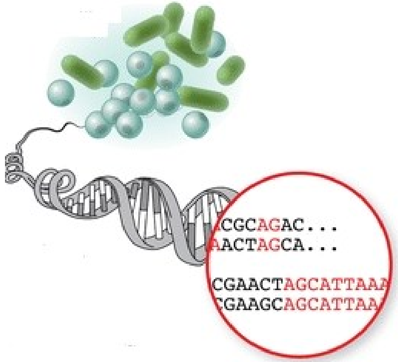
Choosing a significant environment for a given study



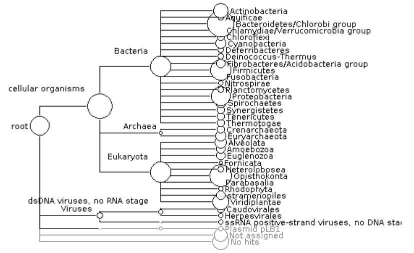
Isolation of a single strain of interest



Genome sequencing and analysis



Analysis of the totality of the extracted DNA



Assembly and analysis of portions of metagenomic DNA

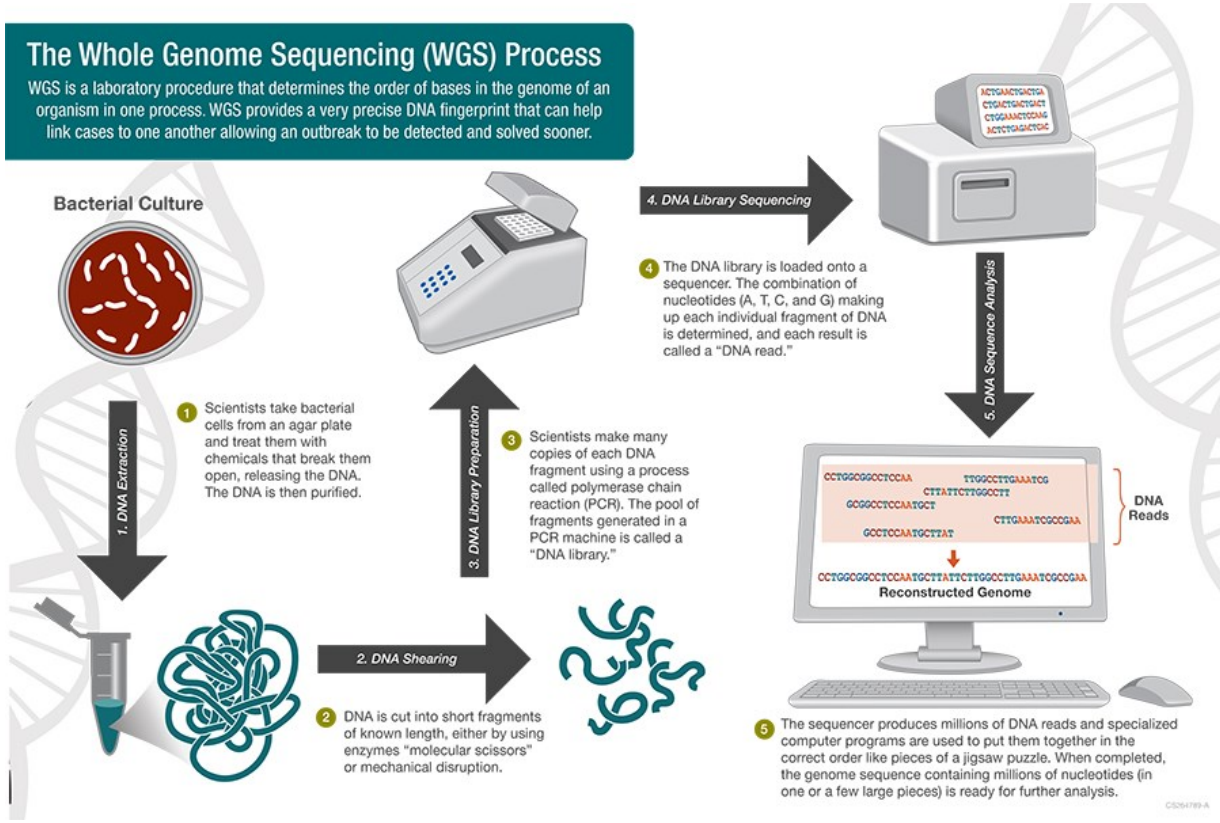
GENOMICS .... **METAGENOMICS**

strong link between the two disciplines

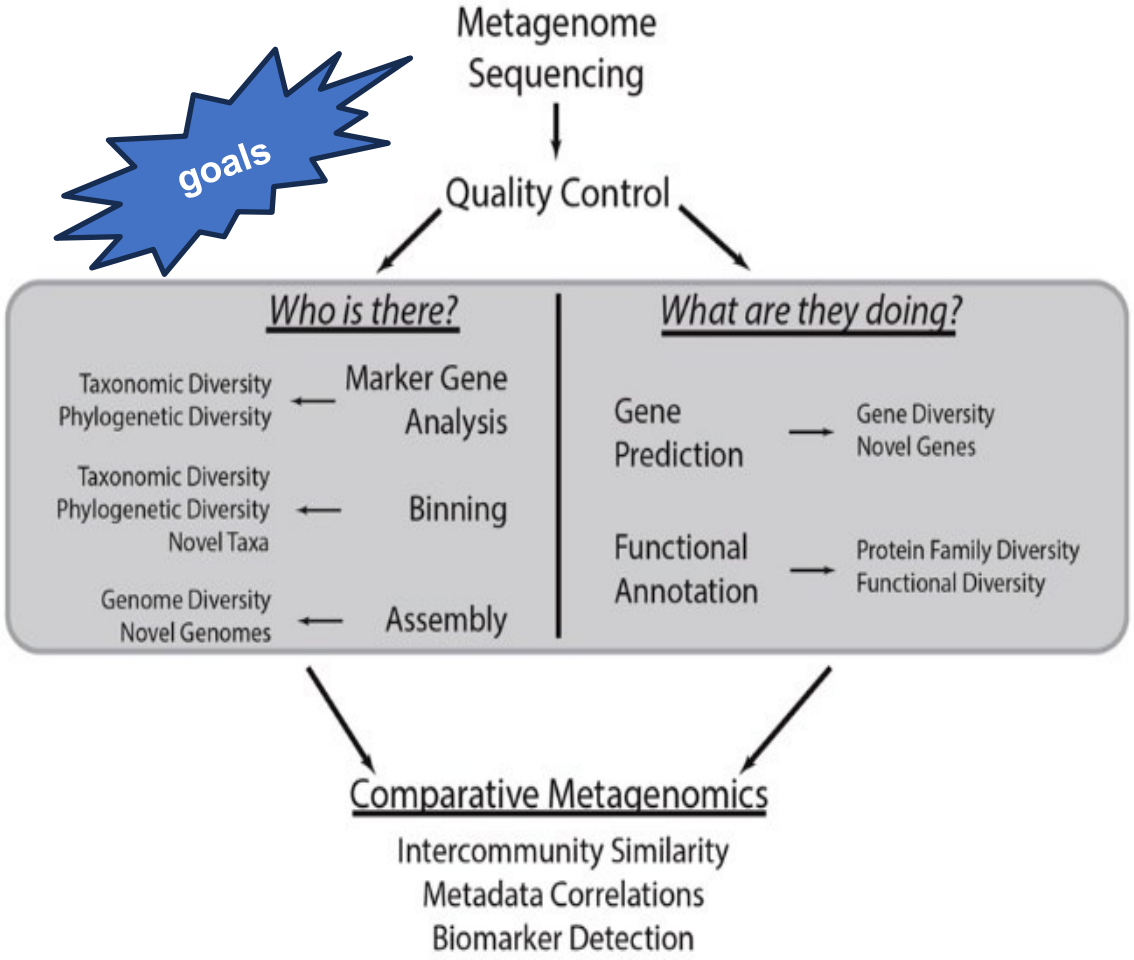
The Whole Genome Sequencing (WGS)

Whole-genome sequencing (WGS) is a comprehensive method for analyzing entire genomes.

