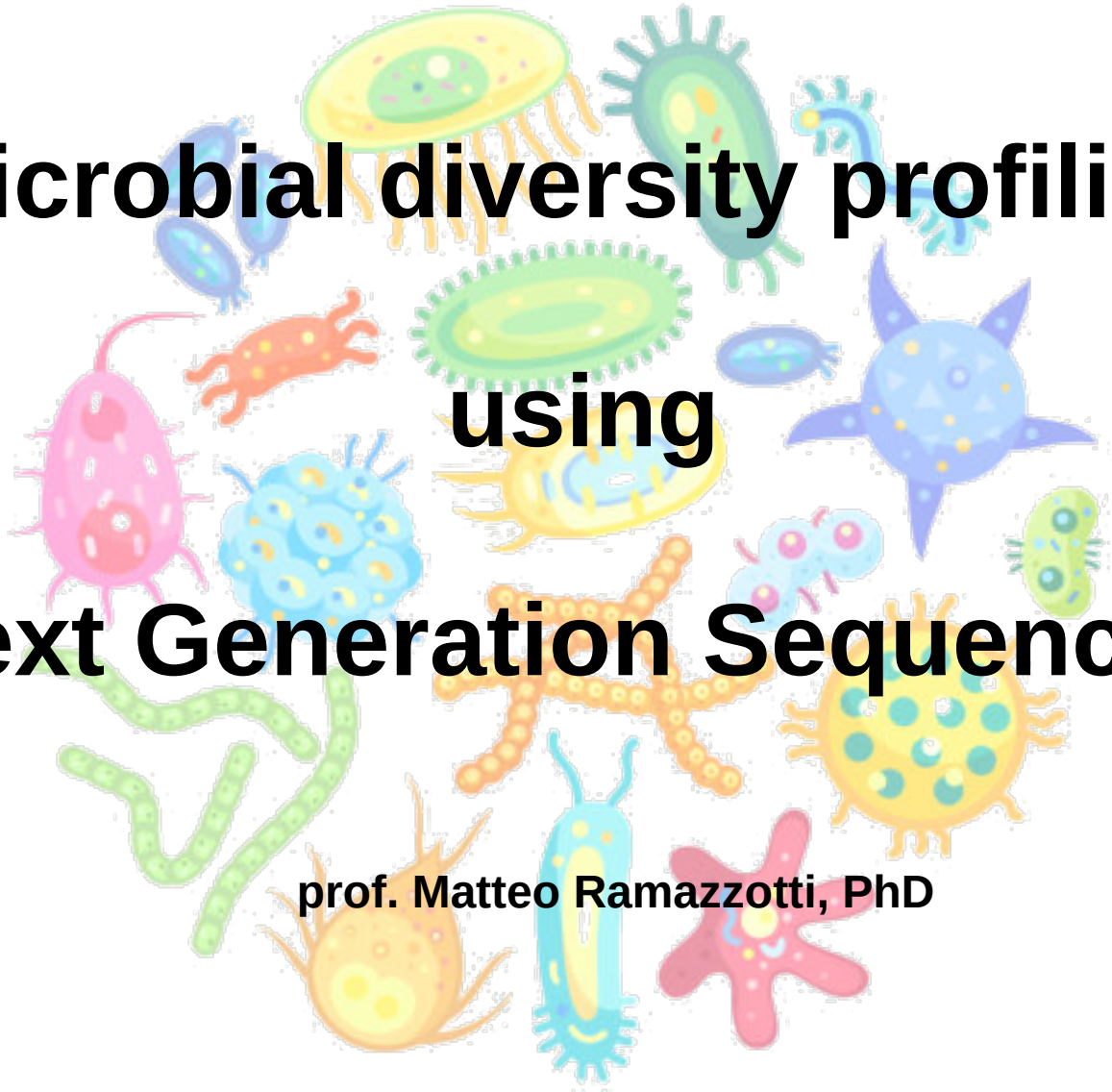


# Microbial diversity profiling using Next Generation Sequencing

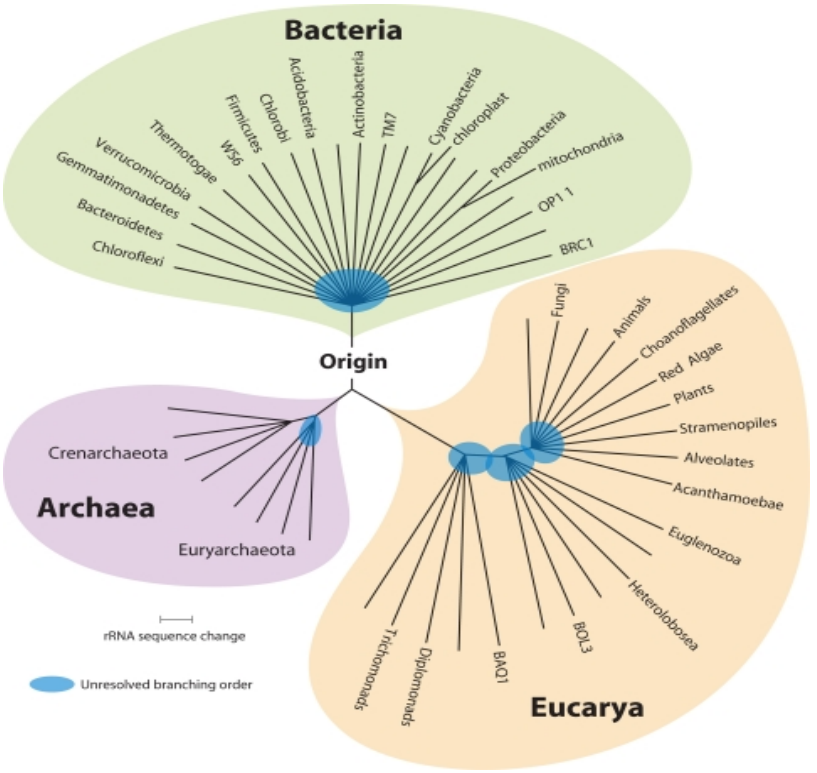
**prof. Matteo Ramazzotti, PhD**

# The diversity of prokaryotes

The prokaryotes consist of

**many millions**

of genetically distinct organisms.




They are phylogenetically distinct from eukaryotes and comprise Bacteria and Archea.

Less than 1% are known because most of them cannot be isolated and cultured.

A combination of techniques is needed to give them a (partial) classification:

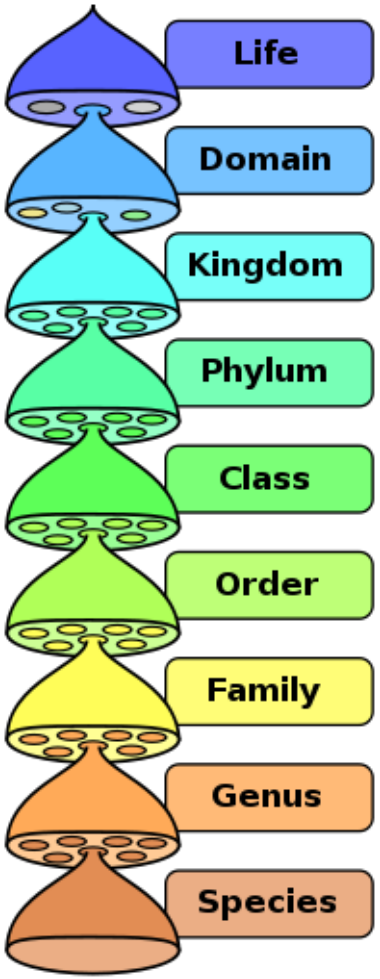
- Morphology
- Motility
- Structural features
- Oxygen requirements
- Metabolism
- Gram stain
- Temperature resistance
- GC%
- Physiology
- Genetics

Woese C PNAS 1990
www.textbookofbacteriology.net

# Definition of species

The definition of “species” has evolved with scientific knowledge and is far to have a stable definition. According to John Wilkins:

- ✔ There is one species concept (and it refers to real species).
- ✔ There are two explanations of why real species are species: ecological adaptation and reproductive reach.
- ✔ There are seven distinct definitions of species plus 27 variations.
- ✔ And there are *n+1* definitions of "species" in a room of *n* biologists.




**Taxonomy** is the practice and science of classification. It organizes *taxa* (singular *taxon*) in hierarchical, progressively inclusive classes termed **ranks**.

**Taxonomic profiling** is the characterization the taxonomic diversity of a population, e.g. a microbial community.

Wilkins J. Theory in Biosciences, 129, 141-148.

# The diversity of prokaryotes: taxonomy

Since January 2000, names of prokaryotes change at a rate approaching 750 validly published names every year.

"List of Prokaryotic names with Standing in Nomenclature" […] provides accurate information about the current status of a name, synonyms, and other useful information."


Genus

Abiotrophia
Family: "*Aerococcaceae*" - (List of genera included in the family)
Suborder:
Order: "*Lactobacillales*" - (List of genera included in the order - List of families included in the order)
Subclass:
Class: "*Bacilli*" - (List of orders and suborders included in the class)
Division or phylum: "Firmicutes" - (see Phylum "Firmicutes" in the file Classification of domains and phyla)
Domain or empire: "*Bacteria*" - (see Domain "Bacteria" in the file Classification of domains and phyla)

Species

***Abiotrophia defectiva*** (Bouvet *et al.* 1989) Kawamura *et al.* 1995, comb. nov. (Type species of the genus).
**Type strain** (see also StrainInfo.net): strain SC10 = ATCC 49176 = CIP 103242 = CCUG 27639 = CCUG 27804 = DSM 9849 = LMG 14740.
**GenBank/EMBL/DDBJ accession number for the 16S rRNA gene sequence of the type strain**: D50541.
**Basonym**: ¤ *Streptococcus defectivus* Bouvet *et al.* 1989.
**Etymology**: L. fem. adj. *defectiva*, deficient.
**Reference**: KAWAMURA (Y.), HOU (X.G.), SULTANA (F.), LIU (S.), YAMAMOTO (H.) and EZAKI (T.): Transfer of *Streptococcus adjac*
*Abiotrophia* gen. nov. as *Abiotrophia adiacens* comb. nov. and *Abiotrophia defectiva* comb. nov., respectively. *Int. J. Syst. Bacteriol.*, 1995, **45**
Original article in IJSEM Online

LPSN is accessible at www.bacterio.net

# The diversity of prokaryotes in databanks (10 years later)

RDP8

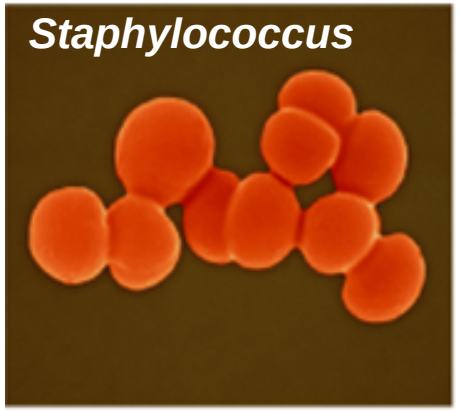
domain Bacteria (0/703222/0) ← 2007
- phylum "Acidobacteria" (0/7334/0)
- phylum "Firmicutes" (0/256748/0)
- phylum "Bacteroidetes" (0/69546/0)
- phylum "Actinobacteria" (0/119442/0)
- phylum "Aquificae" (0/851/0)
- phylum "Caldiserica" (0/44/0)
- phylum "Chlamydiae" (0/307/0)
- phylum "Chlorobi" (0/319/0)
- phylum "Chloroflexi" (0/3764/0)
- phylum "Chrysiogenetes" (0/9/0)
- phylum "Deferribacteres" (0/276/0)
- phylum "Deinococcus-Thermus" (0/851/0)
- phylum "Dictyoglomi" (0/21/0)
- phylum "Fibrobacteres" (0/235/0)
- phylum "Fusobacteria" (0/2213/0)
- phylum "Gemmatimonadetes" (0/738/0)
- phylum "Lentisphaerae" (0/176/0)
- phylum "Nitrospira" (0/818/0)
- phylum "Planctomycetes" (0/4311/0)
- phylum "Proteobacteria" (0/202798/0)
- phylum "Spirochaetes" (0/3354/0)
- phylum "Synergistetes" (0/778/0)
- phylum "Tenericutes" (0/2233/0)
- phylum "Thermodesulfobacteria" (0/94/0)
- phylum "Thermotogae" (0/478/0)
- phylum "Verrucomicrobia" (0/4194/0)
- phylum Bacteria_incertae_sedis (0/271/0)
- phylum OP10 (0/202/0)
- phylum OP11 (0/65/0)
- phylum OD1 (0/103/0)
- phylum BRC1 (0/56/0)
- phylum Cyanobacteria (0/9562/0)
- phylum SR1 (0/2/0)
- phylum WS3 (0/119/0)
- phylum TM7 (0/805/0)

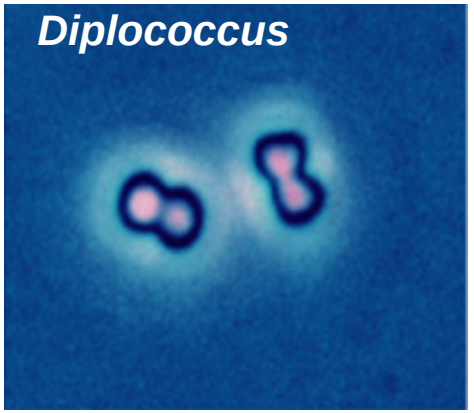
RDP11

domain Bacteria (0/1502575/0) ← 2017
- phylum "Actinobacteria" (0/226279/0)
- phylum "Aquificae" (0/1013/0)
- phylum "Bacteroidetes" (0/180267/0)
- phylum "Caldiserica" (0/230/0)
- phylum "Chlamydiae" (0/800/0)
- phylum "Chlorobi" (0/409/0)
- phylum "Chloroflexi" (0/22363/0)
- phylum "Chrysiogenetes" (0/12/0)
- phylum "Deferribacteres" (0/420/0)
- phylum "Deinococcus-Thermus" (0/2186/0)
- phylum "Dictyoglomi" (0/20/0)
- phylum "Elusimicrobia" (0/403/0)
- phylum "Fibrobacteres" (0/850/0)
- phylum "Fusobacteria" (0/11979/0)
- phylum "Gemmatimonadetes" (0/1638/0)
- phylum "Lentisphaerae" (0/1798/0)
- phylum Nitrospirae (0/1802/0)
- phylum "Planctomycetes" (0/16438/0)
- phylum "Proteobacteria" (0/437996/0)
- phylum "Spirochaetes" (0/11723/0)
- phylum "Synergistetes" (0/1333/0)
- phylum "Tenericutes" (0/5184/0)
- phylum "Thermodesulfobacteria" (0/119/0)
- phylum "Thermotogae" (0/728/0)
- phylum BRC1 (0/406/0)
- phylum Parcubacteria (0/174/0)
- phylum Microgenomates (0/127/0)
- phylum SR1 (0/343/0)
- phylum Candidatus Saccharibacteria (0/2670/0)
- phylum Latescibacteria (0/559/0)
- phylum "Armatimonadetes" (0/1178/0)
- phylum "Verrucomicrobia" (0/10731/0)
- phylum "Acidobacteria" (0/16286/0)
- phylum Firmicutes (0/482511/0)
- phylum Cyanobacteria/Chloroplast (0/26471/0)
- phylum Marinimicrobia (0/1065/0)
- phylum Aminicenantes (0/1543/0)
- phylum Omnitrophica (0/21/0)
- phylum Acetothermia (0/40/0)
- phylum Poribacteria (0/105/0)
- phylum Atribacteria (0/69/0)
- phylum Cloacimonetes (0/220/0)
- phylum Candidatus Calescamantes (0/3/0)
- phylum candidate division WPS-1 (0/311/0)
- phylum candidate division WPS-2 (0/183/0)
- phylum Hydrogenedentes (0/468/0)
- phylum candidate division ZB3 (0/73/0)
- phylum Ignavibacteriae (0/870/0)
- phylum Nitrospinae (0/619/0)

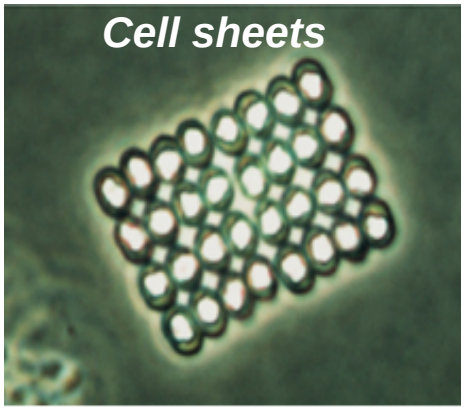
http://rdp.cme.msu.edu/hierarchy/hierarchy_browser.jsp?

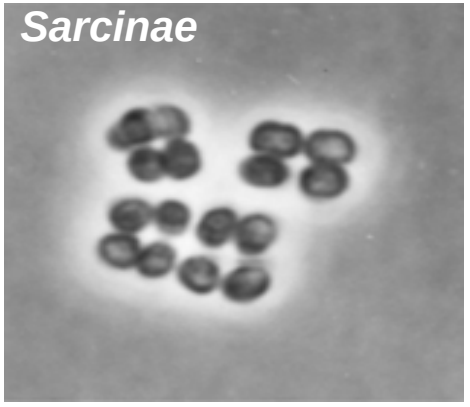
# The diversity of prokaryotes: shapes

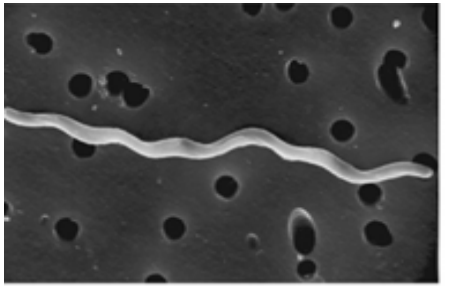


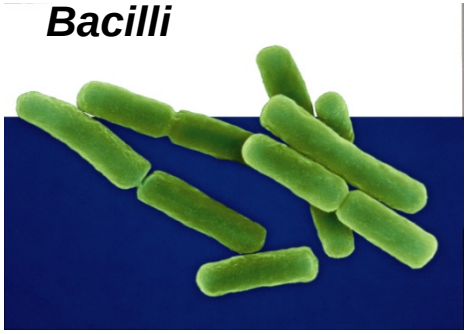



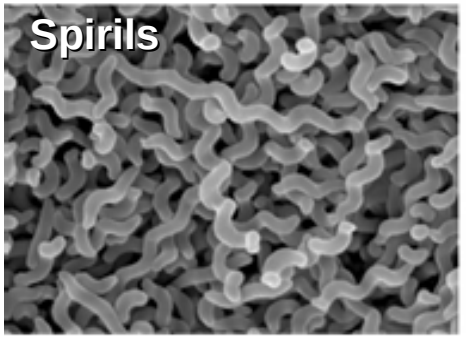


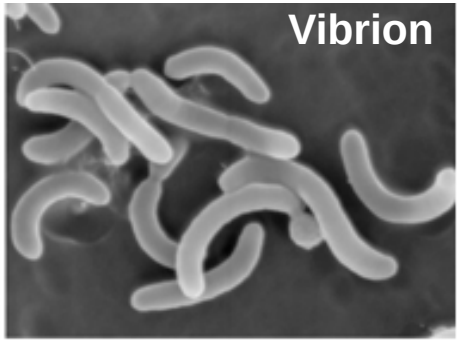




# The diversity of prokaryotes: guidelines

| Macroscopic morphology | Microscopic morphology | Cell component | Growth characteristics | Metabolism | Molecular genetics | Biochemistry |
|---|---|---|---|---|---|---|
| Type of growth | Shape of the cell | Cell wall | Athmospheric requests | Carbon sources | GC% | Exposed antigens |
| Shape of the colony | Dimension of the cell della cell | Gram staining | PH tolerance | Nitrogen sources | DNA sequences | Enzymes |
| Pigments | Internal structures | Capsule | Growth temperature | Sulfur sources | rDNA sequences | Toxins |
| | Accessory structures | | Lifestyle | Type of fermentation | Molecular probes | |
| | | | Sensitivity to antibiotics | Type of respiration | PCR | |
| | | | | End products | | |

# The current definition of a prokaryotic species

A bacterial species is defined as a group of strains having

1. > 70% DDH (DNA/DNA Hybridization) similarity
2. < 5°C Δtm
3. < 5% mol G + C difference of total genomic DNA
4. > 98% 16S rRNA identity (full length gene).

