Introduction to amplicon analyses

Hélène Chiapello et Anne-Laure Abraham

Mars 2016

« Meta-omics » using next-generation sequencing (NGS)

DNA

RNA

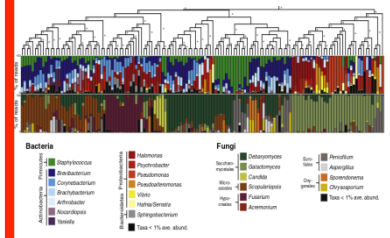
Metagenomics

Metatranscriptomics

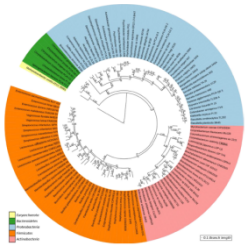
Amplicon sequencing

Shotgun sequencing

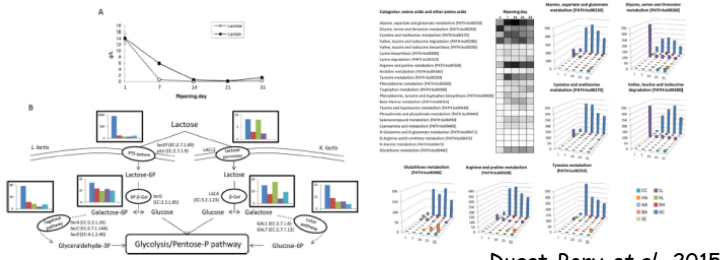
RNA sequencing



Wolfe *et al.*, 2014



Almeida *et al.*, 2014



Dugat-Bony *et al.*, 2015

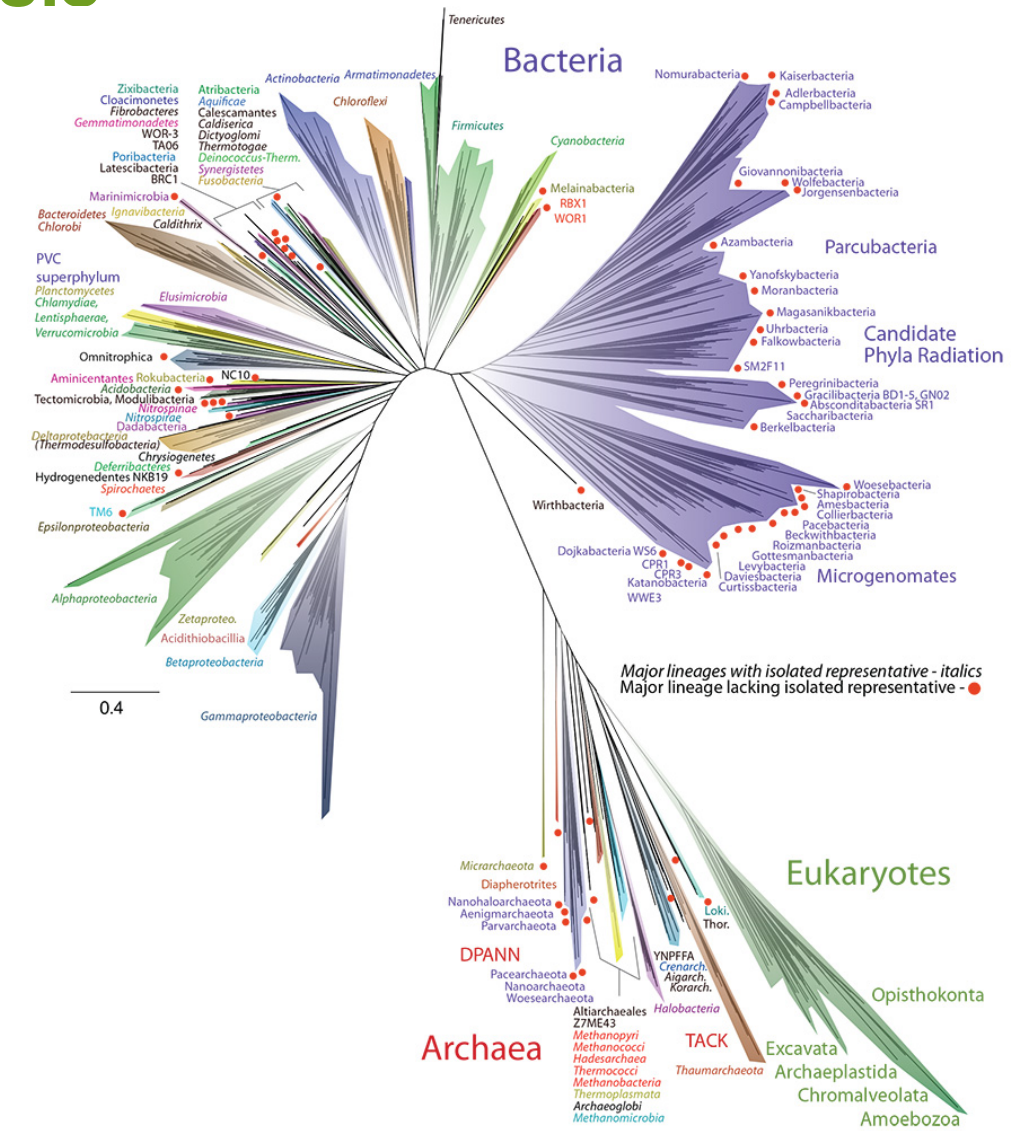
Who is here?