While this method is commonly associated with sequencing human genomes, the scalable, flexible nature of next-generation sequencing (NGS) technology makes it equally useful for sequencing any species, such as agriculturally important disease-related microbes



The revolution of Whole metagenome shotgun sequencing



Leave these two classic approaches, moving towards a third approach based on the sequencing of the entire DNA library.

This approach is not focused on the single gene or the single genome, but examines the entire biodiversity of the microbial community.

Phylogenetic diversity between microbial components and also the level of **intraspecific genetic diversity.**

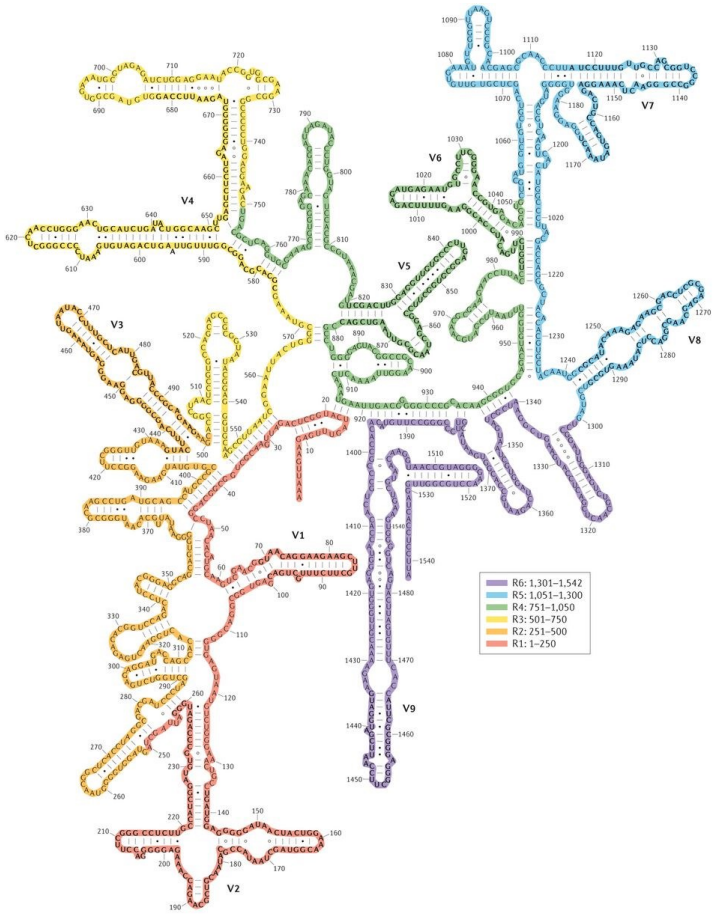
Entire biochemical pathways in a community in detail and to create real "gene collections" represented in the microbiome

Allow the reconstruction of **individual complete genomes** of one or more components

Profiling of a microbial community16S rRNA: why?

- Usually based on single marker genes
- For bacteria, the marker of choice is **16S ribosomal RNA**
- For eukaryotes it is **18S ribosomal RNA**