Modern genomics have proposed two additional measures

5. AAI : average amino acid identity.
6. ANI: average nucleotide identity (>94% ANI ~ > 70% DDH)
7. digital DDH (based on complete genomes) (> 70% dDDH)

In prokaryotes, species definition is “operational”, used to rapidly establish a taxonomic framework based on microbial evolution.

**No biological criteria are involved in the defintion of a prokaryote species**

Thomposon CC et al. BMC Genomics 2013, 14:913

**FOR OUR PURPOSES:**

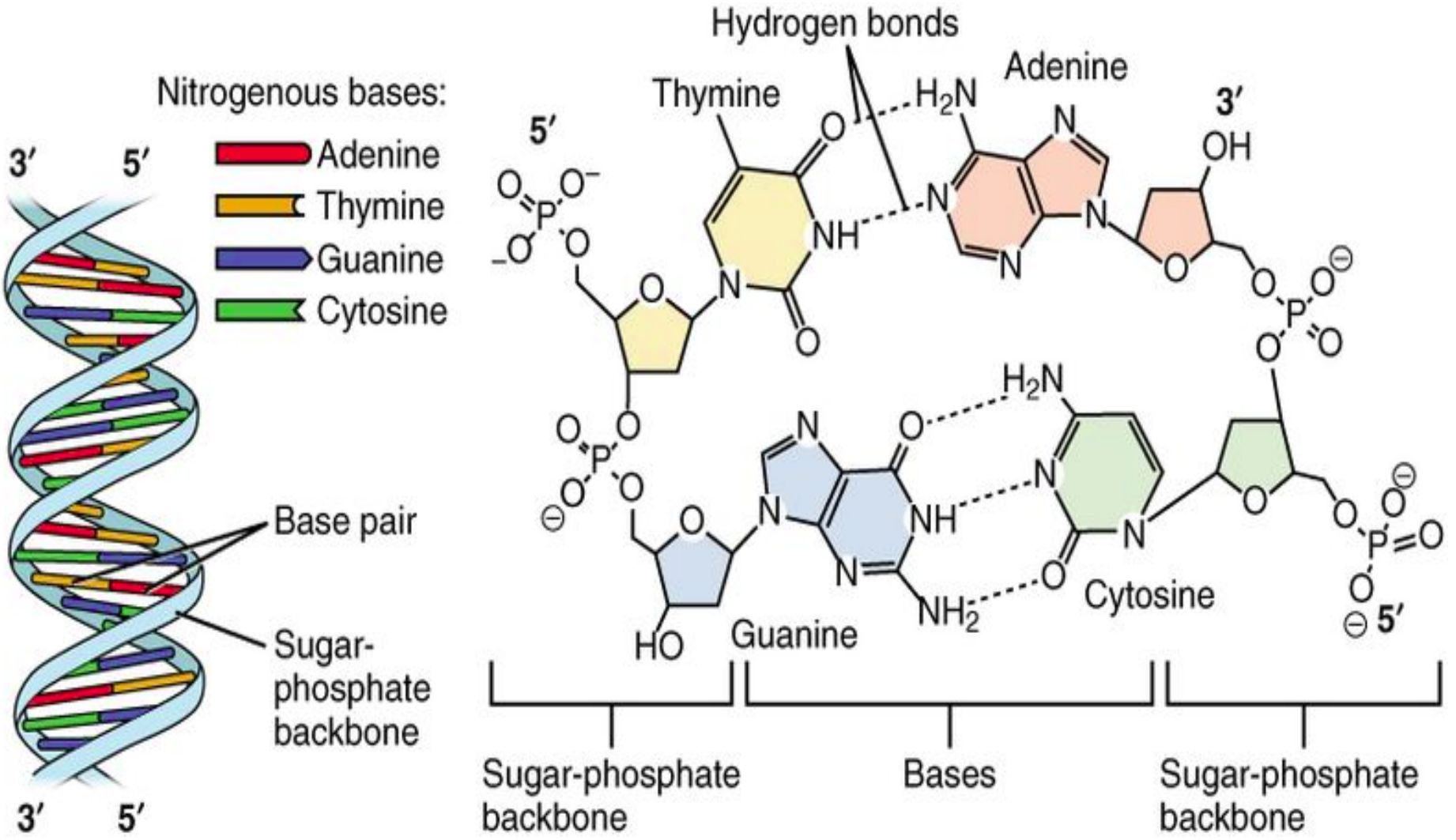
= 1 sequence and one species

**REAL BACTERIA**

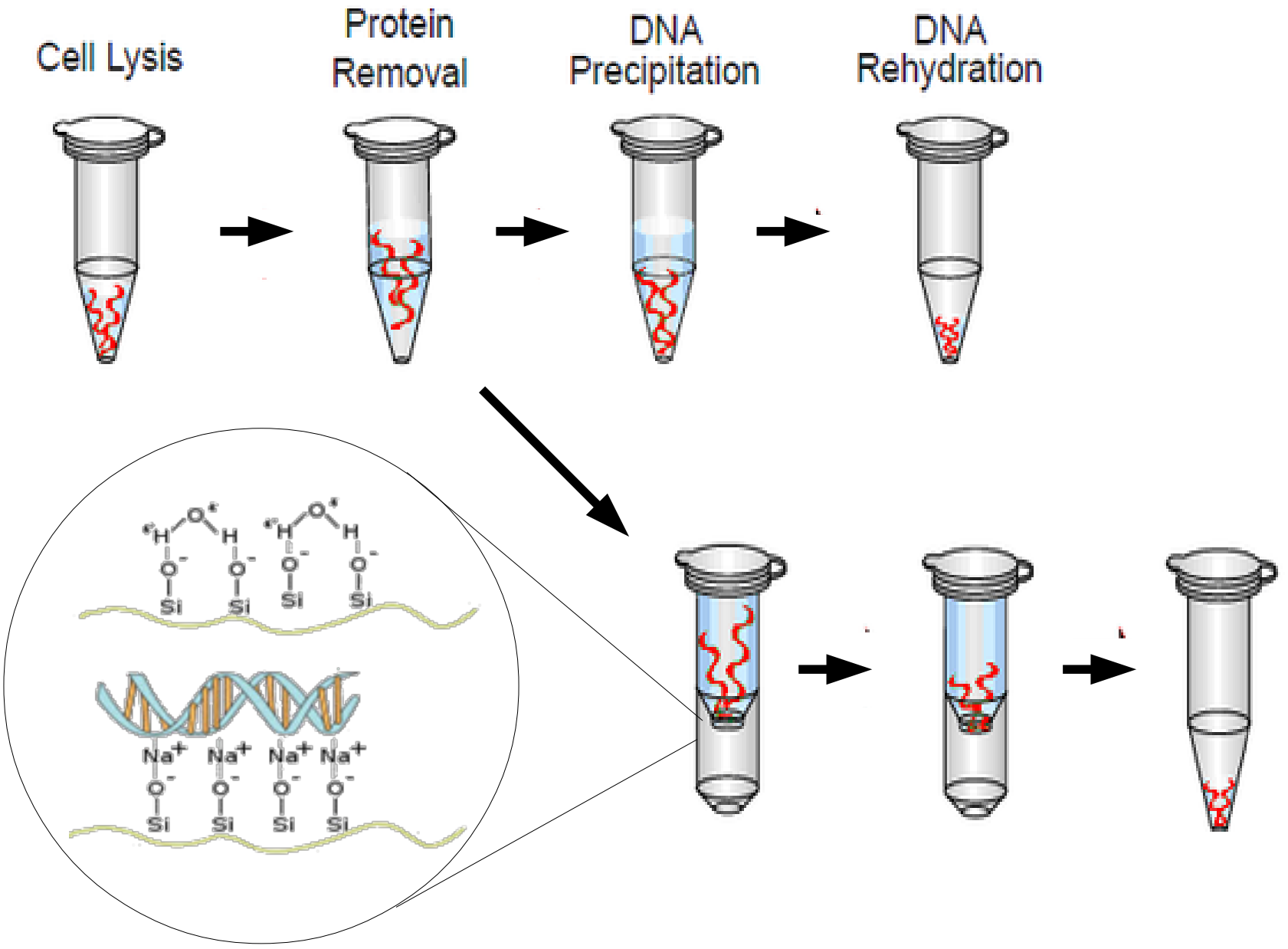
One species can have multiple, different copies of a gene

Two similar species may share an identical sequence

# Chemical structure of DNA

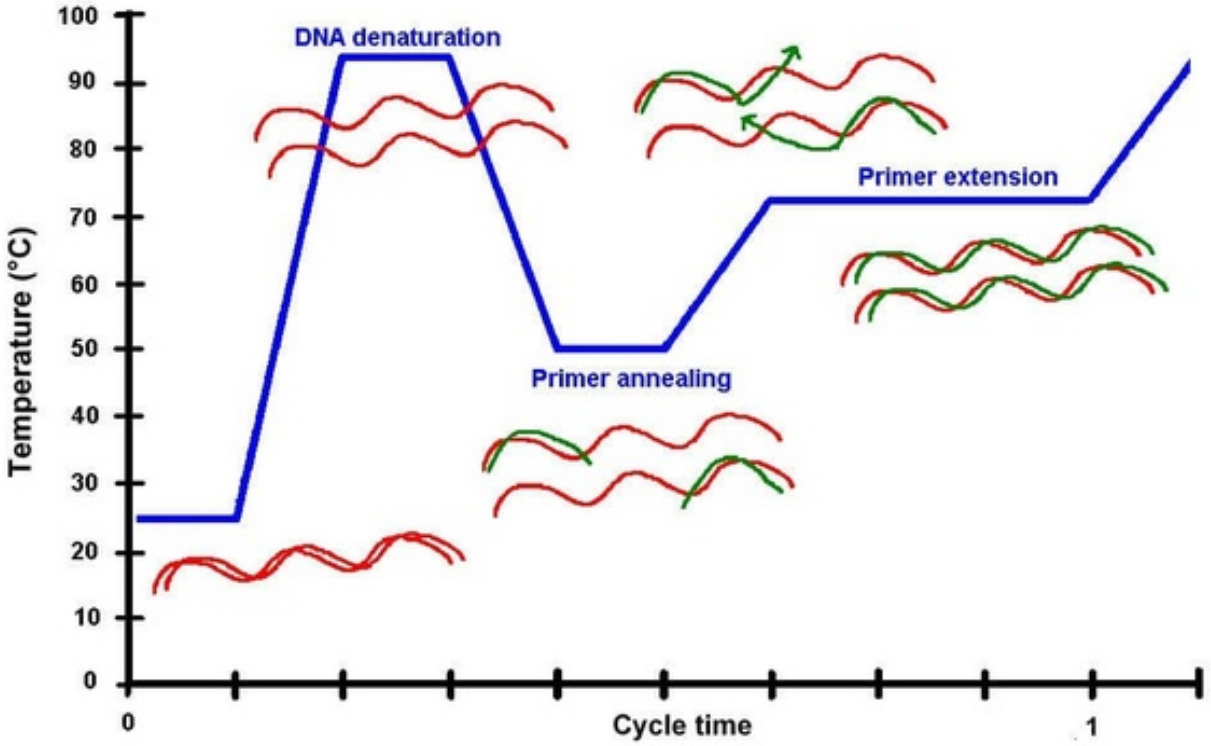


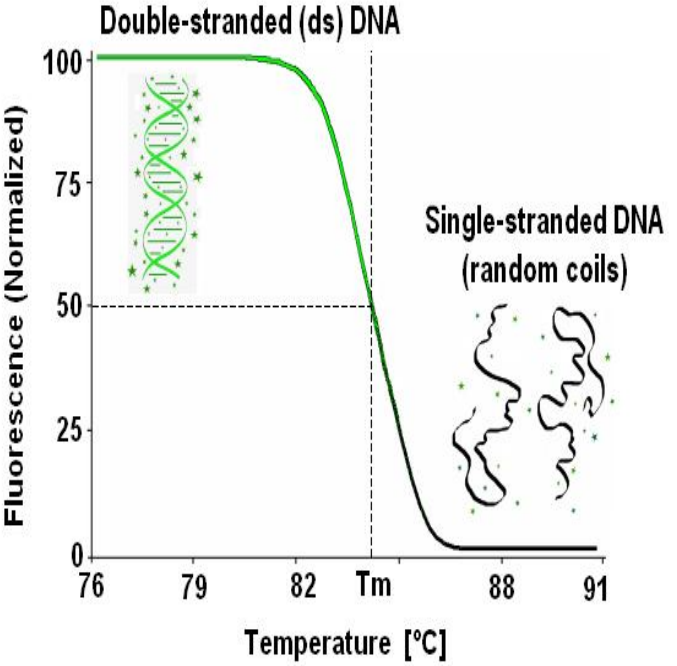
# DNA extraction



# Polymerase Chain Reaction

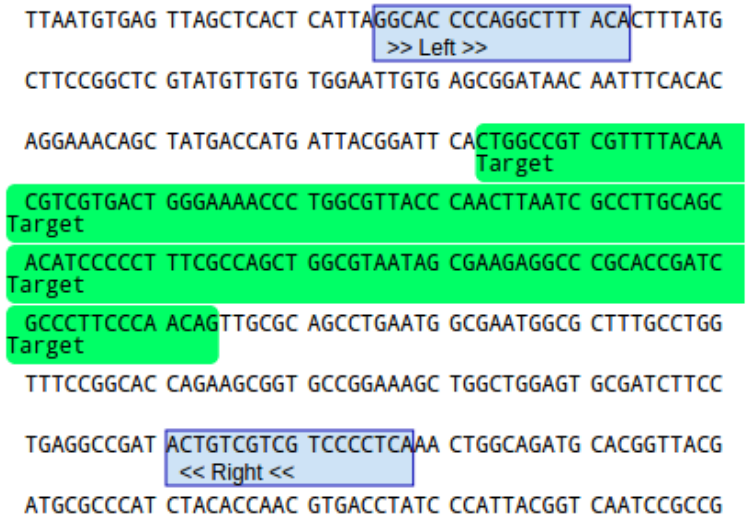
| DNA duplication in vivo | DNA duplication in vitro (PCR) |
|---|---|
| DNA | Extracted DNA |
| Elicase and topoisomerase | Heat |
| DNA primase for 3'OH primers | Synthesized 3'OH primers |
| DNA polymerase | Purified DNA polymerase |
| Nucleotides | Nucleotides |