What can they do?

What are they doing?

Amplicon analysis

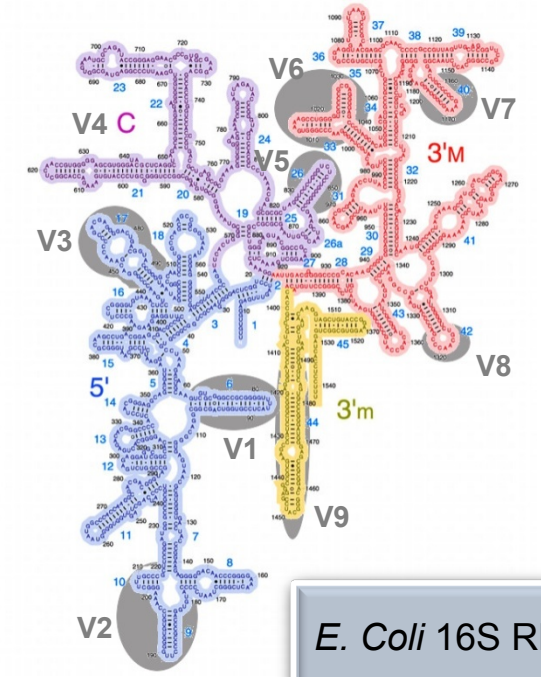
- Provides access to **uncultivated organisms**
- **Expands considerably the known tree of life** due to new genomic sampling of previously unknown microbial lineages



A new view of the Tree of Life
(Hug *et al*. *Nature Microbiology 2016*)

The gene encoding the small subunit of the ribosomal RNA

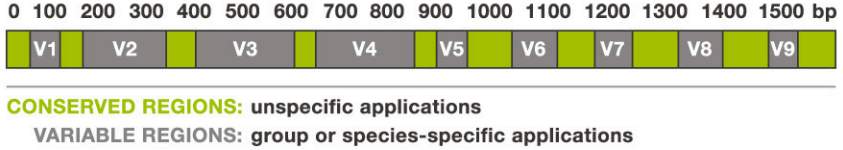
- The most widely used gene in **molecular phylogenetic** studies
- Ubiquist gene : **16S rDNA** in prokayotes ; **18S rDNA** in eukaryotes
- **Gene encoding a ribosomal RNA :** non-coding RNA (not translated), part of the small subunit of the ribosome which is responsible for the translation of mRNA in proteins
- Not submitted to lateral gene transfert
- Availability of databases facilitating comparison (Silva 2015: >22000 typestrains)



E. Coli 16S RNA

1542 bp

9 variable regions



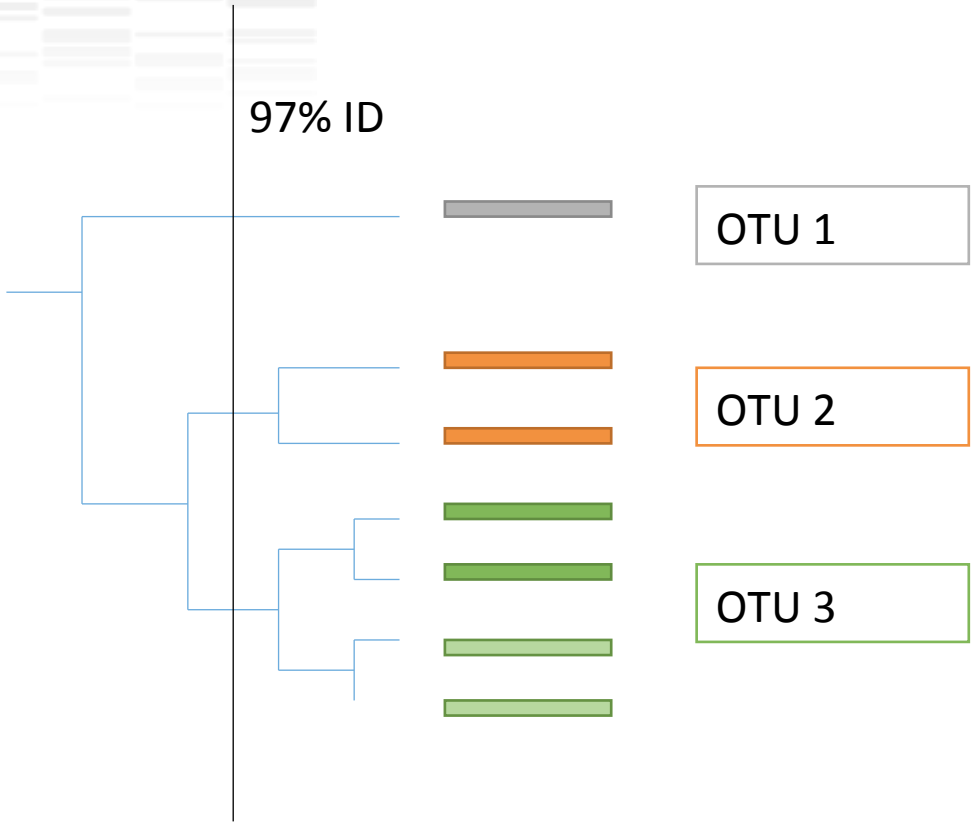
Amplification and sequencing

- **« Universal » primer sets** are used for PCR amplification of the phylogenetic biomarker