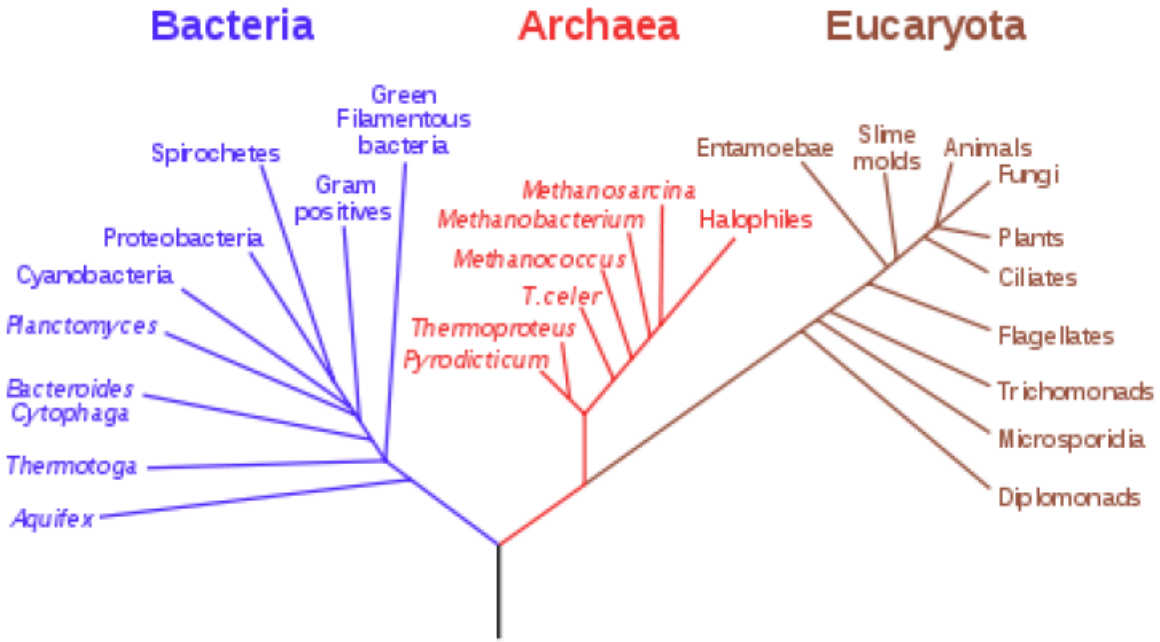
- This is a sequence shared universally by all prokaryotes
- It presents extremely conserved regions, interspersed with highly variable regions (V1-V9)
- These can be amplified and sequenced thanks to the use of degenerate primers designed on their flanking regions
- This bypasses the need to perform a culture



Nature Reviews | Microbiology

16S rRNA: why?

- The variable regions of the 16S rRNA can be considered as "fingerprints" that allow a specific species to be uniquely identified
- There are specific databases such as the Ribosomal Database Project, which allow comparison with over 3 million and 300 thousand already determined sequences of bacteria and Archea



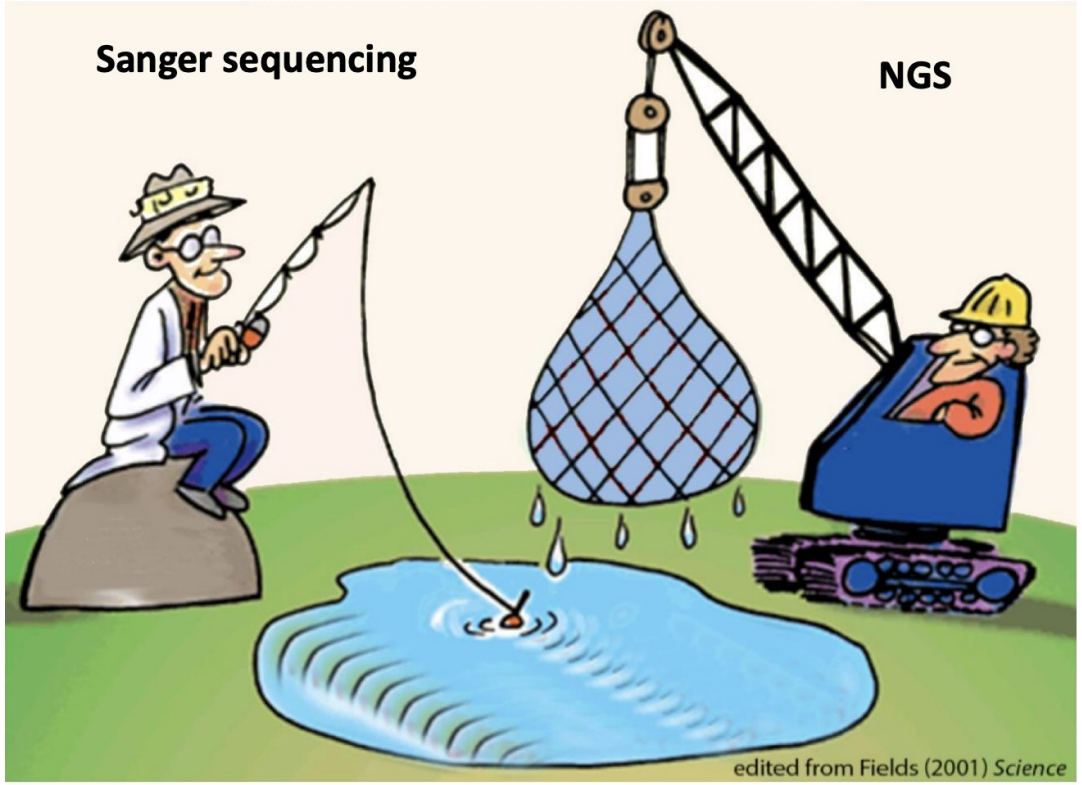
Given the sharing of rRNA sequences by all cellular organisms, the phylogeny based on 16S sequencing has made it possible to redefine the tree of life based on genetic-molecular criteria and no longer just morphological ones