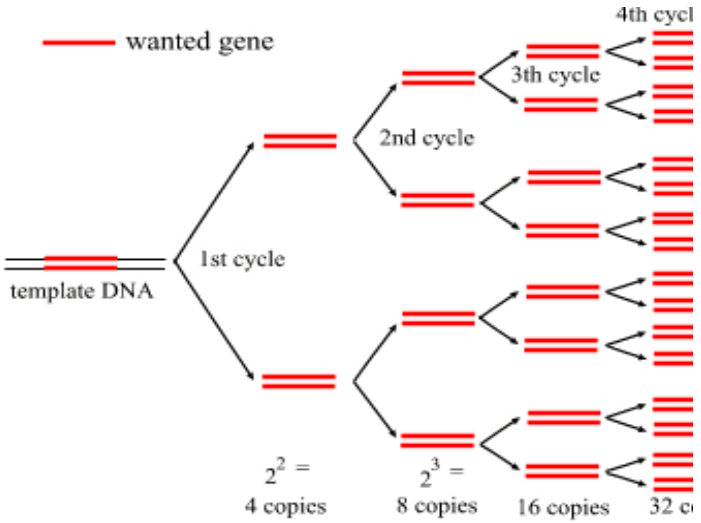


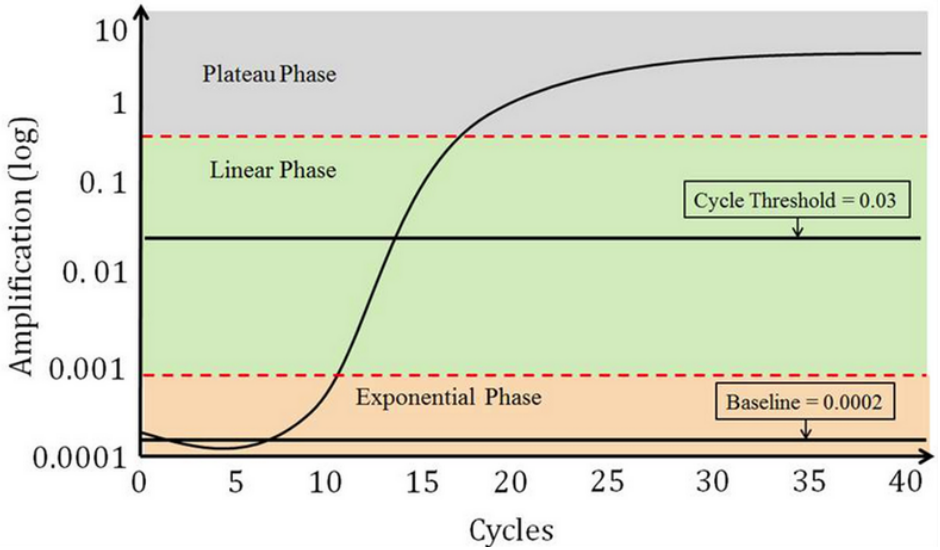



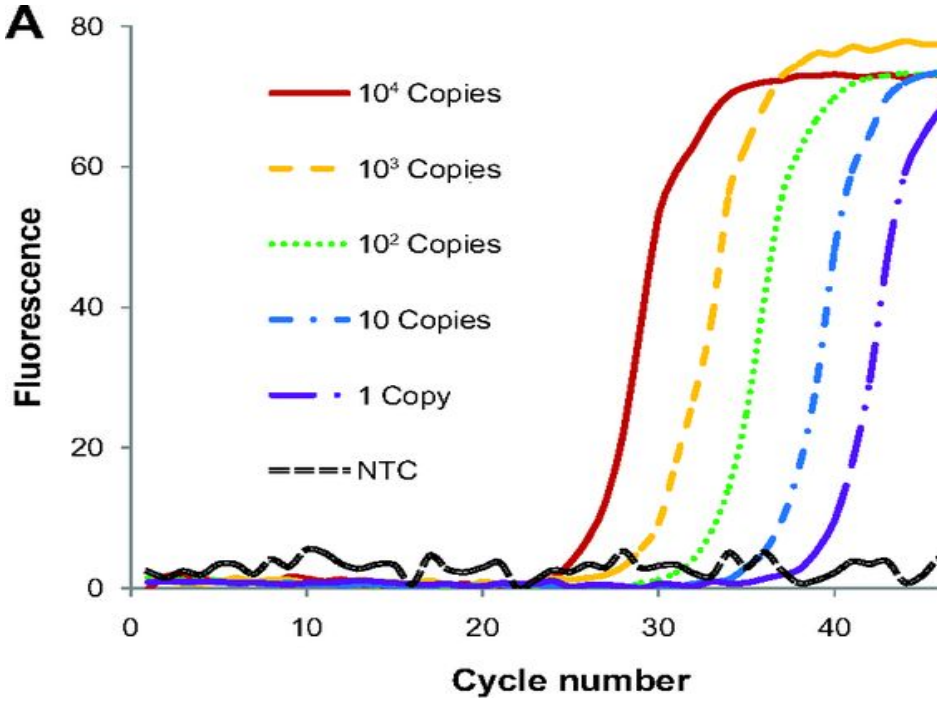

Kari B. Mullis,
Nobel prize 1993

# Chain Reaction: an amplification cascade

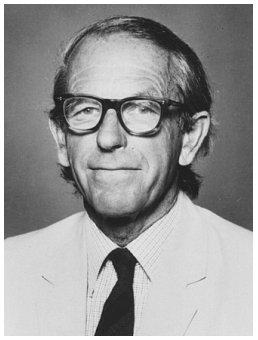


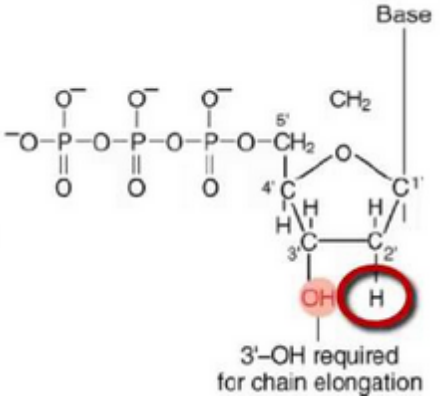



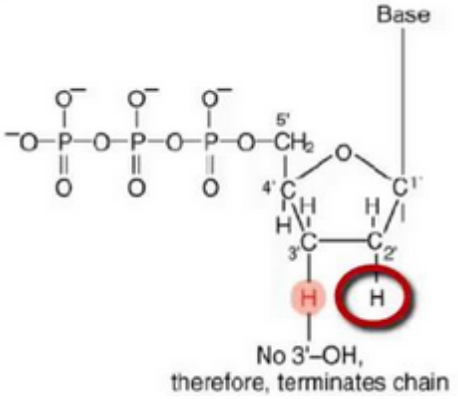



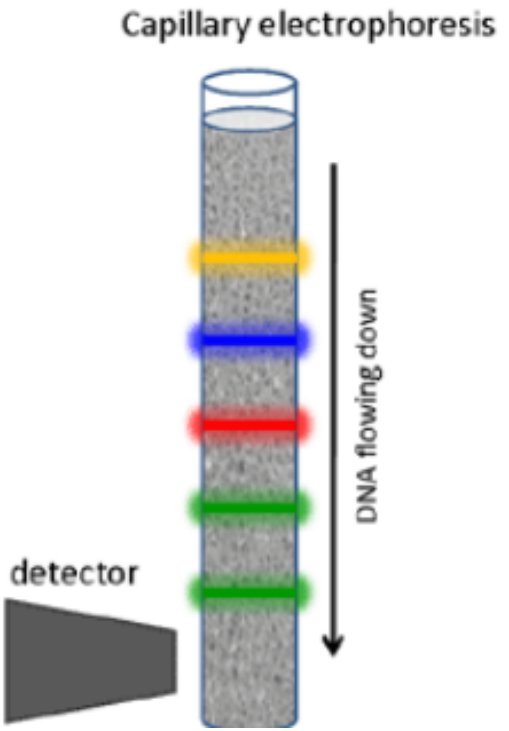
# DNA sequencing with the Sanger method



Frederick Sanger
Nobel prize 1958/1980

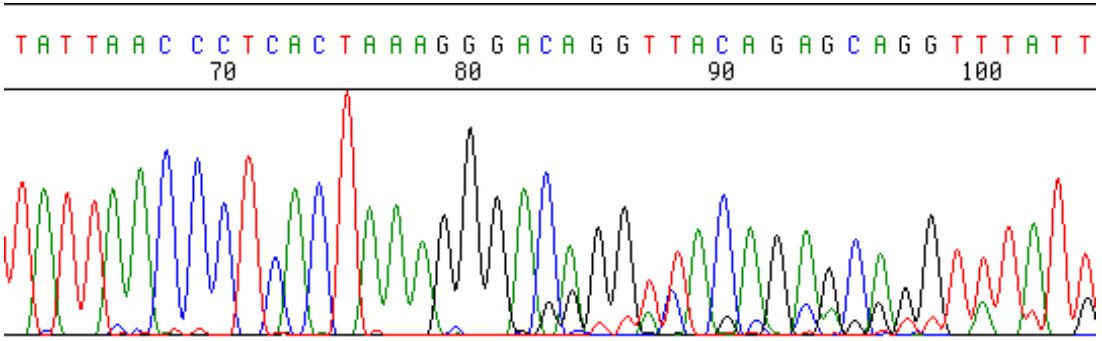


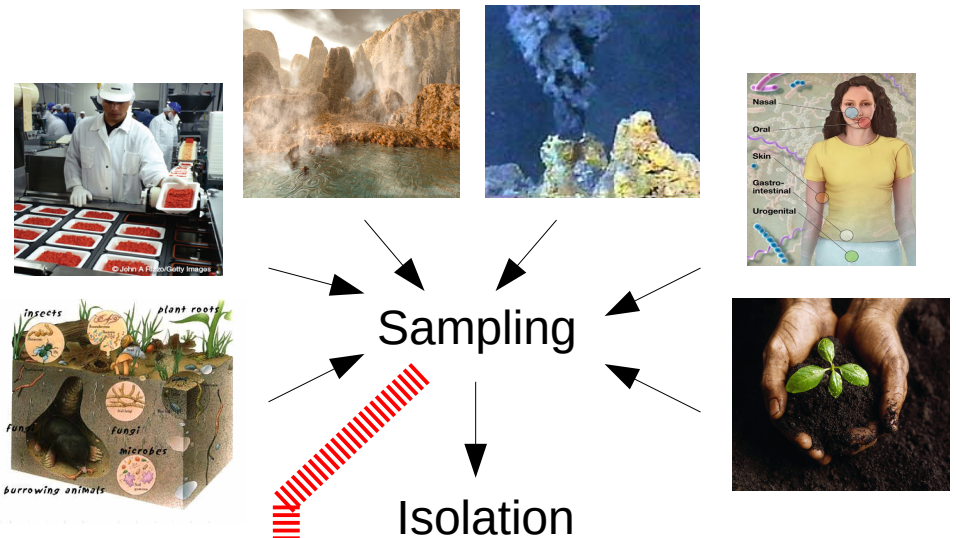




Template Sequence
3'GAGCAAATTCCGATACATTATTGT... 5'
Primer
5'CTCGTTTAAG... 3'

CTCGTTTAAGG
CTCGTTTAAGGC
CTCGTTTAAGGGT
CTCGTTTAAGGGTA
CTCGTTTAAGGGTAT
CTCGTTTAAGGGTATG
CTCGTTTAAGGGTATGT
CTCGTTTAAGGGTATGTA
CTCGTTTAAGGGTATGTAA
CTCGTTTAAGGGTATGTAAT



# A meta-approach to microbial analysis




Sampling

Isolation

Culturing

DNA extraction

PCR

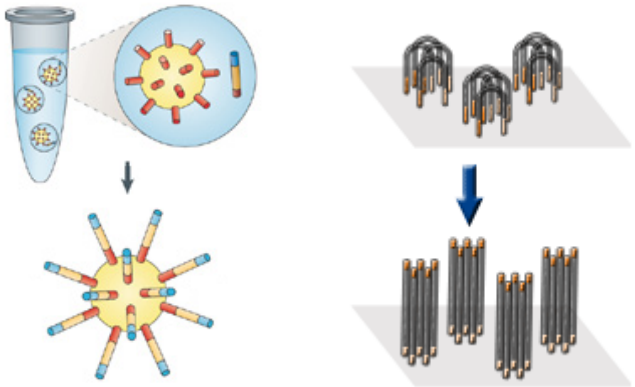
Sanger sequencing

NGS sequencing

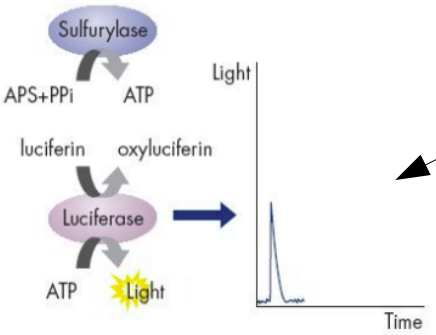
Traditional methods of microbial ecology require organisms collected from an environment to be cultured in the laboratory. From 90% to 99.9% of the cells defy cultivation, and therefore evade identification.

Metagenomics only needs DNA (whole extract or amplified via PCR) and sequencing occurs in parallel: no need for isolation.

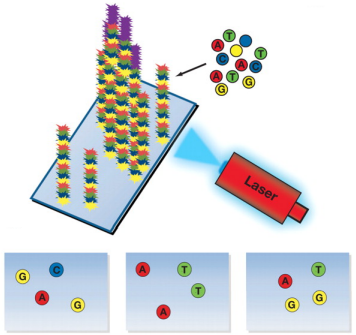
# NGS technologies



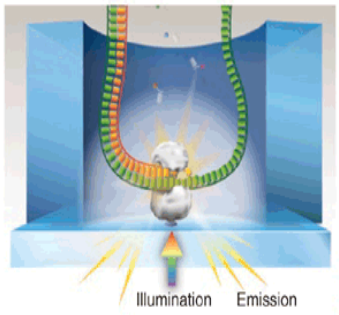
Emulsion PCR Bridge PCR



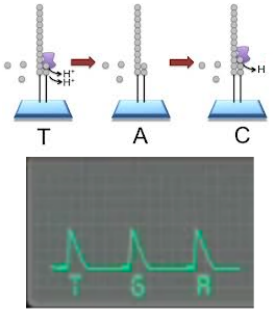

Pyrosequencing



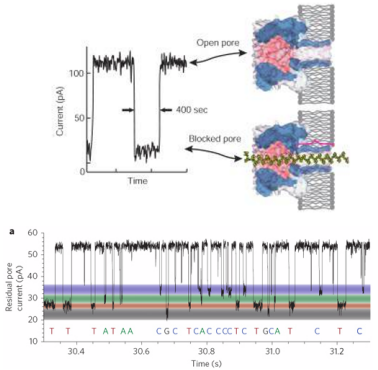

Illumina



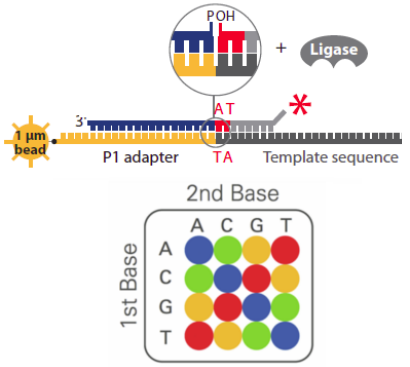

PacBio



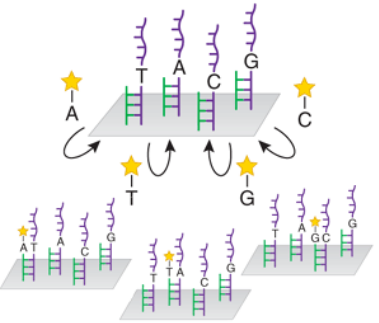

IonTorrent




Oxford nanopores




SOLiD



Helicos

# NGS and metagenomics

| Platform | Instrument | Unit | Reads/unit | Read Length (bp) | Read Type | Error Type |
|---|---|---|---|---|---|---|
| PacBio | PacBio RS II | SMRT Cell | 47000 | 14000 | SR | indel |
| PacBio | PacBio RS | SMRT Cell | 22000 | 4600 | SR | indel |
| Roche 454 | GS FLX+ / FLX | 1 PTP | 700000 | 700 | SR | indel |
| Roche 454 | GS FLX+ / FLX | 1/2 PTP | 350000 | 700 | SR | indel |
| Roche 454 | GS FLX+ / FLX | 1/4 PTP | 125000 | 700 | SR | indel |
| Roche 454 | GS FLX+ / FLX | 1/8 PTP | 50000 | 700 | SR | indel |
| Roche 454 | GS FLX+ / FLX | 1/16 PTP | 20000 | 700 | SR | indel |
| Illumina | MiSeq v3 | Lane | 25000000 | 600 | SR & PE | substitution |
| Illumina | MiSeq v2 Nano | Lane | 1000000 | 500 | SR & PE | substitution |
| Ion | PGM 318 | Chip | 4000000 | 400 | SR | indel |
| Ion | PGM 316 | Chip | 2000000 | 400 | SR | indel |
| Ion | PGM 314 | Chip | 400000 | 400 | SR | indel |
| Roche 454 | GS FLX+ / FLX | 1 PTP | 70000 | 400 | SR | indel |
| Illumina | HiSeq X | Lane | 375000000 | 300 | PE | substitution |
| Illumina | HiSeq 3000/4000 | Lane | 312500000 | 300 | SR & PE | substitution |
| Illumina | NextSeq 500 High-Output | Run | 400000000 | 300 | SR & PE | substitution |
| Illumina | NextSeq 500 Mid-Output | Run | 130000000 | 300 | PE | substitution |
| Illumina | HiSeq Rapid Run | Lane | 150696000 | 300 | SR & PE | substitution |
| Illumina | GAIIx | Lane | 42075000 | 300 | SR & PE | substitution |
| Illumina | MiSeq v2 Micro | Lane | 4000000 | 300 | SR & PE | substitution |
| Illumina | HiSeq High-Output v4 | Lane | 250000000 | 250 | SR & PE | substitution |
| Illumina | HiSeq High-Output v3 | Lane | 186048000 | 250 | SR & PE | substitution |
| Illumina | MiSeq v2 | Lane | 16000000 | 250 | SR & PE | substitution |
| Illumina | MiSeq | Lane | 5000000 | 250 | SR & PE | substitution |
| Illumina | HiScanSQ | Lane | 93024000 | 200 | SR & PE | substitution |
| Ion | Proton I | Chip | 60000000 | 200 | SR | indel |
| SOLiD | 5500xl W | Lane | 266666667 | 100 | SR & PE | A/T Bias |
| SOLiD | 5500 W | Lane | 266666667 | 100 | SR & PE | A/T Bias |
| SOLiD | 5500 | Lane | 81500000 | 100 | SR & PE | A/T Bias |
| SOLiD | 5500xl | Lane | 81500000 | 100 | SR & PE | A/T Bias |

Taken from http://genohub.com/ngs-instrument-guide/

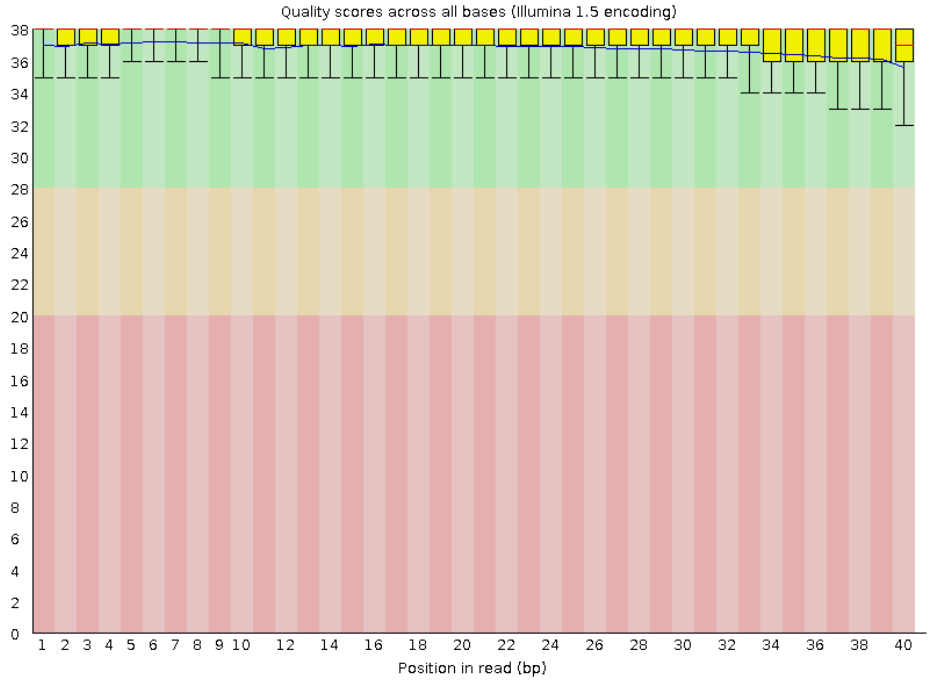
# Adding quality to each sequenced base

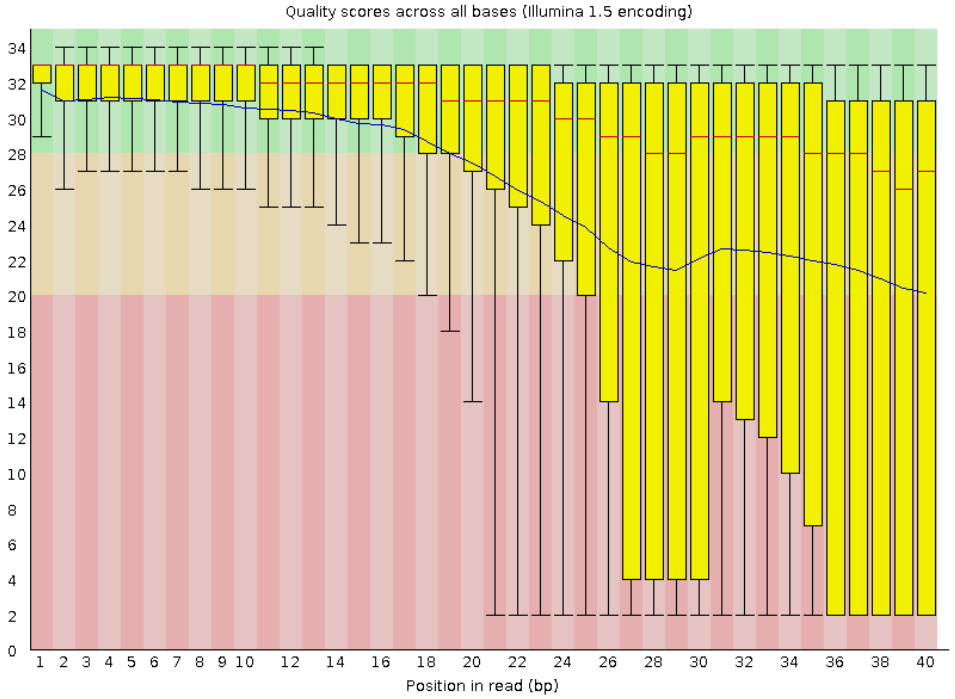
FASTA format

```
>SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCAT
```