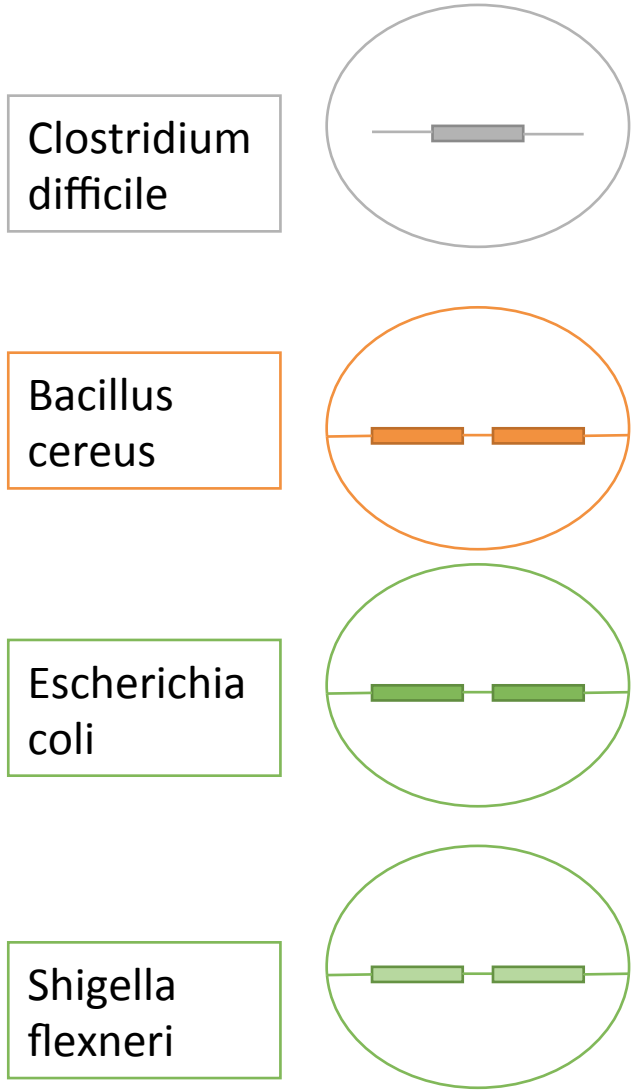
- The primers contain **adaptators** used for the sequencing step and **barcodes** (= tags = MIDs) to distinguish the samples (multiplexing = sequencing several samples on the same run)

- Sequencing is generally performed on **Roche-454** or **Illumina MiSeq** platforms. Roche-454 generally produce ~ 10 000 reads per sample, MiSeq ~ 30 000 reads per sample. Sequence length is >650 bp for pyrosequencing technology (Roche-454) and 2 x 300 bp for the MiSeq technology in paired-end mode.