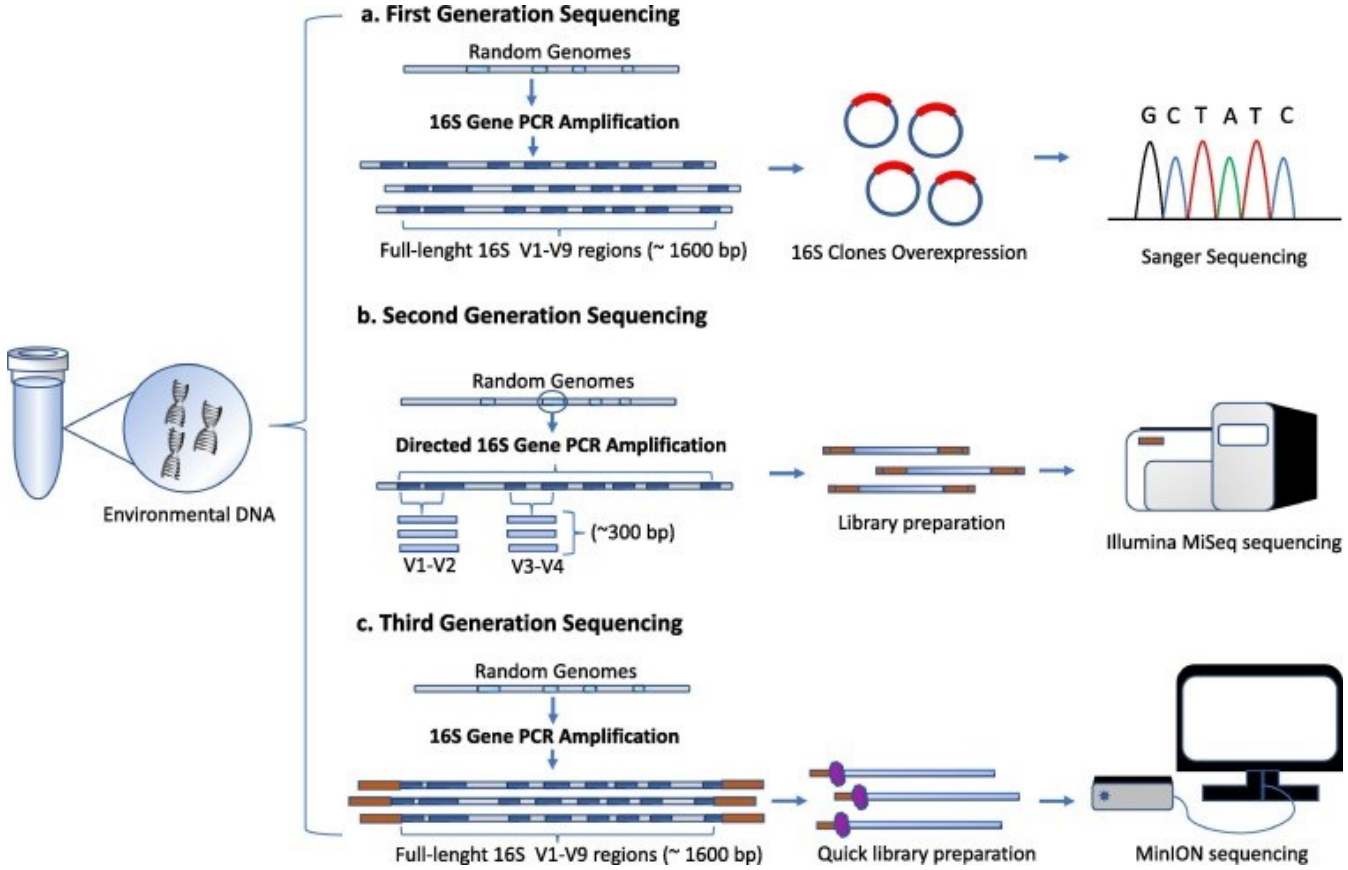
16S rRNA: how?