FASTQ format

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCAT
+
!''*(((***+))%%%++)(%%%%).1***-+*''))**55
```




www.bioinformatics.babraham.ac.uk/projects/fastqc

# The quality score: phred

Introduced by Phil Green's group at the University of Washington in the 1990s, the phred quality estimator **calls each sequenced base with an integer related to the probability for a base to be incorrectly identified**. Born **for ABI chromatograms**, it has been reused in NGS base-callers, with technology specific modifications.

$$p = 10^{-\frac{q}{10}} \qquad q = -10\log_{10}(p)$$

p = error probability for the base

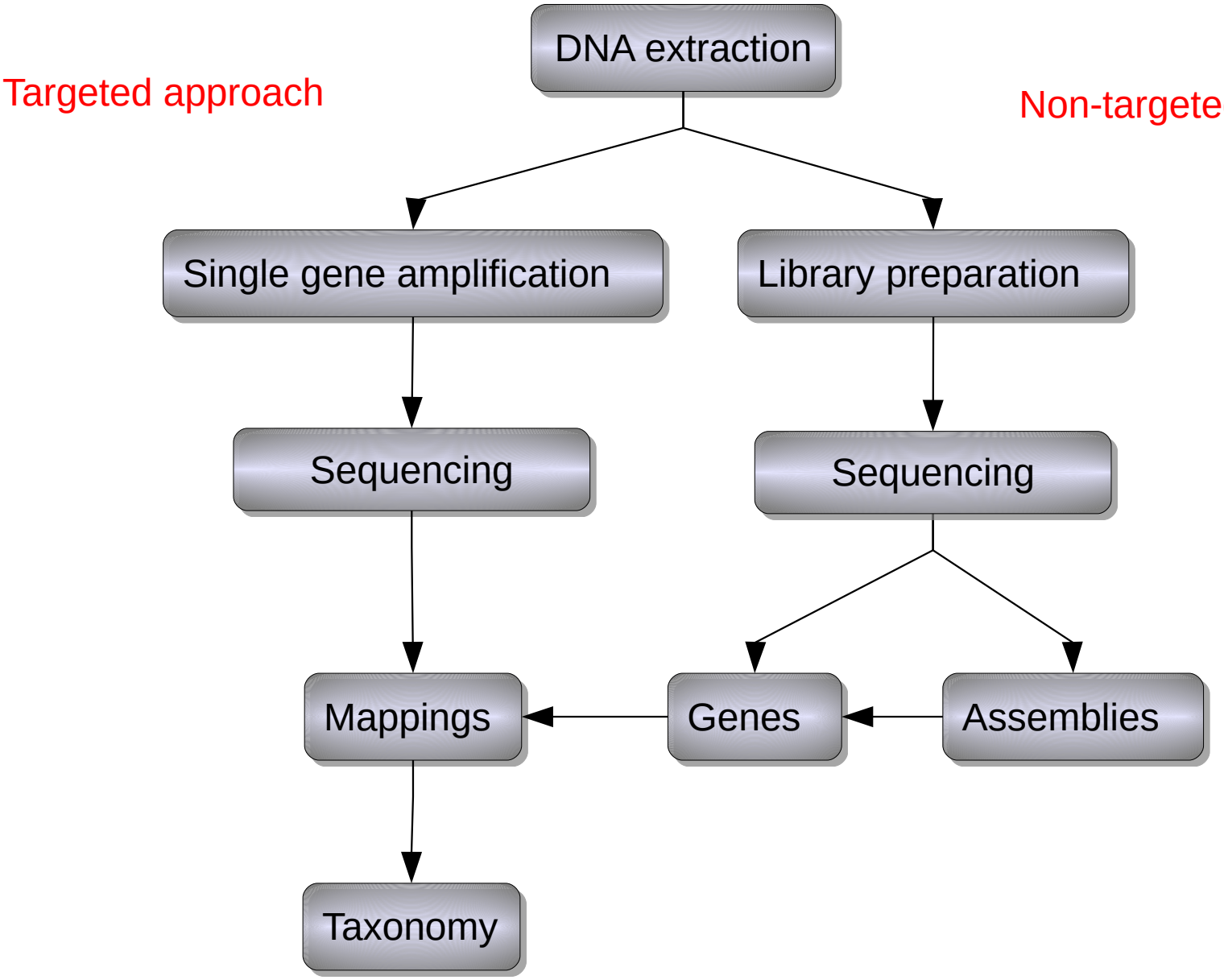
if p = 0.01 (1% chance of error), then q = 20

Phred quality values are rounded to the nearest integer

e.g.
q = 32
$p = 10^{(32/-10)} = 0.00063$

| Quality Value | Error Probability | Probability Called Base is Correct | Description |
|---|---|---|---|
| 10 | 0.1 | 0.9 | error rate of 1 in 10 |
| 20 | 0.01 | 0.99 | error rate of 1 in 100 |
| 30 | 0.001 | 0.999 | error rate of 1 in 1000 |
| 40 | 0.0001 | 0.9999 | error rate of 1 in 10000 |

# Obtaining phylogenetic information from NGS



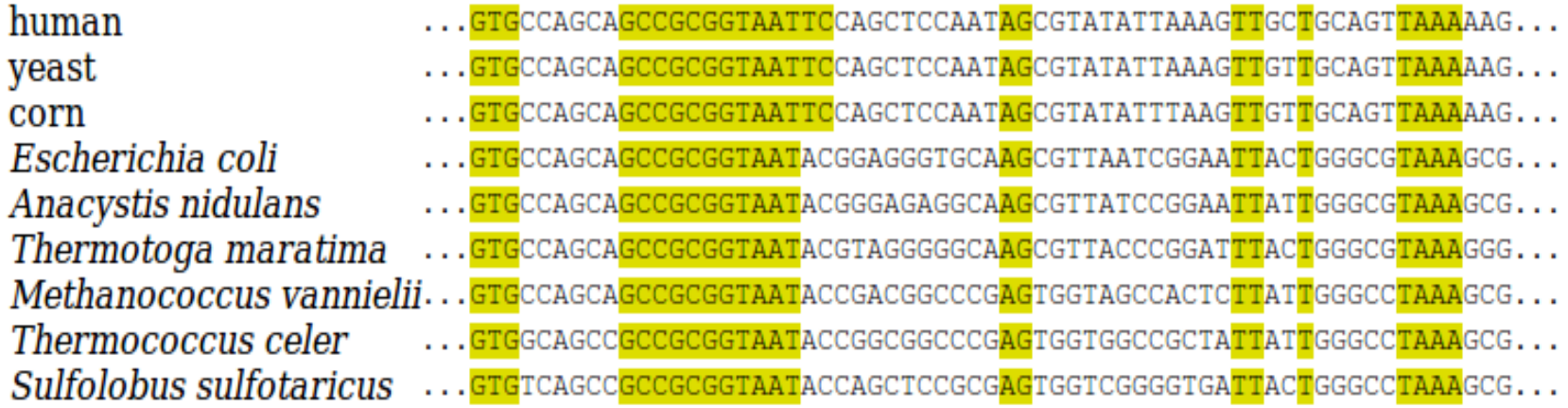
**A good molecular classifier is a regoin in the DNA that:**

- ✔ must be present in all organisms of interest.
- ✔ cannot be subject to transfer between species (lateral transfer).
- ✔ must display an appropriate level of sequence conservation for the divergences of interest.
- ✔ must be sufficiently large to contain a record of the historical information.
- ✔ (optional) singolarity, i.e. being present in only one copy per genome ...

**We need a genetic element that is involved in a central dogma of biology and that may accommodate a significant number of changes though maintaining its fundamental function...**

# Ribosomal DNA as a molecular clock

Ribosomes are RNA/protein complexes that act as the universal protein factory in all living cells. They have been conserved trough billions of years. So some regions have sequences identical in all the three kingdoms of life.




Ribosomal RNAs in Prokaryotes

| Name | Size (bp) | Location |
|---|---|---|
| **5S** | 120 | Large subunit of ribosome |
| **16S** | 1500 | Small subunit of ribosome |
| **23S** | 2900 | Large subunit of ribosome |

Carl Woese in 1990 realized that 16S rDNA could be used for taxonomy and phylogenesis.

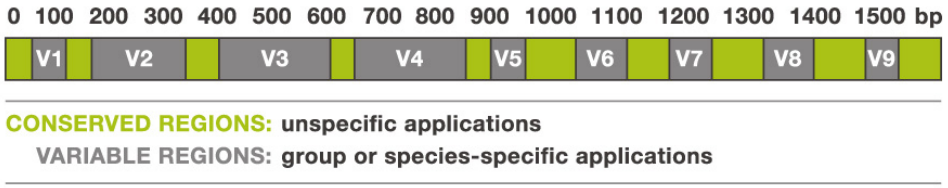
# Ribosomal DNA as a molecular clock: the 16S case

16S rDNA is a perfect candidate because it is:

- present in all prokaryotic organisms

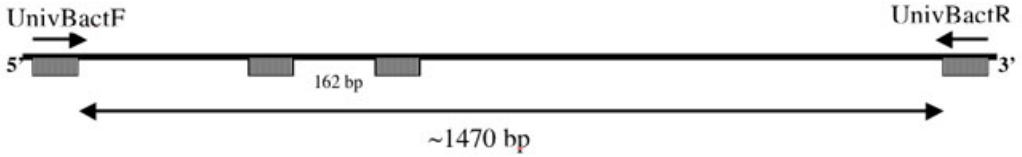
- sufficiently conserved in some regions

- highly variable in other regions




A Simplified map of the 16S rRNA molecule

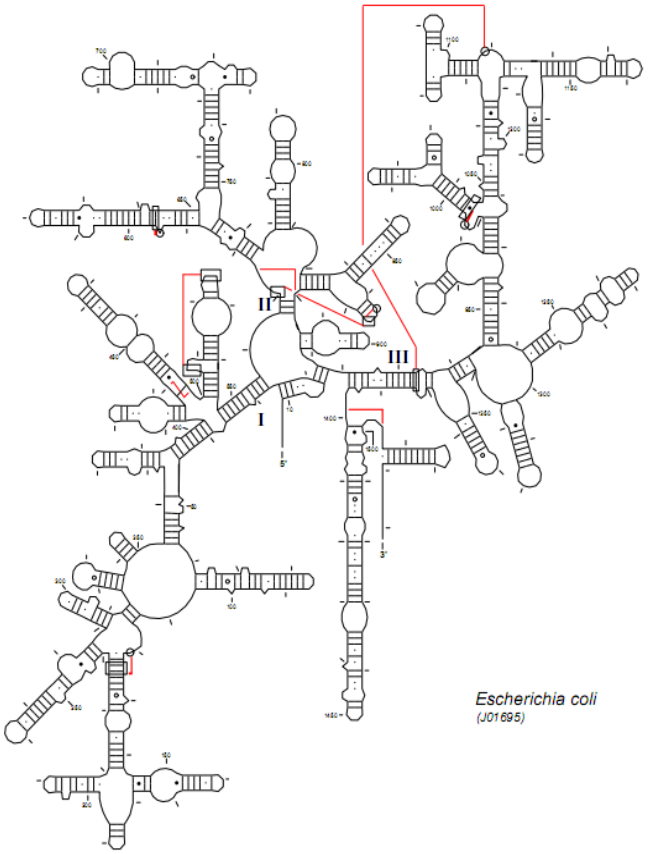
| Primer: | Primer Sequence: | Primer Length | *E. coli* Position |
|---|---|---|---|
| UnivBactF | 5' GAG TTT GAT YMT GGC TC | 17-mer | 9 - 25 |
| UnivBactR | 5' GYT ACC TTG TTA CGA CTT | 18-mer | 1509 - 1492 |



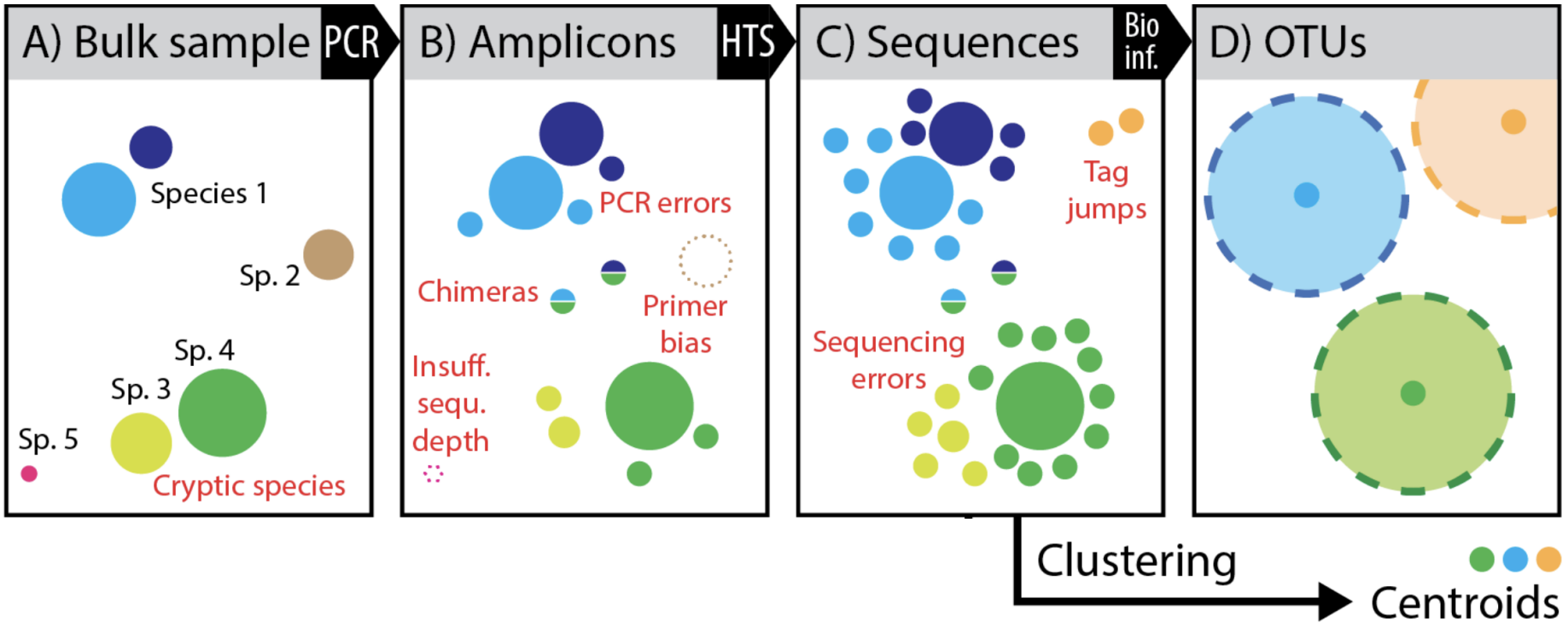

= location of primer sequence and its orientation

= area of broad range sequence conservation (sequence is virtually the same in all bacteria)

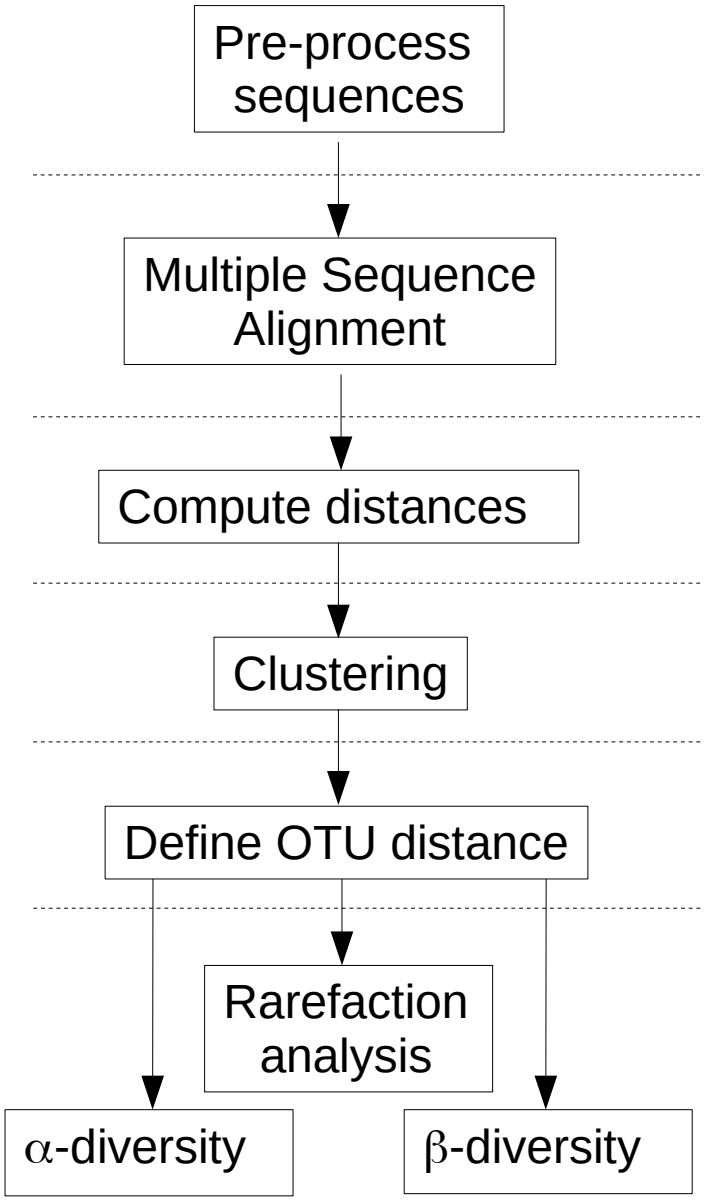
= area between primers (not including the primer sequences)



# Errors during the wet-lab processing



# OTU-based analysis: outline




Clipping, trimming, primer/adaptor removal, QC-check

Using SS-guided aligners (NAST or Infernal)

Given a distance model

Using Neighbor or Weighbor Joining

OTU: operational taxonomic unit

Classical ecology approaches for evaluating population structure and compare niches

# OTU-based analysis: pairwise distance calculation

The general definition of distance for biological sequences is

**d = fraction of mismatches at aligned positions**

So sequences need to be aligned, compared and differences counted and expressed as a fraction of the sequence length. Alternative definitions exists, based e.g. on the number of k-mer shared (no alignment is necessary in this case...)

Substitution models can be introduced now, but no generalized consensus exists.

Problems arise when gaps have to be taken into account