Objective: identifying Operational Taxonomic Units
A proxy for bacterial species

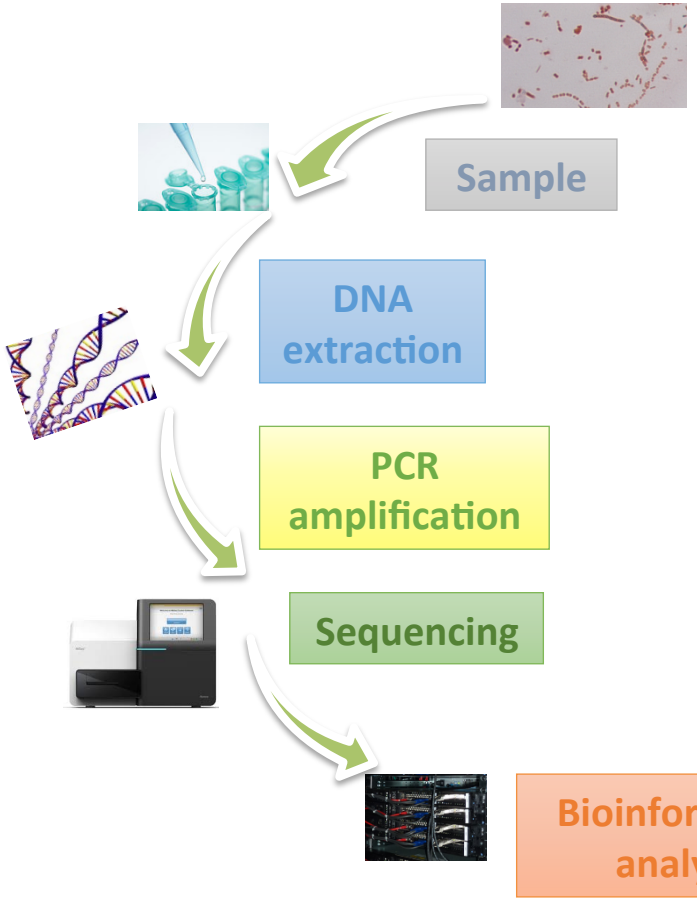


Phylogenetic tree of 16S rRNA



Identify and quantify micro-organisms

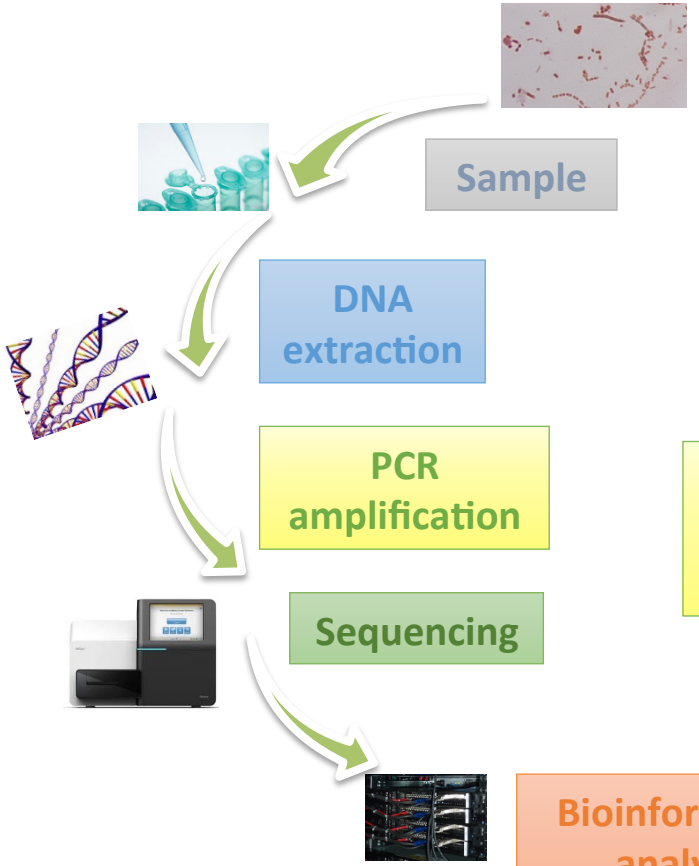
Process



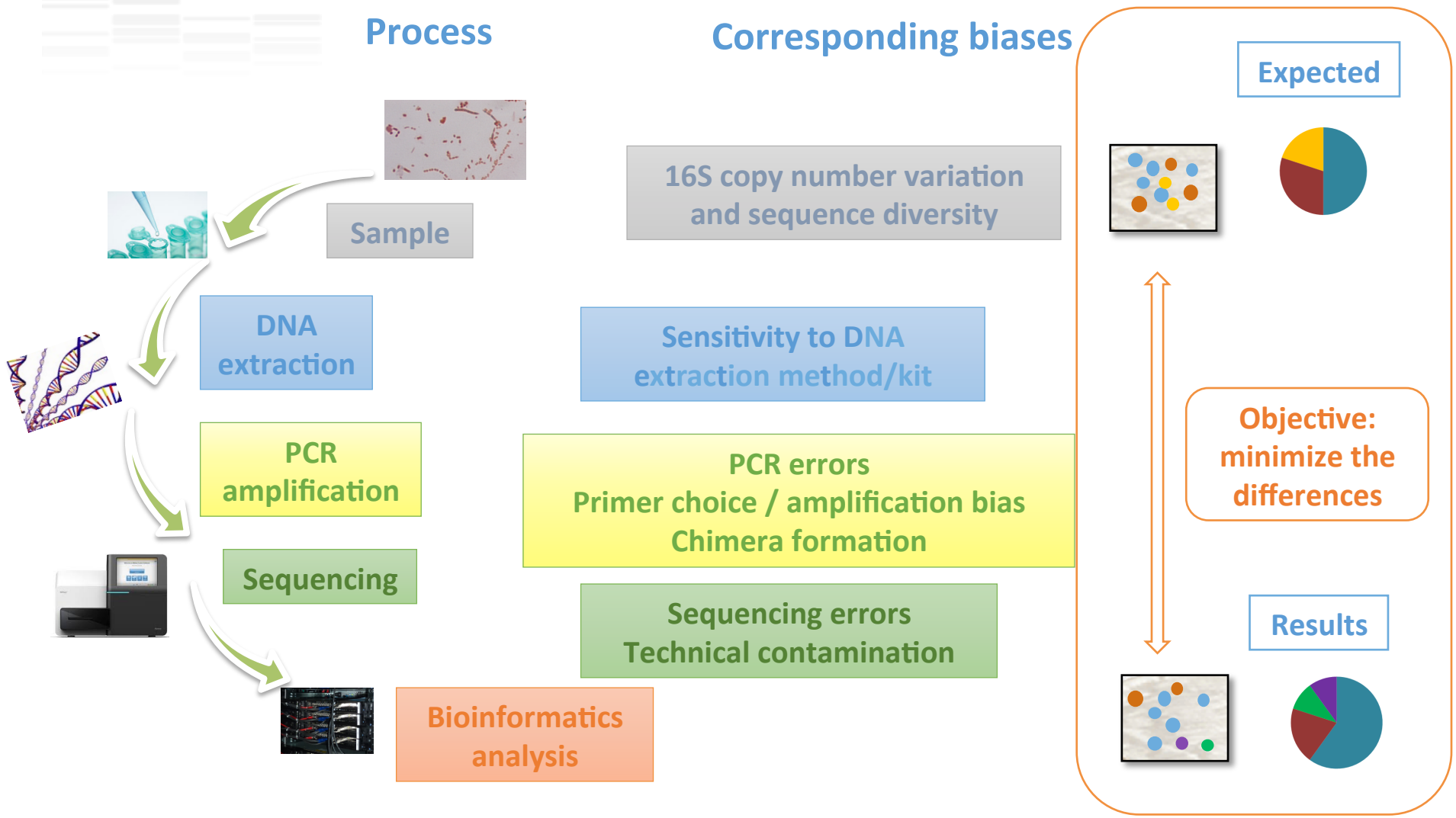
	Affiliation	Sample 1	Sample 2	Sample 3
OTU1	Species A	0	100	0
OTU2	Species B	741	0	456
OTU3	Species C	12786	45	3