- The 9 variable regions are characterized by a different length, but also by a variable degree of diversity
- Their phylogenetic informativeness is therefore different and the length also determines the choice of the type of sequencing to be carried out (for example taking into account the length of the reads)

Region	Position	# b.p.
V1	69-99	30
V2	137-242	105
V3	338-533	195
V4	576-682	106
V5	822-879	57
V6	967-1046	79
V7	1117-1173	56
V8	1243-1294	51
V9	1435-1465	30

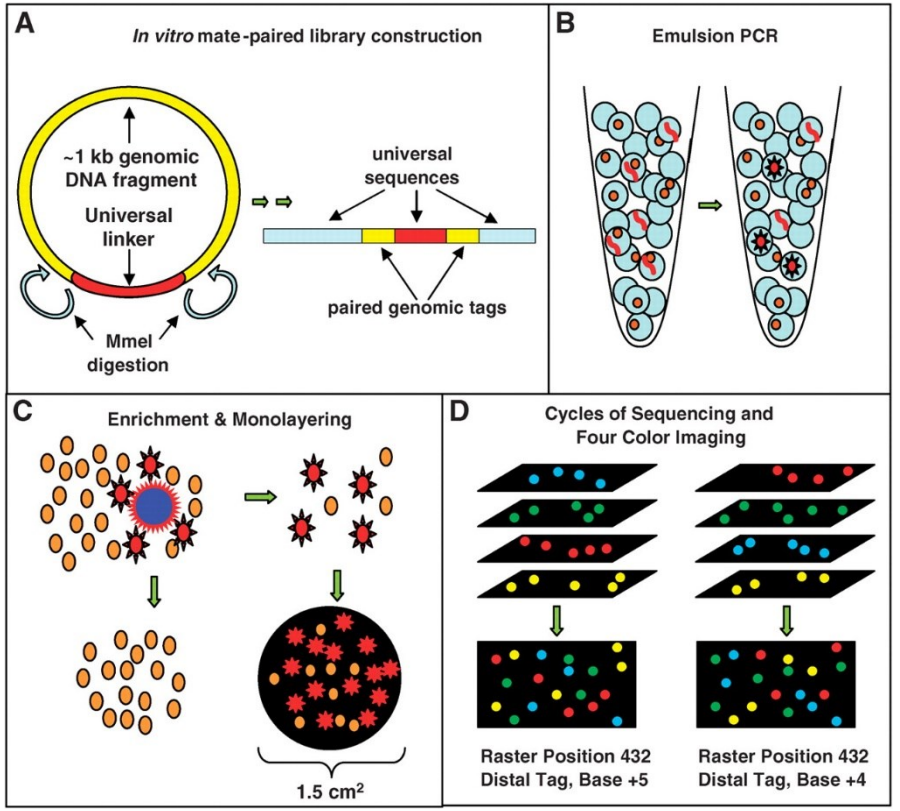
16S rRNA metabarcoding: the different possible approaches





Why do we talk about «next-generation sequencing»?

- The real revolution is given by the very high throughput
- Each sequencing run can generate hundreds of millions of reads (now billions) of limited length
- The reading takes place thanks to the presence of a high resolution optical reader, which records the light signal obtained for each polony, which is physically associated with very precise coordinates of the solid support
- The times are very short and the total cost is enormously lower (cost understood as cost per sequenced base)
- Several technologies have been developed over the years… Illumina, IonTorrent (LifeTechnologies), 454 (Roche), SOLiD (Applied Biosystems), etc







	iSeq 100	MiniSeq	MiSeq Series	NextSeq 550 Series	NextSeq 1000 & 2000
Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)					
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	●	●
Exome & Large Panel Sequencing (enrichment-based)				●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)				●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)				●	●
Targeted Gene Expression Profiling	●	●	●	●	●
miRNA & Small RNA Analysis	●	●	●	●	●
DNA-Protein Interaction Analysis (ChIP-Seq)			●	●	●
Methylation Sequencing				●	●
16S Metagenomic Sequencing		●	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)				●	●
Cell-Free Sequencing & Liquid Biopsy Analysis				●	●
Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11-48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	330 Gb*
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.1 billion*
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