```
ATTTTAA-CAGGATCCCC-AGCTGCA
ATTTTAA--AGGATCCCC-AGCTGCA
ATTTTAAA-AGGATCCCC-AGCTGCA
ATTTTA---AGGATCCCCGAGCTGCA
```

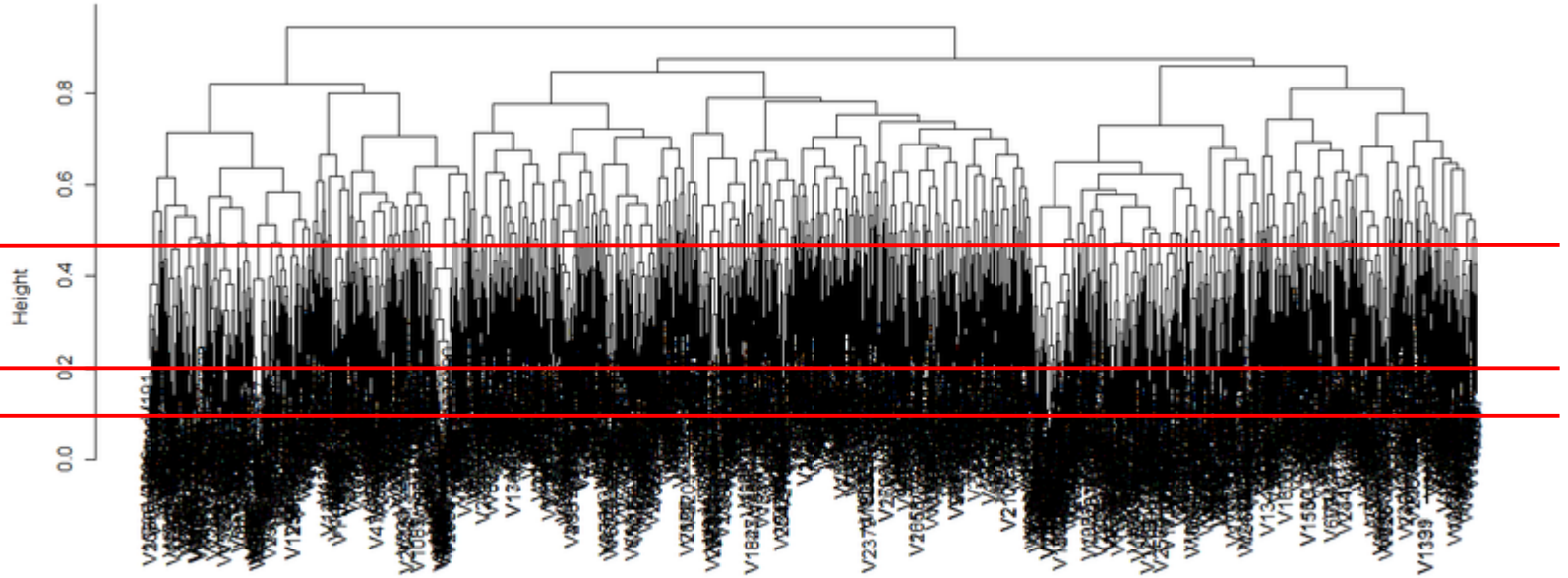
**How to consider this?**

Like this...

1. A single indel event
2. A cluster of indel events
3. No answer, ignore it.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.143 | 0.077 | 0.111 | 0.05 | 0.038 | 0.091 | 0.076 | 0.083 | 0.077 | 0.088 | 0.094 |
| 0.143 | 0 | 0.232 | 0.233 | 0.177 | 0.177 | 0.191 | 0.209 | 0.215 | 0.213 | 0.194 | 0.226 |
| 0.077 | 0.232 | 0 | 0.171 | 0.134 | 0.085 | 0.171 | 0.171 | 0.188 | 0.207 | 0.146 | 0.207 |
| 0.111 | 0.233 | 0.171 | 0 | 0.156 | 0.157 | 0.174 | 0.176 | 0.183 | 0.177 | 0.189 | 0.196 |
| 0.05 | 0.177 | 0.134 | 0.156 | 0 | 0.102 | 0.129 | 0.136 | 0.141 | 0.129 | 0.126 | 0.152 |
| 0.038 | 0.177 | 0.085 | 0.157 | 0.102 | 0 | 0.127 | 0.143 | 0.157 | 0.154 | 0.133 | 0.16 |
| 0.091 | 0.191 | 0.171 | 0.174 | 0.129 | 0.127 | 0 | 0.143 | 0.149 | 0.15 | 0.145 | 0.174 |
| 0.076 | 0.209 | 0.171 | 0.176 | 0.136 | 0.143 | 0.143 | 0 | 0.052 | 0.054 | 0.149 | 0.072 |
| 0.083 | 0.215 | 0.188 | 0.183 | 0.141 | 0.157 | 0.149 | 0.052 | 0 | 0.055 | 0.155 | 0.073 |
| 0.077 | 0.213 | 0.207 | 0.177 | 0.129 | 0.154 | 0.15 | 0.054 | 0.055 | 0 | 0.147 | 0.062 |
| 0.088 | 0.194 | 0.146 | 0.189 | 0.126 | 0.133 | 0.145 | 0.149 | 0.155 | 0.147 | 0 | 0.166 |
| 0.094 | 0.226 | 0.207 | 0.196 | 0.152 | 0.16 | 0.174 | 0.072 | 0.073 | 0.062 | 0.166 | 0 |
| 0 | 0.079 | 0.19 | 0.158 | 0.026 | 0.132 | 0.053 | 0.053 | 0 | 0.026 | 0.053 | 0.026 |
| 0.1 | 0.232 | 0.22 | 0.2 | 0.156 | 0.169 | 0.166 | 0.069 | 0.078 | 0.074 | 0.17 | 0.109 |
| 0.083 | 0.22 | 0.22 | 0.186 | 0.143 | 0.157 | 0.172 | 0.05 | 0.055 | 0.046 | 0.163 | 0.066 |
| 0.071 | 0.186 | 0.183 | 0.159 | 0.106 | 0.138 | 0.129 | 0.107 | 0.112 | 0.093 | 0.139 | 0.126 |
| 0.076 | 0.217 | 0.207 | 0.183 | 0.144 | 0.151 | 0.162 | 0.05 | 0.045 | 0.05 | 0.157 | 0.064 |
| 0.081 | 0.227 | 0.207 | 0.184 | 0.146 | 0.162 | 0.147 | 0.066 | 0.061 | 0.062 | 0.159 | 0.087 |
| 0.087 | 0.202 | 0.134 | 0.183 | 0.1 | 0.139 | 0.127 | 0.137 | 0.149 | 0.141 | 0.14 | 0.164 |
| 0.076 | 0.216 | 0.207 | 0.18 | 0.139 | 0.156 | 0.145 | 0.063 | 0.059 | 0.062 | 0.153 | 0.083 |
| 0.114 | 0.238 | 0.232 | 0.205 | 0.178 | 0.183 | 0.186 | 0.08 | 0.082 | 0.083 | 0.187 | 0.092 |
| 0.158 | 0.284 | 0.256 | 0.239 | 0.202 | 0.231 | 0.225 | 0.133 | 0.134 | 0.139 | 0.224 | 0.161 |
| 0.105 | 0.223 | 0.195 | 0.212 | 0.176 | 0.164 | 0.154 | 0.094 | 0.098 | 0.102 | 0.164 | 0.111 |
| 0.061 | 0.181 | 0.122 | 0.163 | 0.093 | 0.127 | 0.113 | 0.116 | 0.131 | 0.117 | 0.14 | 0.143 |
| 0.109 | 0.233 | 0.195 | 0.196 | 0.162 | 0.183 | 0.164 | 0.11 | 0.099 | 0.112 | 0.173 | 0.133 |
| 0.065 | 0.207 | 0.171 | 0.169 | 0.127 | 0.137 | 0.148 | 0.047 | 0.049 | 0.043 | 0.147 | 0.049 |
| 0.103 | 0.245 | 0.207 | 0.198 | 0.165 | 0.176 | 0.162 | 0.106 | 0.098 | 0.096 | 0.176 | 0.129 |

# OTU-based analysis: defining the OTU distance




Stringent OTU definition => many OTUs

Intermediate OTU definition => less OTUs

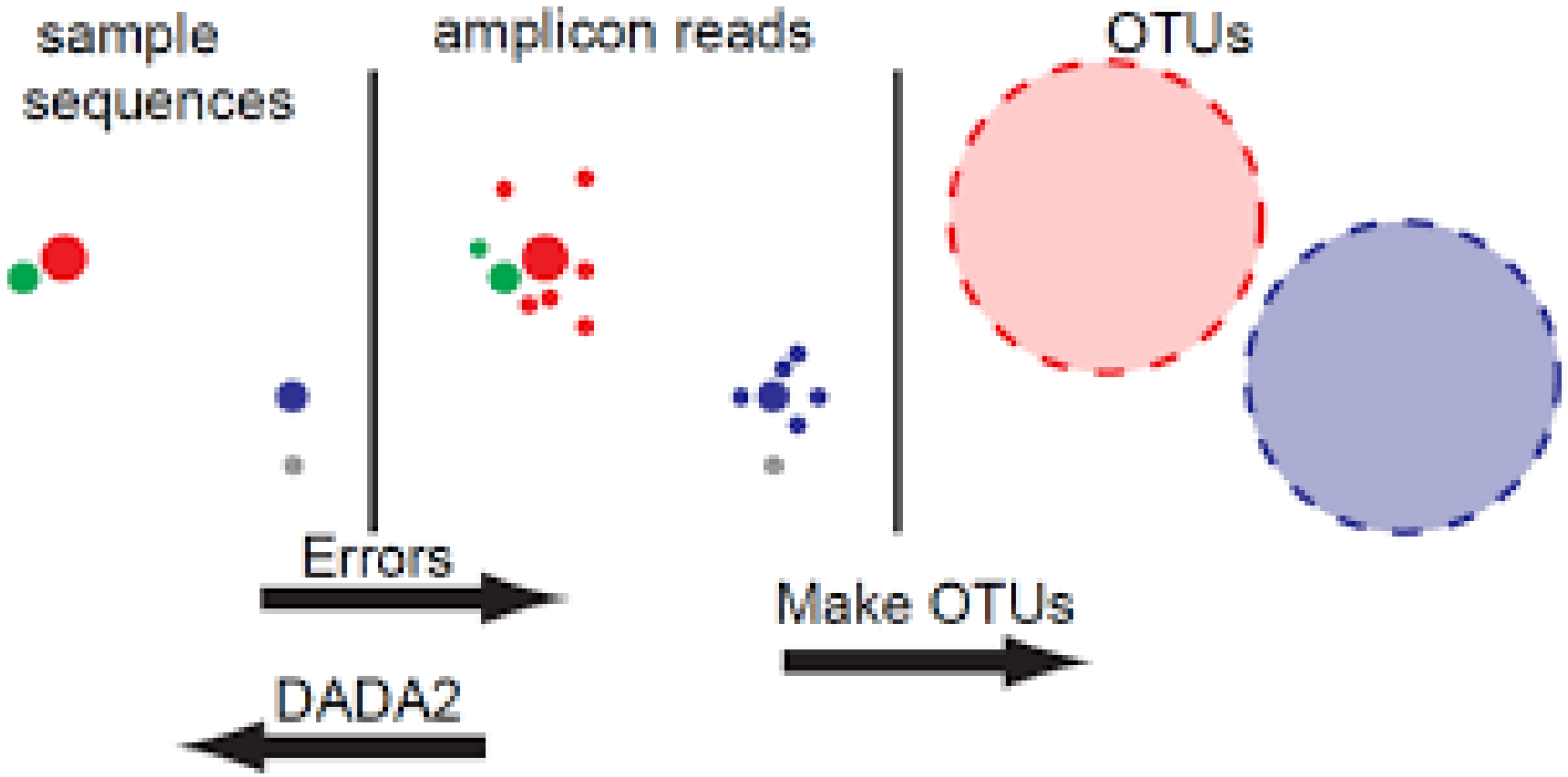
Relaxed OTU definition => few OTUs

In full length 16S rDNA analysis, 3% distance is considered as a proxy for species definition.

**But we are dealing with short reads, no universal criterion is in use, we have to explore different definitions**

**Each OTU will have a variable number of sequences (i.e. individuals) assigned.**

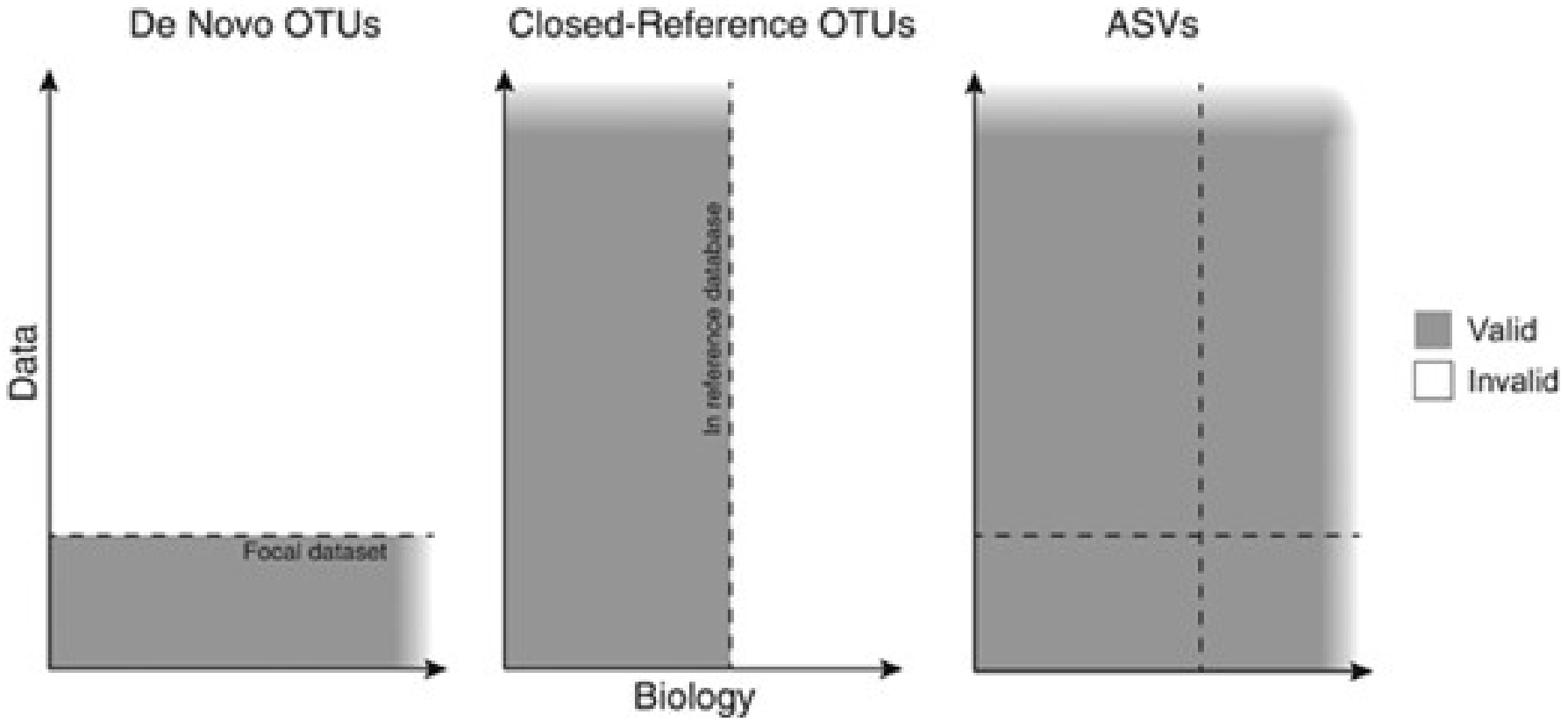
# OTU vs. ASV


sample sequences
amplicon reads
OTUs
Errors
DADA2
Make OTUs

# OTU vs. ASV

| | ASVs | De novo | Closed-ref |
|---|---|---|---|
| Precise | ✓ | ~ | ~ |
| Tractable | ✓ | ~ | ✓ |
| Reproducible | ✓ | X | ✓ |
| Comprehensive | ✓ | ✓ | X |

# OTU vs. ASV


De Novo OTUs
Closed-Reference OTUs
ASVs
Data
Focal dataset
In reference database
Biology
Valid
Invalid

# OTU vs. ASV

| OTU | ASV |
|---|---|
| Can be subject to reference bias | Reference is not used until taxonomy assignment |
| OTU tables cannot be combined between studies | ASV tables can be compared across studies |
| Represented by a consensus sequence | Represented by an exact sequence |
| Can represent multiple species with different sequences | If it represents multiple species, it is because they share the sequence |
| Subject to chimeric sequences | Subject to chimeric sequences |
| Chimera detection can be complex and may require reference bias | Chimera detection is simple and reference-free |

## α-diversity: within a sample

**Chao1** richness index is a estimation of species abundance taking into account rare sequences.

$$S_{chao1} = S_{obs} + \frac{n_1 (n_1 - 1)}{2 (n_2 + 1)}$$

**Shannon** diversity index is related to the effective species number given an OTU definition

$$\hat{H}_{shannon} = \sum_{i=1}^{S_t} \frac{\hat{C}\pi_i \ln \left( \hat{C}\pi_i \right)}{1 - \left( 1 - \hat{C}\pi_i \right)^N}$$

## β-diversity: between samples

**Jaccard** dissimilarity index evaluate the fraction of OTUs shared by two communities

$$D_{Jaccard} = \frac{S_{AB}}{S_A + S_B - S_{AB}}$$

**Yue & Clayton theta** similarity coefficient (dissimilarity index) evaluate the fraction of OTUs shared by two communities

$$D_{\Theta_{YC}} = 1 - \frac{\sum_{i=1}^{S_T} a_i b_i}{\sum_{i=1}^{S_T} (a_i - b_i)^2 + \sum_{i=1}^{S_T} a_i b_i}$$

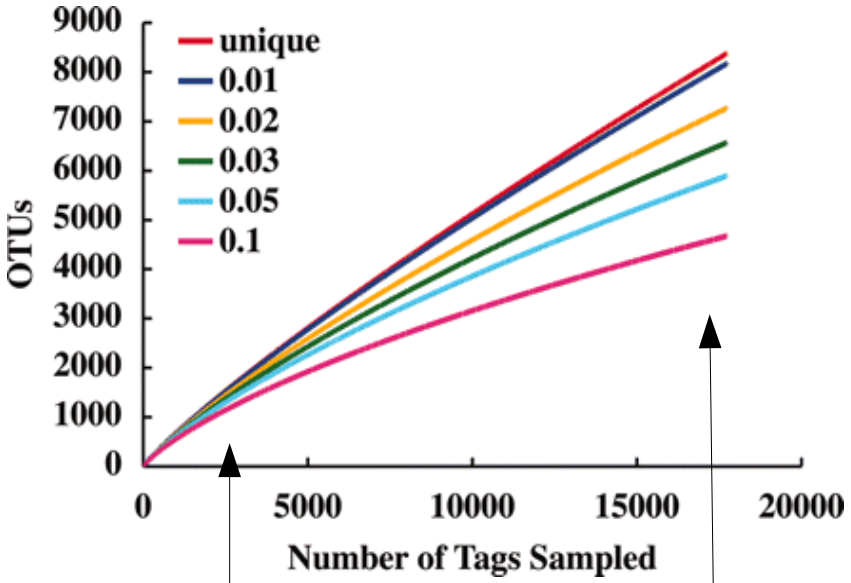
# OTU-based analysis: the rarefaction curves

Rarefaction is a technique to compare population estimators (e.g. richness). Rarefaction curves are plots of value of the estimator as a function of the number of individuals sampled.

To build a rarefaction curve the following procedure is used:

- given N sequences (the sample) and
- an initial richness value of R=0:

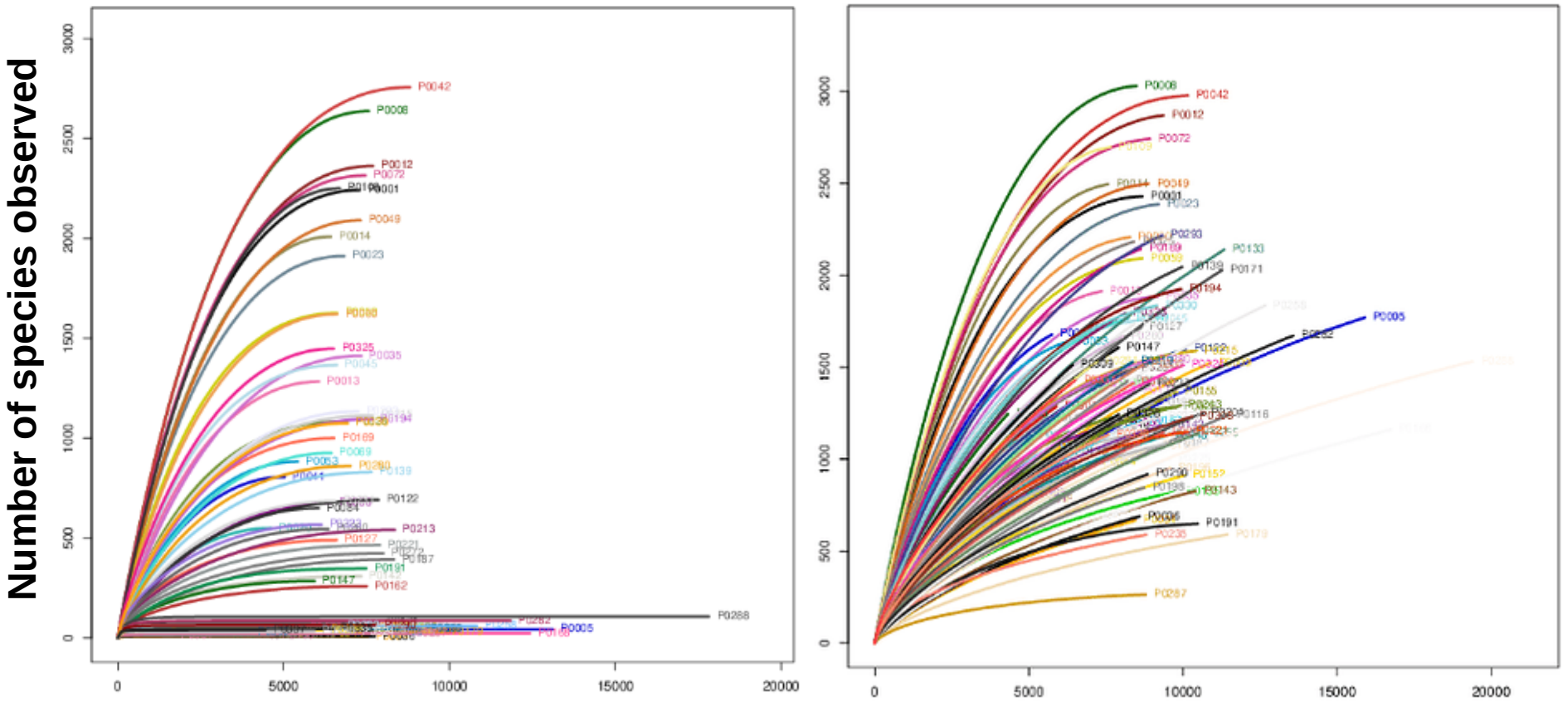
1. Take a random sample with n% individuals
2. Calculate richness R'
3. Accumulate richness with R = R+R'
4. go to 1.




On the left the steep slope indicates that a large fraction of the species diversity remains to be discovered.

On the right, a flat curve indicate richness saturation and a growing curve indicate an insufficient sampling.

# Importance of saturation in ecological analyses



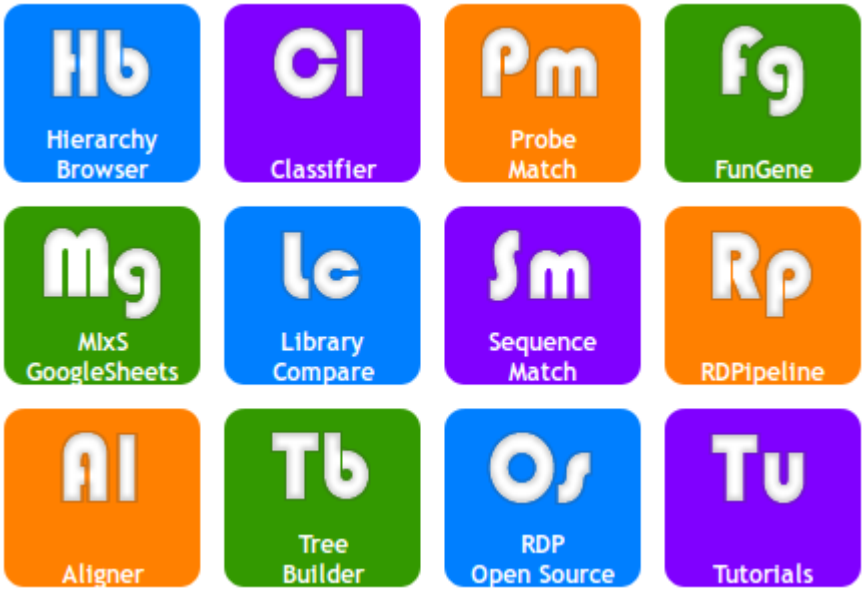

Complete !!!

Insufficient !!!

# Taxonomy and the RDP database


RDP Release 11, Update 4 :: May 26, 2015

3,224,600 16S rRNAs :: 108,901 Fungal 28S rRNAs

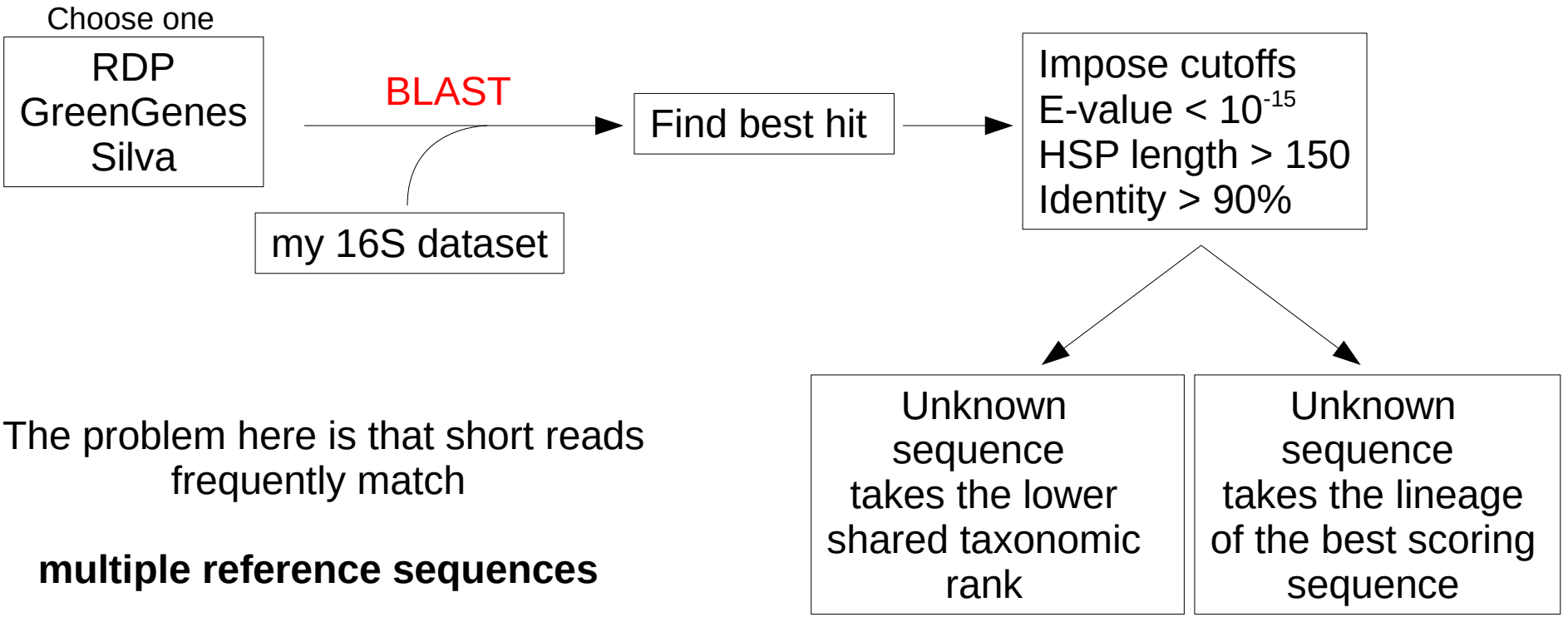



RDP offers a number of online tools for the analysis of 16S reads, from classification to tree building and probe design.

Each sequence is framed into its taxonomic assignment using infernal SS aligner.
The taxonomy is kept consistent by periodical revision.

# Taxonomic assignment with BLAST

BLAST allows a fast and efficient euristic search into 16S rDNA databases. The best-hit method is usually applied.




The problem here is that short reads frequently match

**multiple reference sequences**

with identical scores, so the degree and confidence of the assignment depend on the the taxonomic spread of the matches.

# Taxonomic assignment with k-mer search

One of the simplest and fastest way or searching in a database is to pre-index it using words of defined length. Length should be optimized for sensitivity and speed.

Given the size length of the word (k) the search is space is $4^k$ .

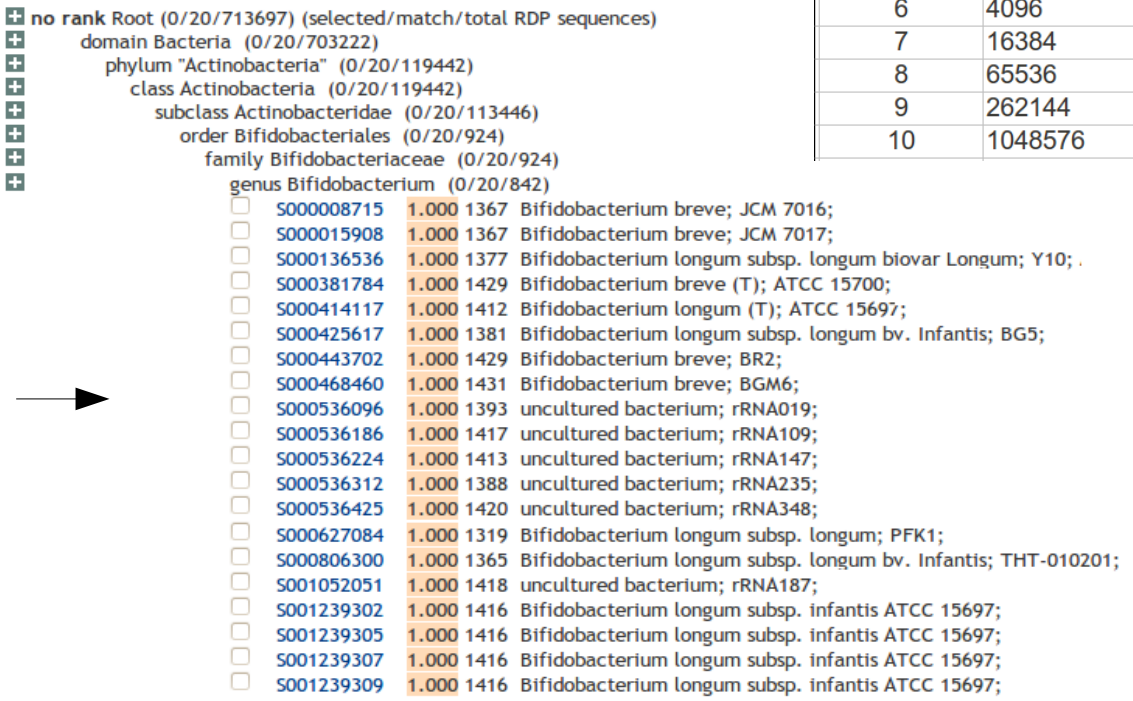
| Length | Space |
|---|---|
| 1 | 4 |
| 2 | 16 |
| 3 | 64 |
| 4 | 256 |
| 5 | 1024 |