Identify and quantify micro-organisms

Process	Corresponding biases
Sample	16S copy number variation and sequence diversity
DNA extraction	Sensitivity to DNA extraction method/kit
PCR amplification	PCR errors Primer choice / amplification bias Chimera formation
Sequencing	Sequencing errors Technical contamination
Bioinformatics analysis	



Identify and quantify micro-organisms



Biases

Biological biases

- Variable number of 16S gene copies
- Sequence diversity among the same organism
- Some 16S sequences are common to multiple species, and sequence diversity differs among phyla

Technical biases

- PCR error
- Sequencing error
- PCR amplification biases
- Chimera formation
- DNA extraction method/kit
- Technical contamination (between runs or inside run)
- Low DNA quantity
- DNA sequencer choice

Human biases

- Sample Contamination
- Choice of variable region for amplification
- Primer choice

Biological biases

Biological biases

- **Variable number of 16S gene copies**
- **Sequence diversity among the same organism**
- **Some 16S sequences are common to multiple species, and sequence diversity differs among phyla**

- Gene copy number spans over an order of magnitude, **from 1 to up to 15 in Bacteria, but only up to 5 in Archaea**

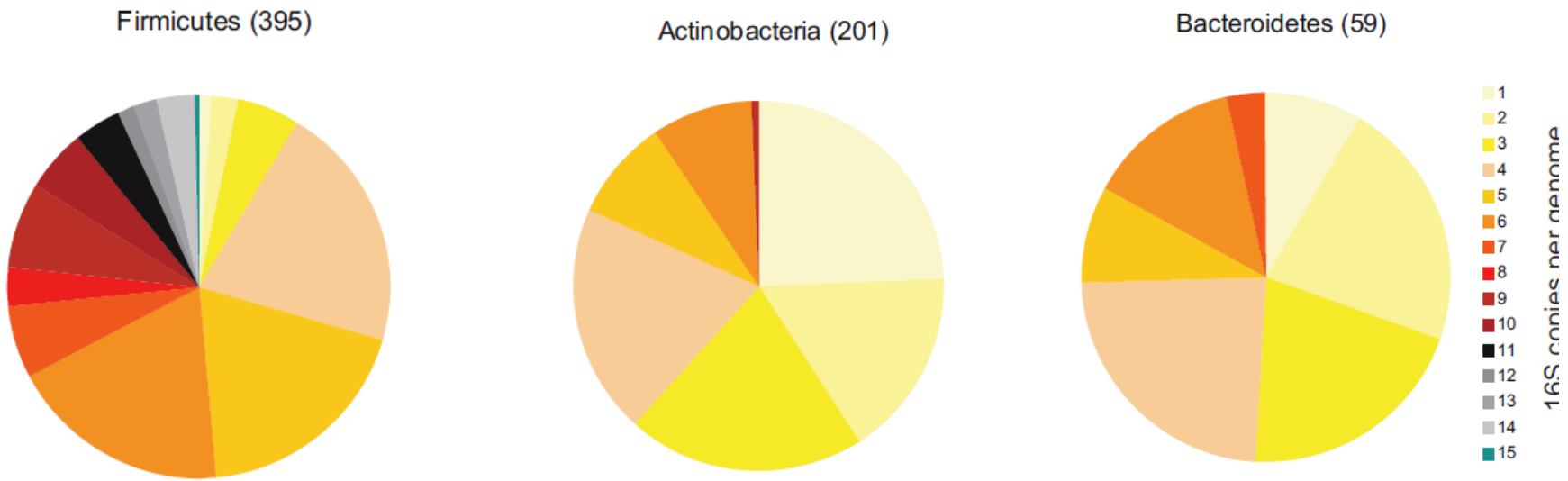
- Only a **minority** of bacterial genomes harbors **identical 16S rRNA gene copies**
- **Sequence diversity increases with increasing copy numbers.**
- While certain taxa harbor dissimilar 16S rRNA genes, others contain sequences common to multiple species.