liadomani
NAZIONALE
ESA E RESILIENZA

SUS-MIRRI.IT

An overview of Illumina platforms – different solutions for different objectives

- These are the so-called "benchtop sequencers", i.e. benchtop sequencers designed to meet the needs of small laboratories
- Not suitable for all applications
- However, note the times required to complete a run and output. You can try to compare this potential with the times and costs that were required to sequence the human genome twenty years ago...

An overview of Illumina platforms – different solutions for different objectives



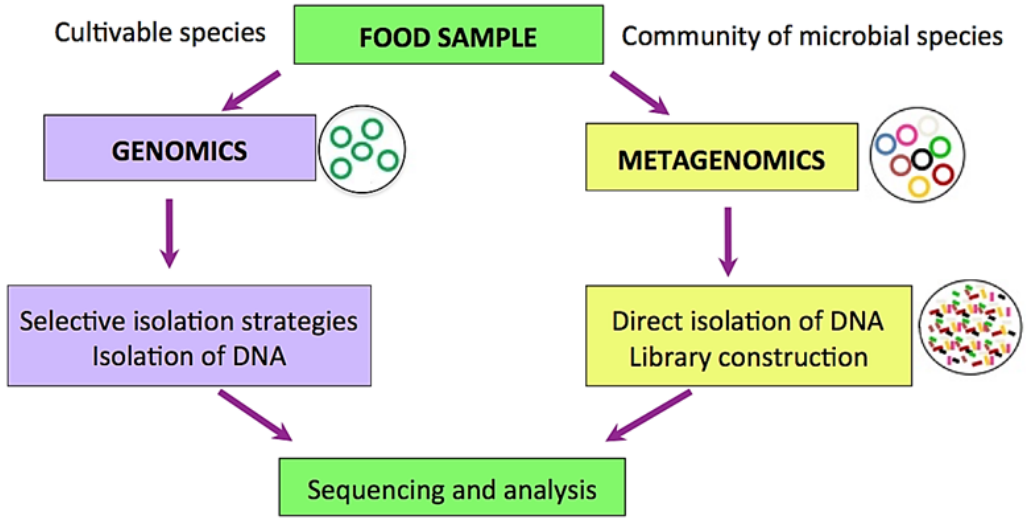
	NextSeq 550 Series	NextSeq 1000 & 2000	NovaSeq 6000
Popular Applications & Methods	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)			●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●
Exome & Large Panel Sequencing (enrichment-based)	●	●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)	●	●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)	●	●	●
Chromatin Analysis (ATAC-Seq, ChIP-Seq)	●	●	●
Methylation Sequencing	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)	●	●	●
Cell-Free Sequencing & Liquid Biopsy Analysis	●	●	●
Run Time	12–30 hours	11-48 hours	~13 - 38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	120 Gb	330 Gb*	6000 Gb
Maximum Reads Per Run	400 million	1.1 billion*	20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 x 250**

- These are the so-called "production-scale" sequencers, designed for sequencing centers
- The type of possible applications is decidedly different from those seen previously
- Sequencing approaches on even large genomes
- Very high processivity

Take home message: Genomics vs. Metagenomics

Genomics advantages:

1. Sequence the genome of one organism at a time
2. Use cultures to isolate microbe of interest



Metagenomic advantages:

1. Extract sequence data from microbial communities as they exist in nature
2. Bypass the need for culture techniques
3. Sequence all DNA in sample
4. Select DNA based on universal sequences

Fig. 2 Schematic representation of the differences between genomics and metagenomics.

✓**The application of multi-omics in food safety and quality has the potential to answer questions traditional microbiological methods could not address.**

✓**Approaching the food ecosystem from different angles allows for a "holistic" representation of which microorganisms are present, how they behave, how they interact and which are the phenotypic manifestations in this complex arena.**

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
SUS-MIRRI.IT

ENEA

manuela.costanzo@enea.it

annamaria.bevivino@enea.it