| 6 | 4096 |
| 7 | 16384 |
| 8 | 65536 |
| 9 | 262144 |
| 10 | 1048576 |

Each sequence of a database of 16S sequences

**ATACGCA**TCGACT
A**TACGCAT**CGACT
AT**ACGCATC**GACT
ATA**CGCATCG**ACT
ATAC**GCATCGA**CT
ATACG**CATCGAC**T
ATACGC**ATCGACT**

Kmer decomposition

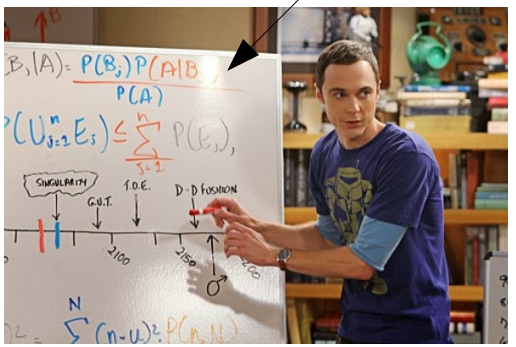
Unassigned sequence




Score database sequences by number of shared kmers or a fraction over the length

# Taxonomic assignment with naïve bayesian classifiers

Developed at RDP, is it based on the conditional probability of assigning a sequence to a rank given a set of rank specific sequences in a training set. The probability is based on the frequencies of $w_{mers}$ (optimal length 8).




Basically, we want to calculate P(G|S) i.e. the probability of having e.g. the sequence S in the genus G.

We can use the Bayes Theorem $P(G|S) = P(S|G)\frac{P(G)}{P(S)}$

Given the database, they are constant and can be removed

Given a dataset of N sequences, count w-mer occurrences (n) and derive the **prior** probability of a w-mer to occur in N.

$$P(\text{w-mer}) = \frac{n(\text{w-mer}) + 0.5}{N + 1}$$

Given a dataset of M sequences assigned to a genus G, count w-mer occurrences (m) and derive the genus conditioned probability.

$$P(\text{w-mer}|G) = \frac{m(\text{w-mer}) + P(\text{w-mer})}{M + 1}$$

Given a sequence S, count w-mer occurrences (v) and derive the genus conditioned probability of that sequence.

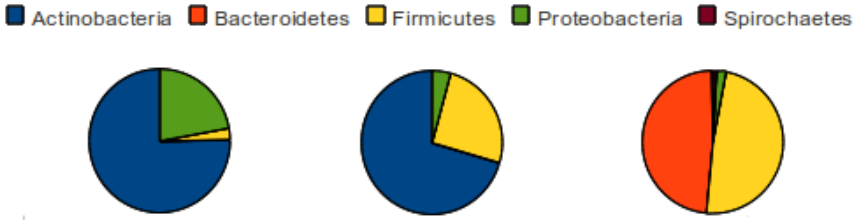
$$P(S|G) = \prod P(\text{w-mer}|G)$$

**Do this for each genus, sort by P and retain the most probable genus**

# Counts at different taxonomic ranks

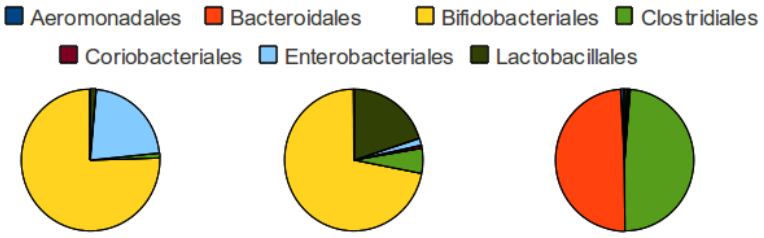
The main application of taxonomic assignment is to evaluate the samples in terms of rank counts. When done for all the ranks, one can appreciate the actual biological variability, if any, in the sample of interest.

**Analysis at the phylum level**




| PHYLUM | BB01 | BB02 | BC01 |
|---|---|---|---|
| Total | 11391.00 | 4752.00 | 9171.00 |
| **Actinobacteria** | 75.03 | 70.35 | 0.32 |
| **Bacteroidetes** | 0.12 | 0.08 | 47.46 |
| **Firmicutes** | 2.55 | 25.23 | 47.60 |
| **Proteobacteria** | 22.18 | 4.31 | 2.15 |
| **Spirochaetes** | 0.00 | 0.00 | 0.87 |

**Analysis at the order level**




| ORDER | BB01 | BB02 | BC01 |
|---|---|---|---|
| Total | 11391.00 | 4752.00 | 9170.00 |
| Aeromonadales | 0.02 | 0.00 | 0.92 |
| Bacteroidales | 0.12 | 0.08 | 47.25 |
| Bifidobacteriales | 75.01 | 69.02 | 0.00 |
| Clostridiales | 1.11 | 5.49 | 46.76 |
| Coriobacteriales | 0.00 | 0.67 | 0.28 |
| Enterobacteriales | 21.98 | 1.47 | 0.41 |
| Lactobacillales | 1.33 | 19.30 | 0.56 |

# Limitations of the 16S marker

- ✔ A wide variability in copy number has been observed even within closely related organisms

- ✔ Different amplified regions have different classification performances.

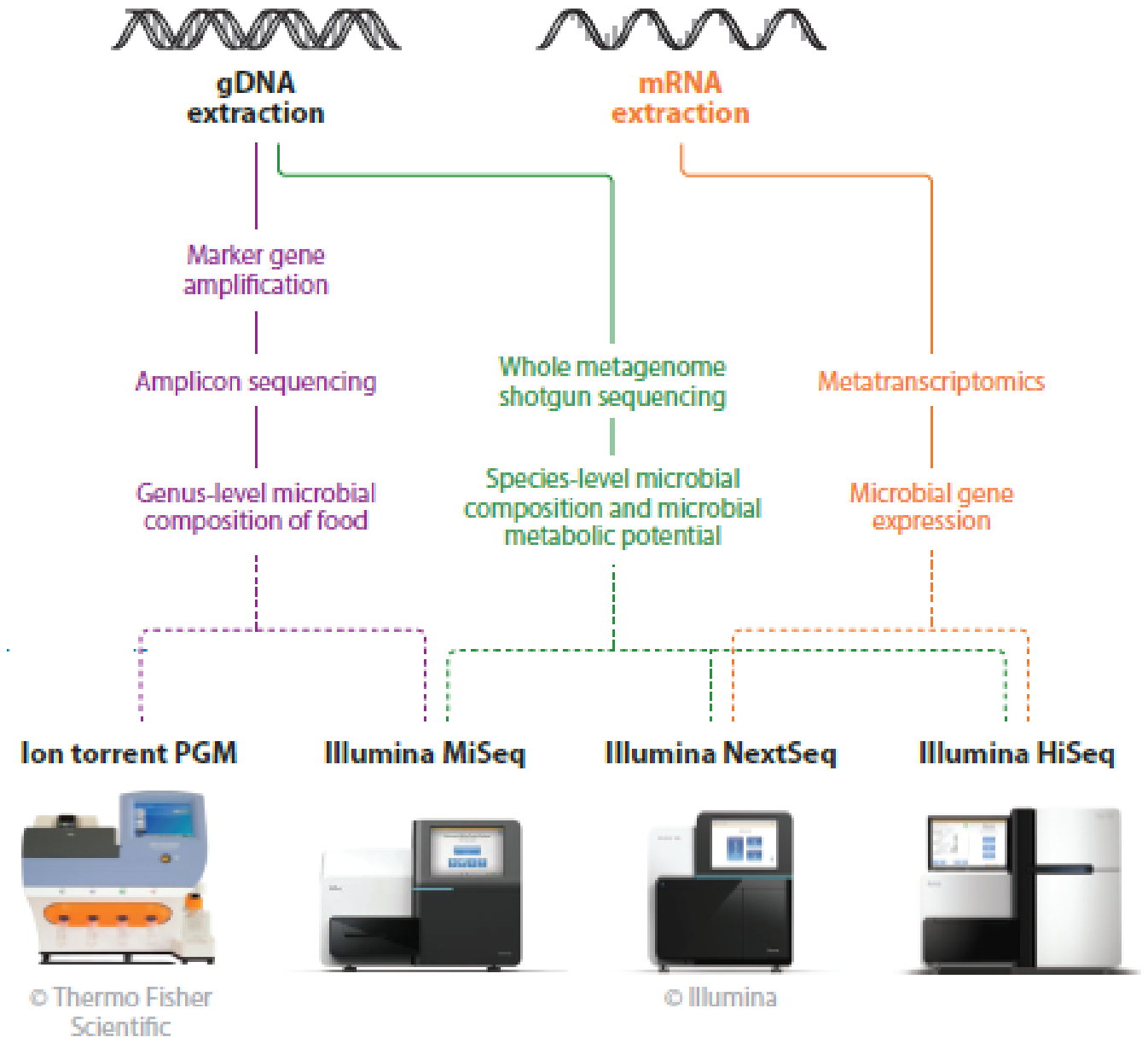
- ✔ Some primers couples covers more “unknown” organisms than other.

- ✔ It has been well documented that evolutionarily distant SSU rRNA genes that are similar in nucleotide composition have been consistently - but nevertheless incorrectly - placed close together in phylogenetic trees.

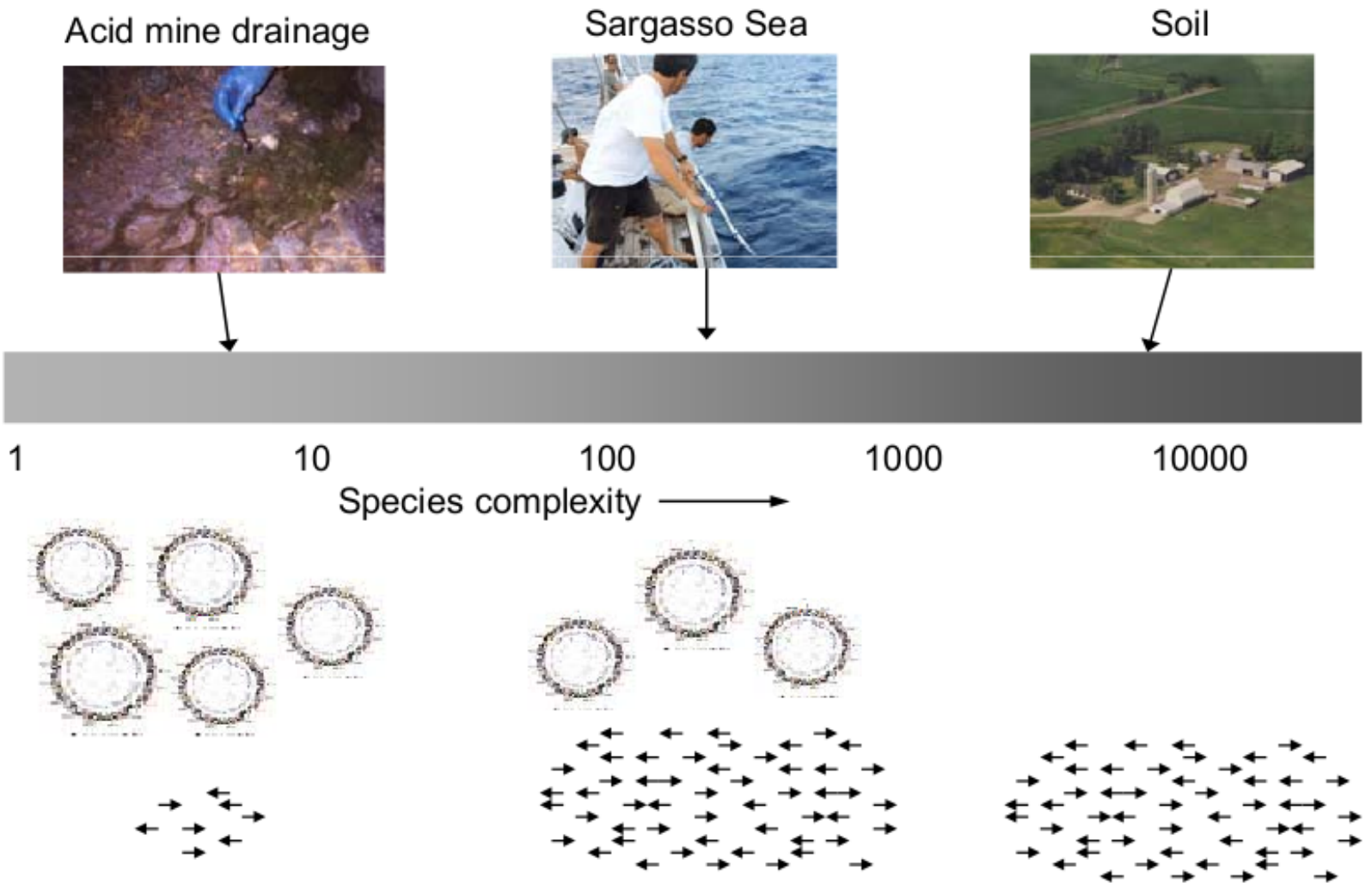
  And anyway,

- ✔ Inferring the phylogeny of organisms from any single gene carries some risks and should be corroborated by the use of other phylogenetic markers.

# Targeted vs untargeted exploration



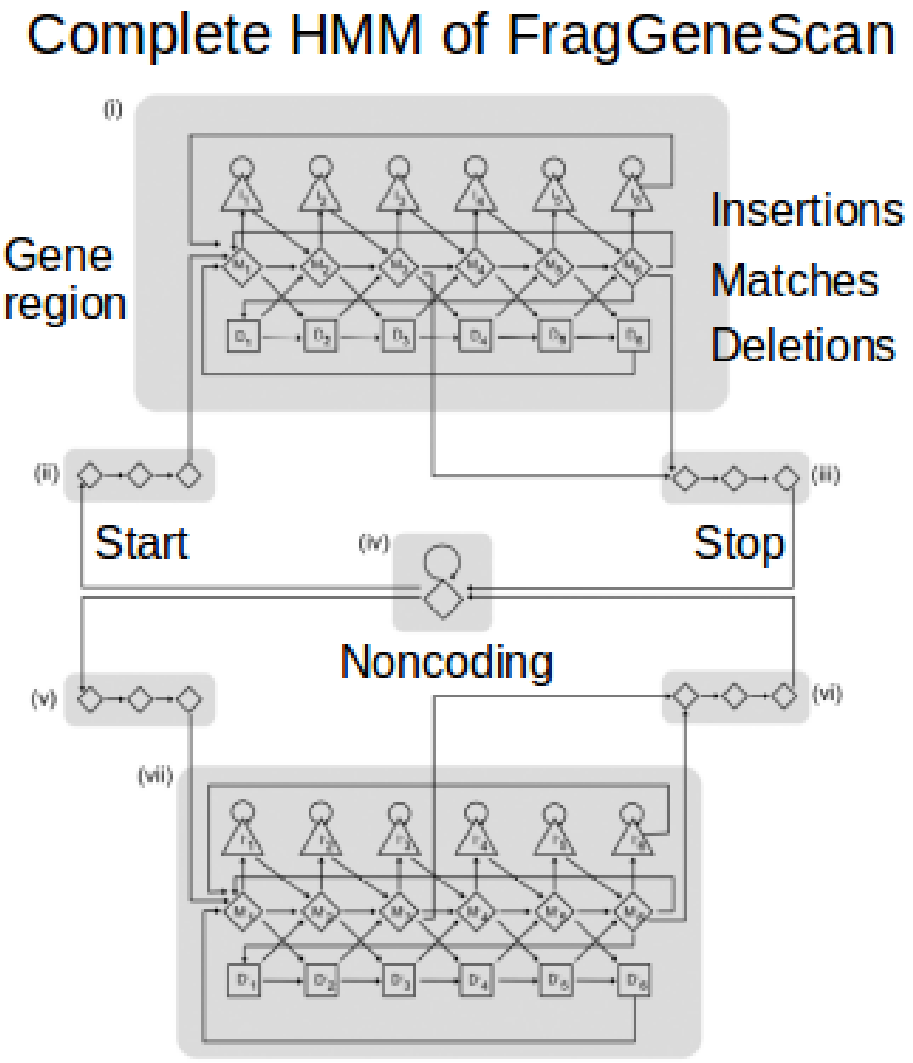
# Metagenomics, diversity and sampling efforts



# Finding genes in untargeted reads

Several programs have been developed to address the issue of finding actual genes in the context metagenomics. Such programs incorporate HMM, GC content models and k-mer composition to predict:

1. Ribosome binding sites
2. Start and stop codons
3. Composition of genes (and non-coding regions)
4. Sequencing errors




**FragGeneScan** have been specifically developed for predicting genes in short reads (< 100 bp) so it **can be used directly on raw (preprocessed) data !!!**


Noguchi et al. DNA Res. 2008, 15, 387-396
Rho M et al. Nucl. Acids Res 2010, gkq747
Zhu W et al. Nucl. Acids Res 2010, 38(12):e132
Kelley DR et al. Nucl. Acids Res 2012, 40(1): e9

# Many probable hits...should we take the top?

The lower the E-val
The better the target?

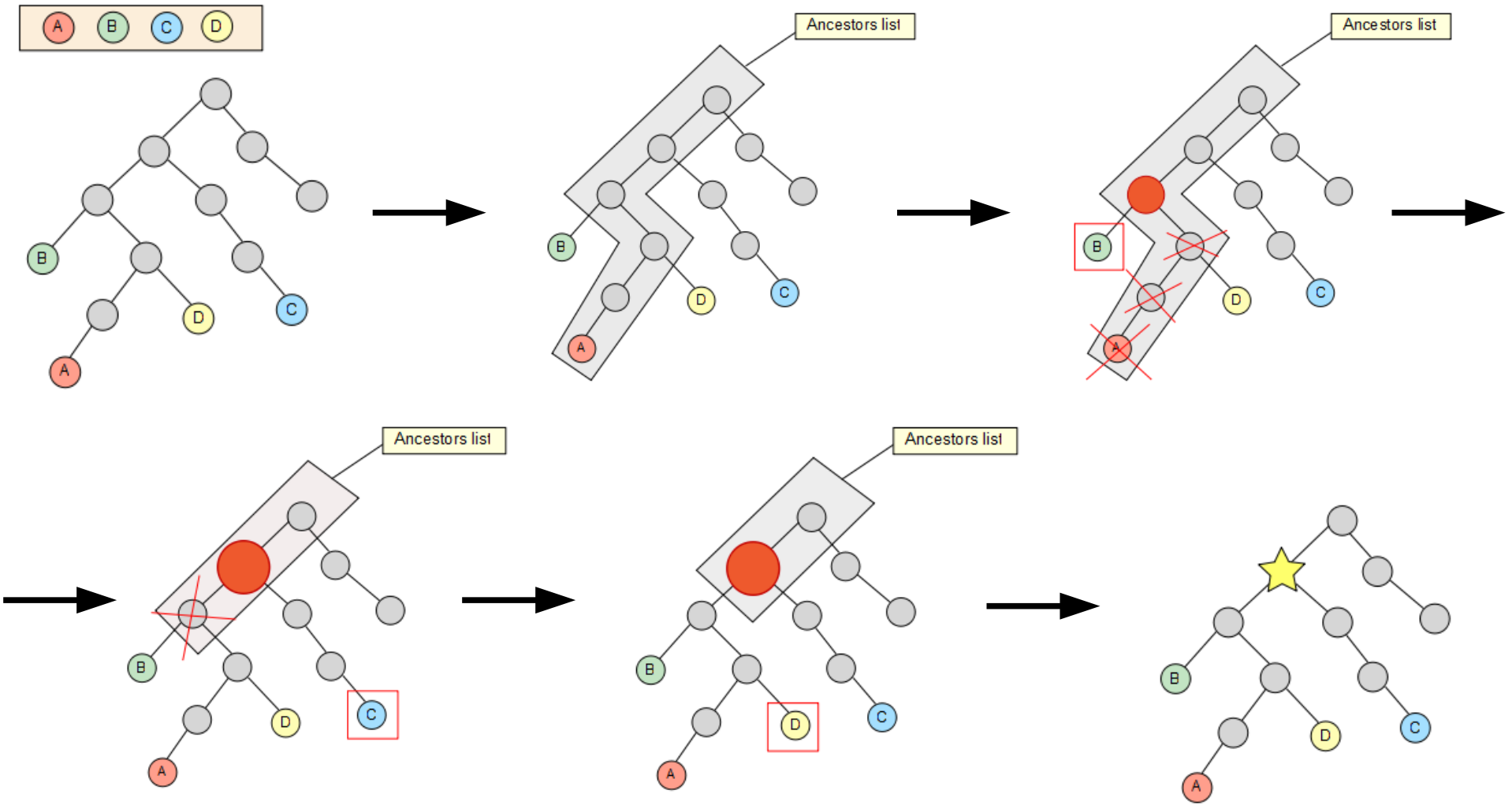
The longer the alignment
the better the target?

The higher the identity%
the better the target?

| Q.name | Q.length | E-val | A.Length | Q.start | Q.end | gaps | strand | A.id% | T.name | T.st | T.en |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HWI-ST170:227:5:1:2755-2167 | 77nt | 4.9e-24 | 76 | 77 | 2 | 0 | - | 85.53% | YP_003142869.1 | 833 | 908 |
| HWI-ST170:227:5:1:2755-2167 | 77nt | 2.3e-15 | 43 | 77 | 35 | 0 | - | 88.37% | YP_606609.1 | 776 | 818 |
| HWI-ST170:227:5:1:2755-2167 | 77nt | 8.5e-15 | 33 | 75 | 43 | 0 | - | 93.94% | YP_004033514.1 | 823 | 855 |
| HWI-ST170:227:5:1:2755-2167 | 77nt | 8.5e-15 | 33 | 75 | 43 | 0 | - | 93.94% | YP_618644.1 | 823 | 855 |
| HWI-ST170:227:5:1:2755-2167 | 77nt | 8.5e-15 | 33 | 75 | 43 | 0 | - | 93.94% | YP_812565.1 | 823 | 855 |
| HWI-ST170:227:5:1:2755-2167 | 77nt | 1.2e-13 | 43 | 77 | 35 | 0 | - | 86.05% | YP_001667032.1 | 776 | 818 |
| HWI-ST170:227:5:1:2755-2167 | 77nt | 1.2e-13 | 40 | 75 | 36 | 0 | - | 87.50% | YP_001093961.1 | 763 | 802 |
| HWI-ST170:227:5:1:2755-2167 | 77nt | 6.6e-12 | 31 | 77 | 47 | 0 | - | 90.32% | YP_002354340.1 | 788 | 818 |
| HWI-ST170:227:5:1:2755-2167 | 77nt | 3.6e-10 | 31 | 77 | 47 | 0 | - | 87.10% | YP_002435333.1 | 803 | 833 |
| HWI-ST170:227:5:1:2755-2167 | 77nt | 1.7e-12 | 41 | 75 | 35 | 0 | - | 85.37% | YP_003912539.1 | 808 | 848 |
| HWI-ST170:227:5:1:2755-2167 | 77nt | 1.2e-13 | 31 | 77 | 47 | 0 | - | 93.55% | YP_004195006.1 | 806 | 836 |
| HWI-ST170:227:5:1:2755-2167 | 77nt | 3.6e-10 | 31 | 77 | 47 | 0 | - | 87.10% | YP_004368623.1 | 728 | 758 |
| HWI-ST170:227:5:1:2755-2167 | 77nt | 3.6e-10 | 31 | 77 | 47 | 0 | - | 87.10% | NP_903294.1 | 755 | 785 |

Targets are different, and probably in different organisms...but what if they have all the same function?

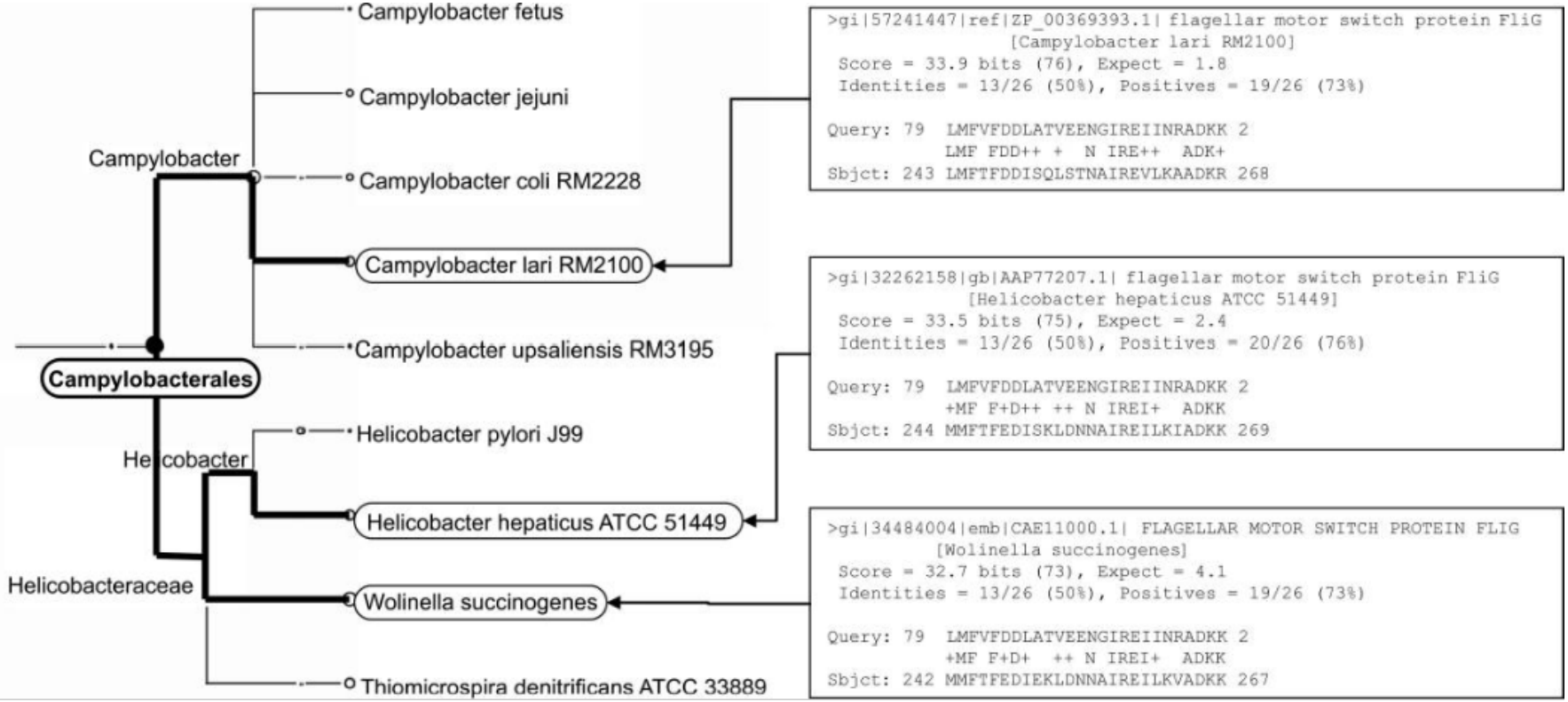
# The LCA (Lowest Common Ancestor) algorithm




This algorithm is of paramount importance in metagenomics since it allows to attribute functions and taxonomies to ambiguous assignments, that are frequent due to the shortness of the reads and orthology of genes.

# Taxonomy from untargeted

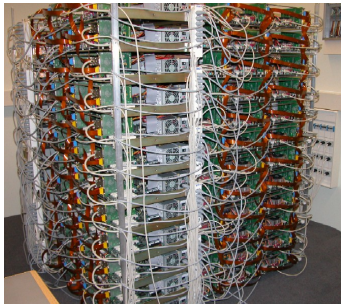
MEGAN (now in its 6th version) analyzes BLAST-like output for each read to establish a taxonomy of the read, and then show the result in the taxonomy tree. Multiple samples can also be visualized.




LCA can be applied every time we need to quantify something that can be placed in a predefined, hierarchical stricture such as taxonomy.

Huson DH et al Genome Research 2011, 21:1552-1560
Mitra S et al Bioinformatics 2009, 25:1849–1855

# Exponential need of calculation power



**Other alternatives?**

# Taxonomic markers from multiple alignments

We all know that proteins can be grouped in many different ways according to function, structure, orthology. In general, sequences that share common features can be represented as multiple sequence alignments.

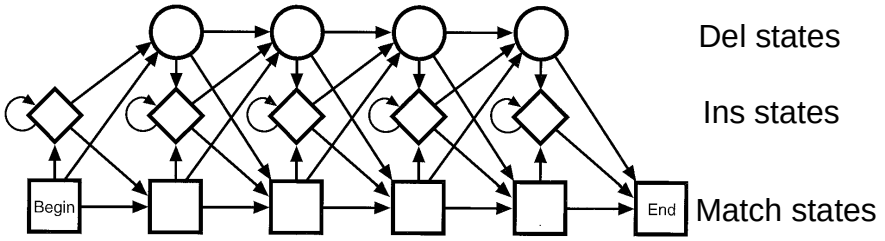

A large number of databases exist with different scopes that collect MSAs.