Vetrovsky *et al.*, Plos one (2013) ; Angly *et al.*, Microbiome (2014)

Biais quantification

Biological biases

- **Variable number of 16S gene copies**
- Sequence diversity among the same organism
- Some 16S sequences are common to multiple species, and sequence diversity differs among phyla



- **CopyRighter, new software which uses these estimates to correct 16S rRNA amplicon microbial profiles and associated quantitative (q)PCR total abundance.**

Vetrovsky *et al.*, Plos one (2013)

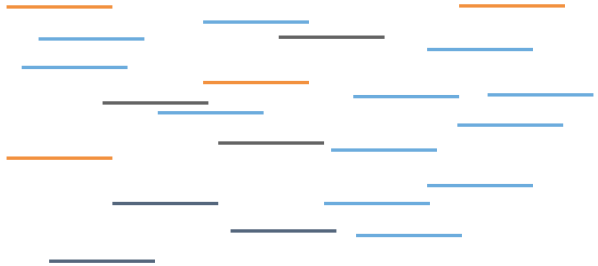
Smets *et al.*, PeerJ (2015), Angly *et al.*, Microbiome (2014)

Technical biases

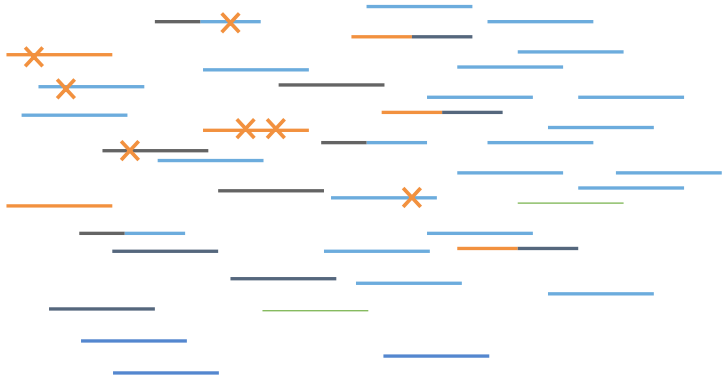
Sample

sequencing

Perfect world



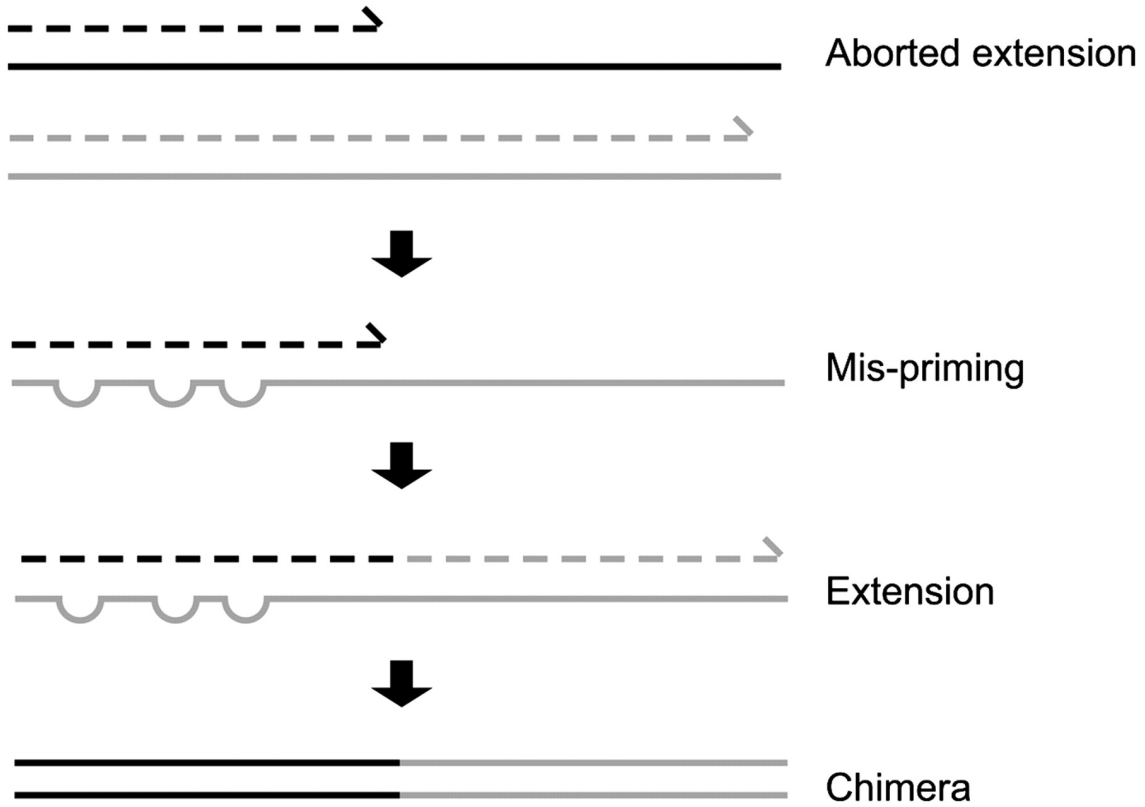
Reality



Technical biases

- PCR error
- Sequencing error
  - 454 : 0.6%
  - MiSeq : 0.3%
- PCR amplification biases
- Chimera formation
- DNA extraction method/kit
- Technical contamination (between runs or inside run)
- Low DNA quantity

Formation of chimeric sequences during PCR.