```
                    10        20        30        40        50        60
                    ....*....|....*....|....*....|....*....|....*....|....*....|.
Feature 1             #  #              #  #           #  #              #   #
1TOT_A          7 YTCNECKh--HVe---TRWHCTVCe------DYDLCINCYNTk-------SH--THKMVKW 47
gi 7023094     95 ISCDGCDe--IAp--wHRYRCLQCs------DMDLCKTCFLGgvk---peGHgdDHEMVNM 142
gi 50750334   201 VRCRVCKtfpITg---LRYRCLKCl------NFDLCQVCFFTgrh---skPHksSHPVVEH 249
gi 50257626   582 SECTICLtalFS----NRFKCVSCp------KFDLCRSCYQKvd-----eIHp-AHAFLSL 626
gi 47222763    98 IICDSCKkhgIMg---MRWKCKVCf------DYDLCTQCYMNn-------KHdlSHAFERY 142
gi 40743717 1022 RVCNNCLk-eFDe--gKMVSCADCd------DFDLCITCILGhk----hgHHp-SHTFVLL 1068
gi 51261627    15 PPCKGCSs-yLMe---PYIKCAECgp----pEFLLCLQCFSGfe----ykKHqsDHSYEIM 63
gi 16944480   367 RTCNCCIq-dLPe--aEFVHCQTCd------DFDLCKVCFAKnr-----hGHhpKHAFSPI 413
gi 42546497   336 RTCNCCVq-eHPe--aEFLHCRMCe------DFDLCQSCFARds-----hGHhpKHSFAPA 382
gi 40745179   373 IICDGCNaegLA----VQYHCADCe------DYDLCQSCYKAgtrc-gykGHt-YHLEFNA 421
gi 49648418   505 FVCDYCLe-pISe---ARFHCQSCv------DFDVCSSCYPSra-----kSHaqKHGFVQL 550
gi 56512242    87 YLCDECQy-aIMpl-sYVFECNTCe------NFTLCKKCFKKg-------KH--EHPLKKM 130
gi 42555336  1607 SFCDACLm-nNYg---LRFHCDVCw------DFDLCFKCYRSqn-----iIHp-KHSFTNP 1651
gi 47207916   134 VTCDGCEg-pVVg---TRFKCSVCp------NYDLCSACQAKg-------THt-EHPLLPI 176
gi 42551863  1612 MFCDCCLl-dIYg---FYYTCSTCf------ECDLCDKCYLSvs-----kIHp-AHSSFMK 1656
gi 40744696  1669 YSCDICLs-tIWg---PVYECETCl------DFTACKKCHGRin-----lYH--GHLRLEN 1712
gi 30725524   289 IRCDGCGvlpITg---PRFKSKVKe------DYDLCTICYSVm-------GN--EGDYTRM 331
gi 7512750    167 FKCDNCGiepIQg---VRWHCQDCppemsldFCDSCSDCLHEt------dIHkeDHQLEPI 218
gi 7023094    144 FTCDHCQg-lIIg---RRMNCNVCd------DFDLCYGCYAAkky---syGHlpTHSITAH 191
gi 47223744  1889 YACDHCQg-lIVg---SRINCNVCe------DFDLCFGCYNAkky---pdSHlpTHRITVY 1936
gi 60467539  3292 FSCDLCNinpITg---KRWNCSNCg------DFDLCNQCYQNpe-----kDHpkDHIFKEF 3338
gi 30678519  2616 YCCDGCStvpILr---RRWHCTVCp------DFDLCEACYEVldadrlppPHtrDHPMTAI 2667
gi 50758066  1840 YACDHCQg-vIIg---RRMNCNVCd------DFDLCYGCYSAkky---sdSHlpTHSITVY 1887
gi 4758520   2706 VTCDGCQmfpINg---SRFKCRNCd------DFDFCETCFKTk-------KHntRHTFGRI 2750
gi 60465335    40 YSCNGCGs-eIWppkqERYACNECs------NFDLCSECYRKem----ilINgtQEEKDKL 89
gi 17530949   281 NSCAGCRkehIVg---IRFRCQVCr------DISLCLPCFAVgfa---ggRHepGHRMCEV 329
gi 1176014    258 YICHTCGn-eSIn---VRYHNLRAr------DTNLCSRCFQEgh----fgAN--FQSSDFI 302
gi 46852178     7 VSCDACLkgnFRg---RRYKCLICy------DYDLCASCYESgat---ttRHttDHPMQCI 55
```

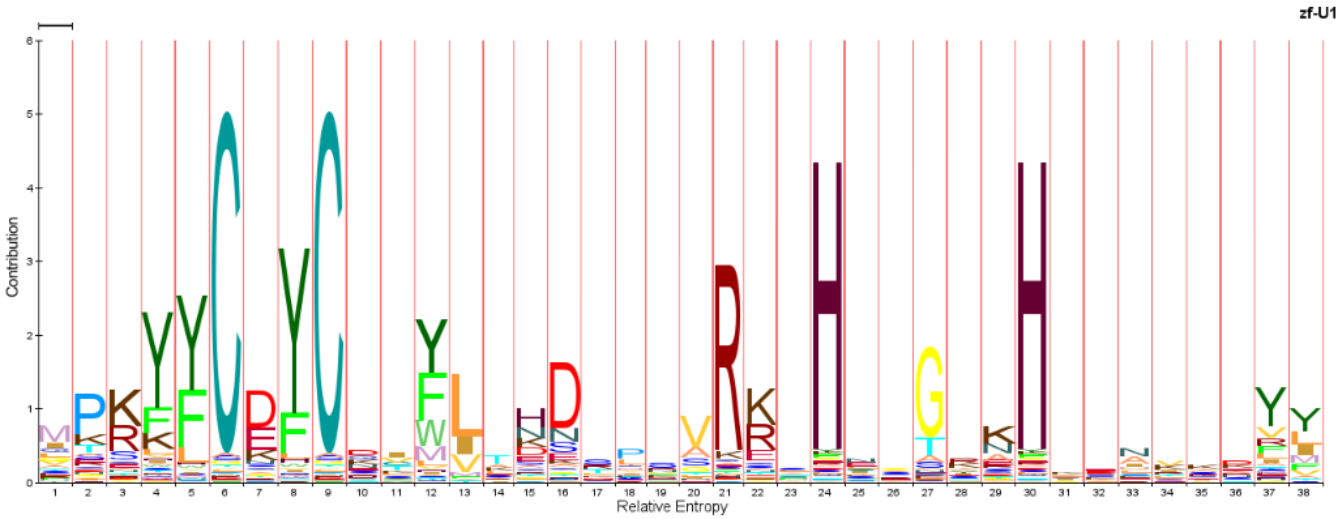
Multiple Sequence Alignments allow to catch sequence conservation in the aligned sequences and identify columns that contribute to define the reason for which they are grouped together.

# Summarizing a marker: HMM

```
RU1C_HUMAN/1-38       MPKFYCDYCDTYLTHDSPSVRKTHCSGRKHKENVKDYY
RU1C_HUMAN/1-38 (SS)  --S-B-TTT--B-S--SHHHHHHHT---HHHHHHHHHT
P90815_CAEEL/1-38     MPKYYCDYCDTFLTHDSPSVRKTHNGGRKHKDNVRMFY
Q9ZR19_ARATH/1-38     MPRYYCDYCDTYLTHDSPSVRKQHNAGYKHKANVRIYY
WBP4_MOUSE/8-44       QPKKFCDYCKCWIADNRPSV.EFHERGKNHKENVARRI
Q8IAW3_PLAF7/1-38     MPKYYCEYCDIYLTHSSPVGRRQHIHGRKHISAKIEYF
YNS7_SCHPO/8-44       IPKYYCKYCQIFV.KDTPFARRSHEQTYKHQDAIKKVM
RU1C_YEAST/1-38       MTRYYCEYCHSYLTHDTLSVRKSHLVGKNHLRITADYY
RU1C_SCHPO/1-38       MPRYLCDYCQVWLTHDSQSVRKAHNAGRAHIQNVQDYY
Q9HEK5_NEUCR/1-38     MPKFFCDYCDVYLTHDSMSVRKAHNSGRNHLRNVVDYY
Q8IPW7_DROME/3-39     GKSYYCDYCCFLKNDL.NVRKLHNGGIAHAIAKSNYL
Q8SR72_ENCCU/1-38     MPRYFCEFCNKMLLNDRLRSRRMHCGGAKHGLMRKAYY
```



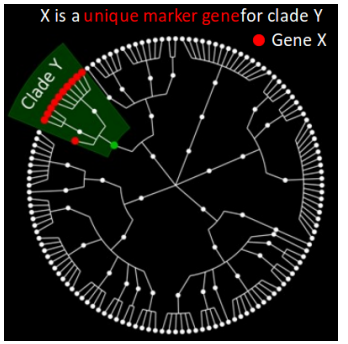

Model positions trough a series of interconnected states with pre-computed emission probabilities

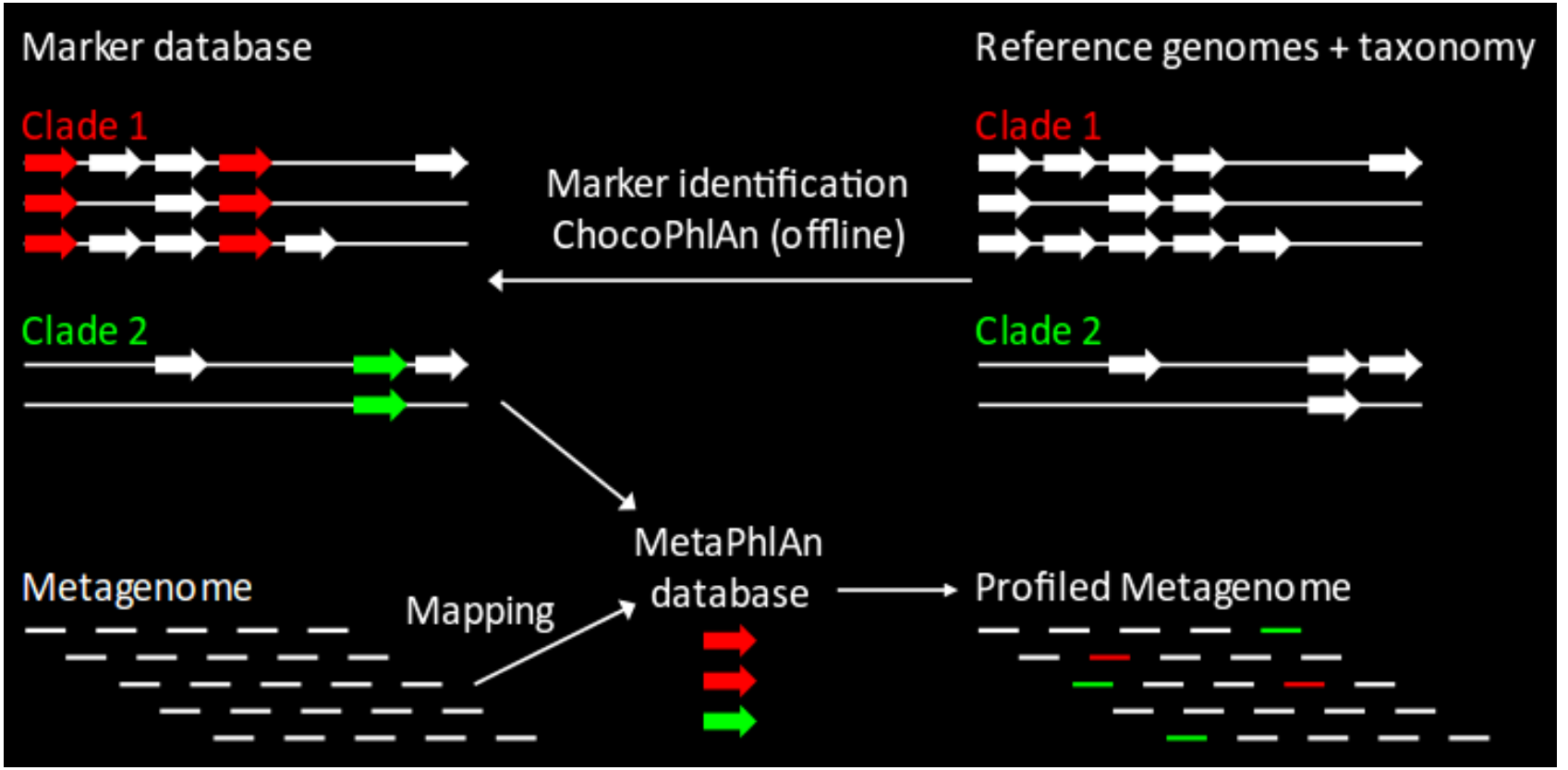



http://pfam.sanger.ac.uk/

# Clade specific marker genes: improved taxonomy

MetaPhlAn2 (2012) relies on ~1M unique clade-specific marker genes pre-identified from ~17,000 reference genomes (~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic), allowing to compute abundance at the “species”-level for bacteria, archaea, eukaryotes and viruses.



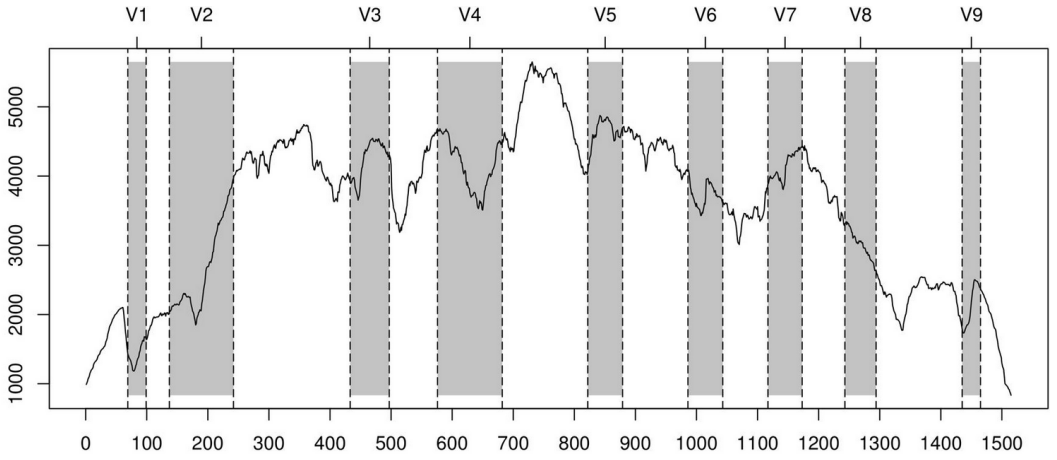

Segata Net al. Nat Methods. 2012, 9: 811-814

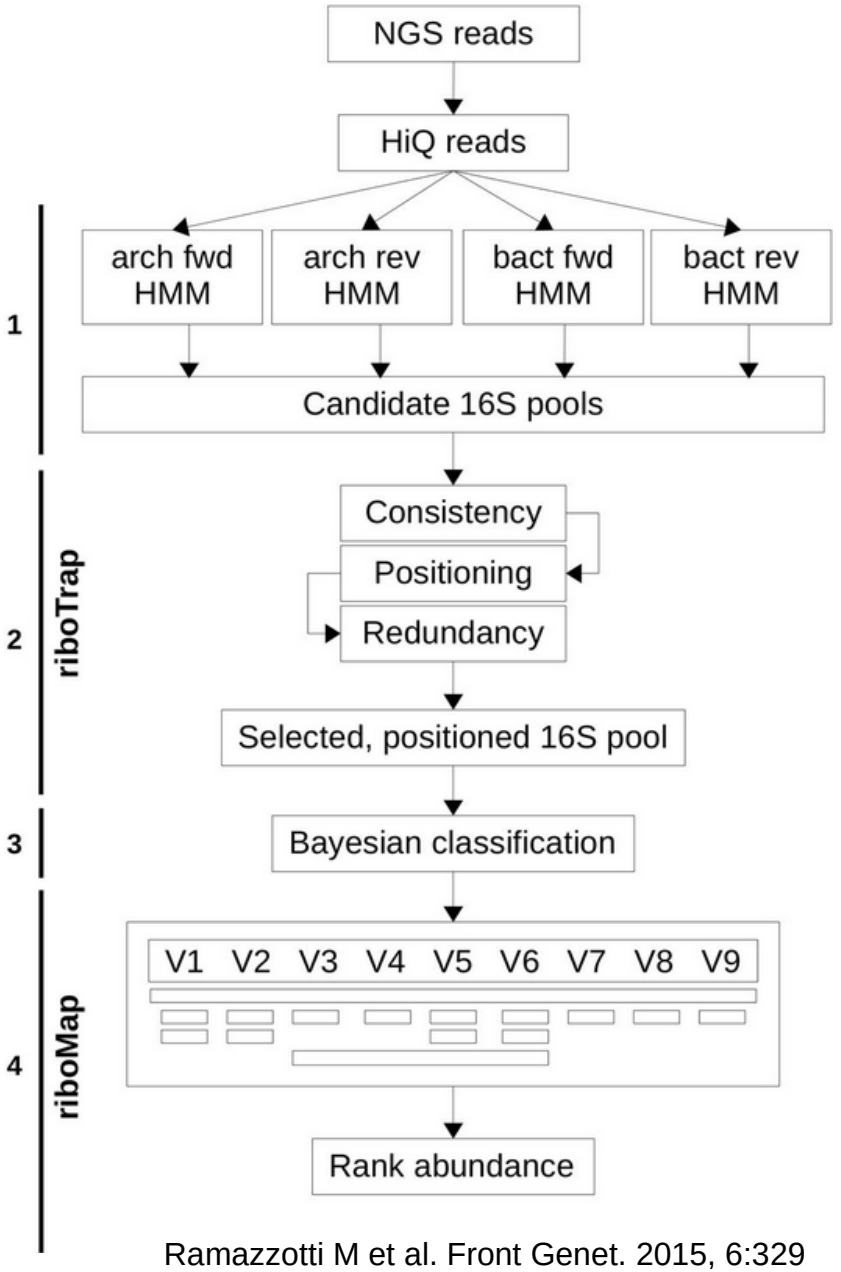
# Using 16S in non-targeted metagenomics

RiboFrame solves the issue of bias-by-amplification when using 16 rDNA.

Reads overlapping the 16S rDNA genes are identified using Hidden Markov Models and a taxonomic assignment is obtained by naïve Bayesian classification.

All reads identified as ribosomal are coherently positioned in the 16S rDNA gene, allowing the use of the topology of the gene (i.e., the secondary structure and the location of variable regions) to guide the abundance analysis.

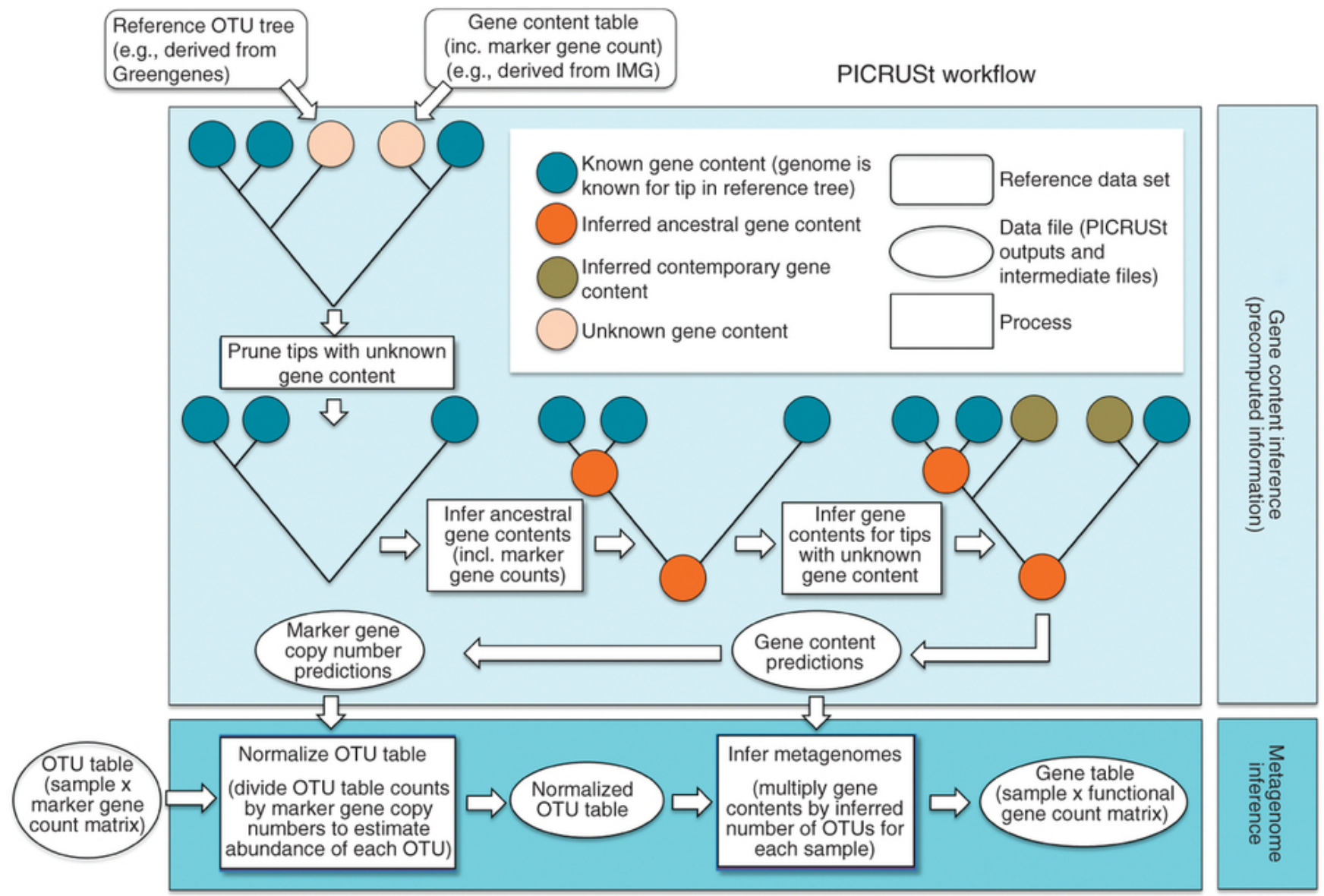



Ramazzotti M et al. Front Genet. 2015, 6:329

# Estimating community functions from genes

# Estimating community functions from taxonomy




Langille MG et al. Nat Biotechnol. 2013, 31(9):814-21

# MGnify: the EBI metagenomics resource

MGnify

| | | | |
|---|---|---|---|
| 1763 | studies | 112898 | amplicon |
| 109884 | samples | 7026 | assemblies |
| 141860 | analyses | 2035 | metabarcoding |
| | | 17961 | metagenomes |
| | | 1739 | metatranscriptom |


https://www.ebi.ac.uk/metagenomics/