- Up to **70% of chimeric sequences** in the **unique** amplicon pool of PCR-amplified samples
- **Chimera: from 5 to 45% of reads**
- Even after treatment with traditional chimera detection tools, chimeras are continuously detected in databases like RDP, SILVA, and Greengenes.

Haas *et al.* Genome Res. (2011) Schloss *et al.*, Plos one (2011)

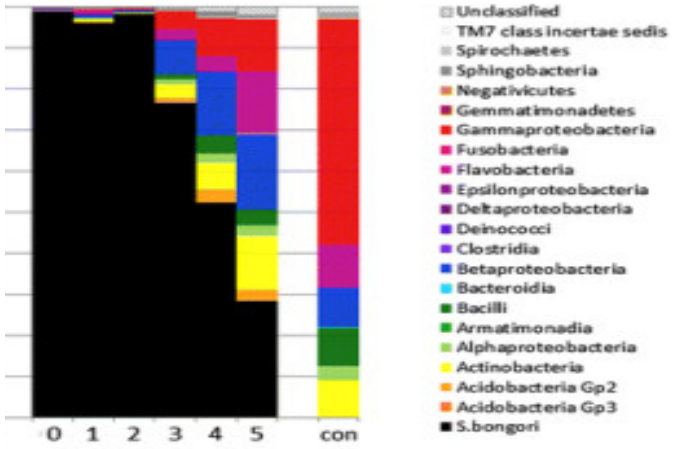
Mysara *et al.* , AEM (2014) ;)

Contamination, low DNA quantity

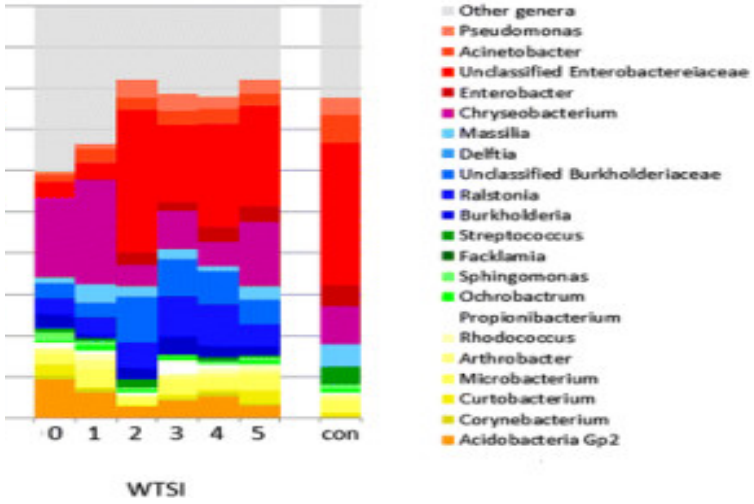
Technical biases

- DNA extraction method/kit
- Technical contamination (between runs or inside run)
- Low DNA quantity

- **Contaminating DNA is ubiquitous** in commonly used DNA extraction kits and other laboratory reagents
- Contaminating DNA varies greatly in composition between different kits and kit batches
- This contamination critically impacts results obtained from samples containing a low microbial biomass.
  They recommend at least 10^3 to 10^4 cells.



S. bongori



Without S. bongori

Salter *et al.*, BMC Biology (2014)

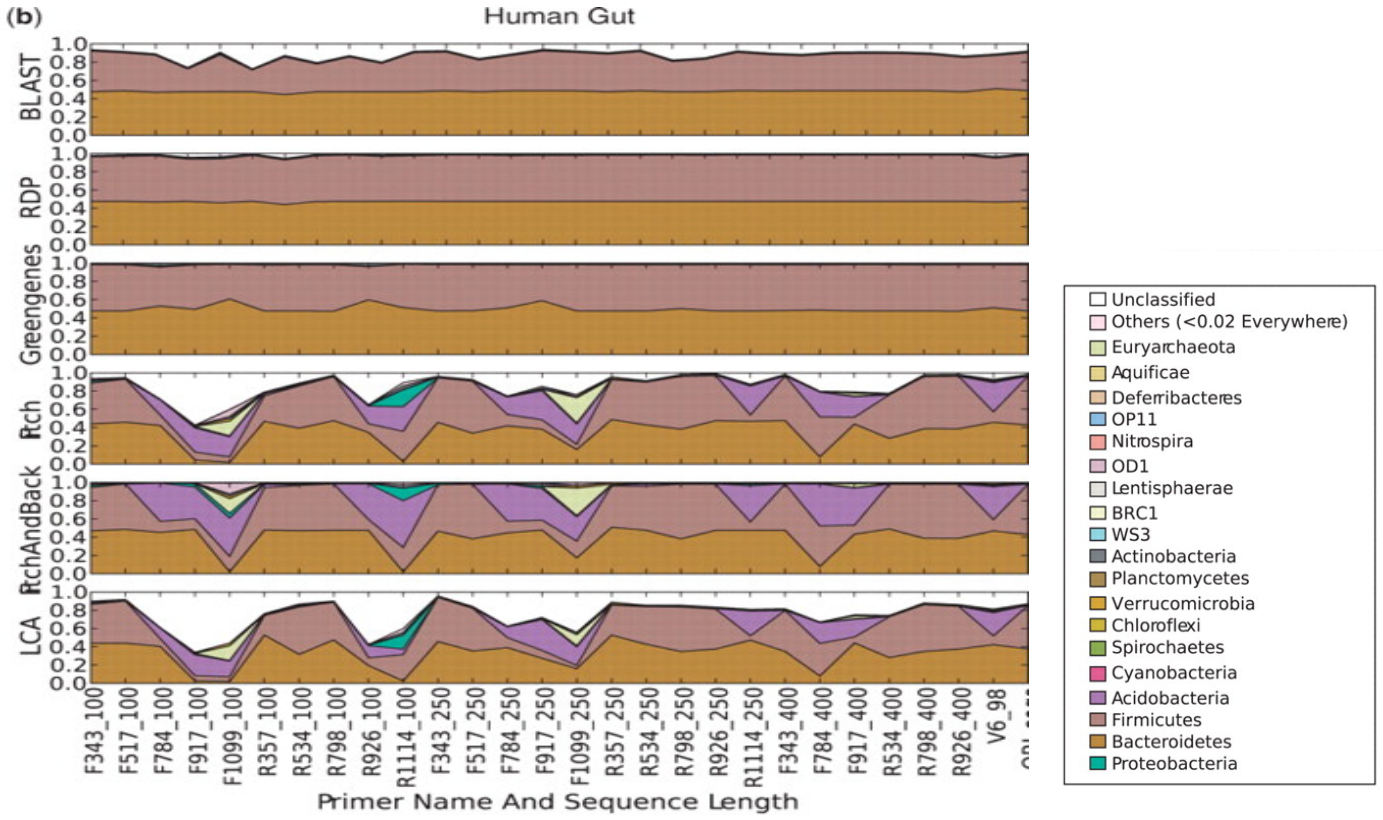
Choice of 16S amplification regions

Human biases

- Sample Contamination
- **Choice of variable region for amplification**
- Primer choice

- Most methods are **sensitive to the region of the 16S rRNA gene** that is targeted for sequencing
- The hypervariable region targeted for sequencing plays a critical role in **influencing the composition of pyrotag communities**

Compositions at the phylum level for Human gut and, using a range of different methods (separate subpanels within each group).



Liu *et al.*, NAR (2008) ; Cruaud *et al.* AEM (2014) ; Kumar *et al.* Plos one (2011)

Choice of 16S amplification regions

Human biases

- Sample Contamination
- **Choice of variable region for amplification**
- Primer choice

- The chimera formation rates for the 16S **V1/V2/V3 region: 22.1–38.5%**
- **V4/V5 region: 3.68–3.88%**
- Chimeric **hot spots** located in **conserved regions**

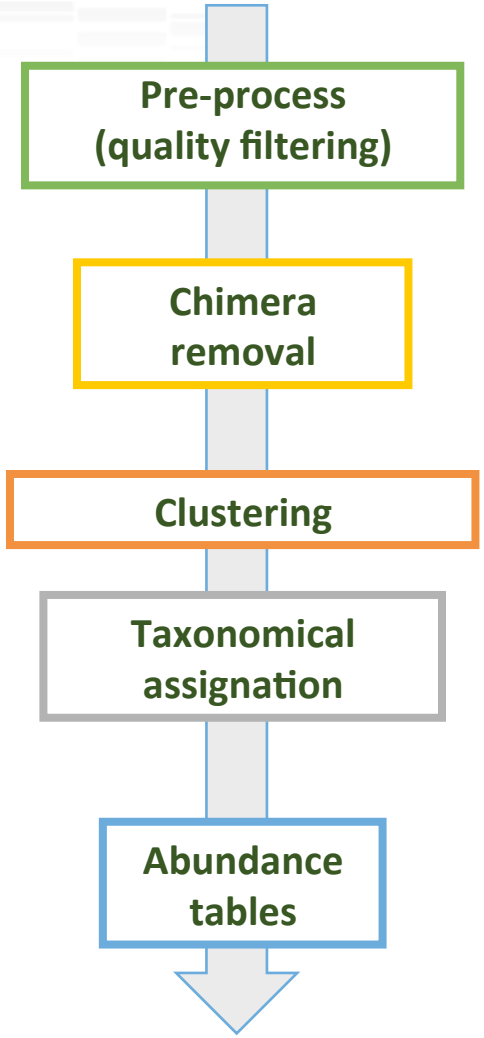
- **Stéphane Chaillou:**
  - **V1-V3: adapted to firmicutes**
  - **V4-V6: adapted to enterobacteria and actinobacteria**

Shin *et al.*, Journal of Microbiology (2014)

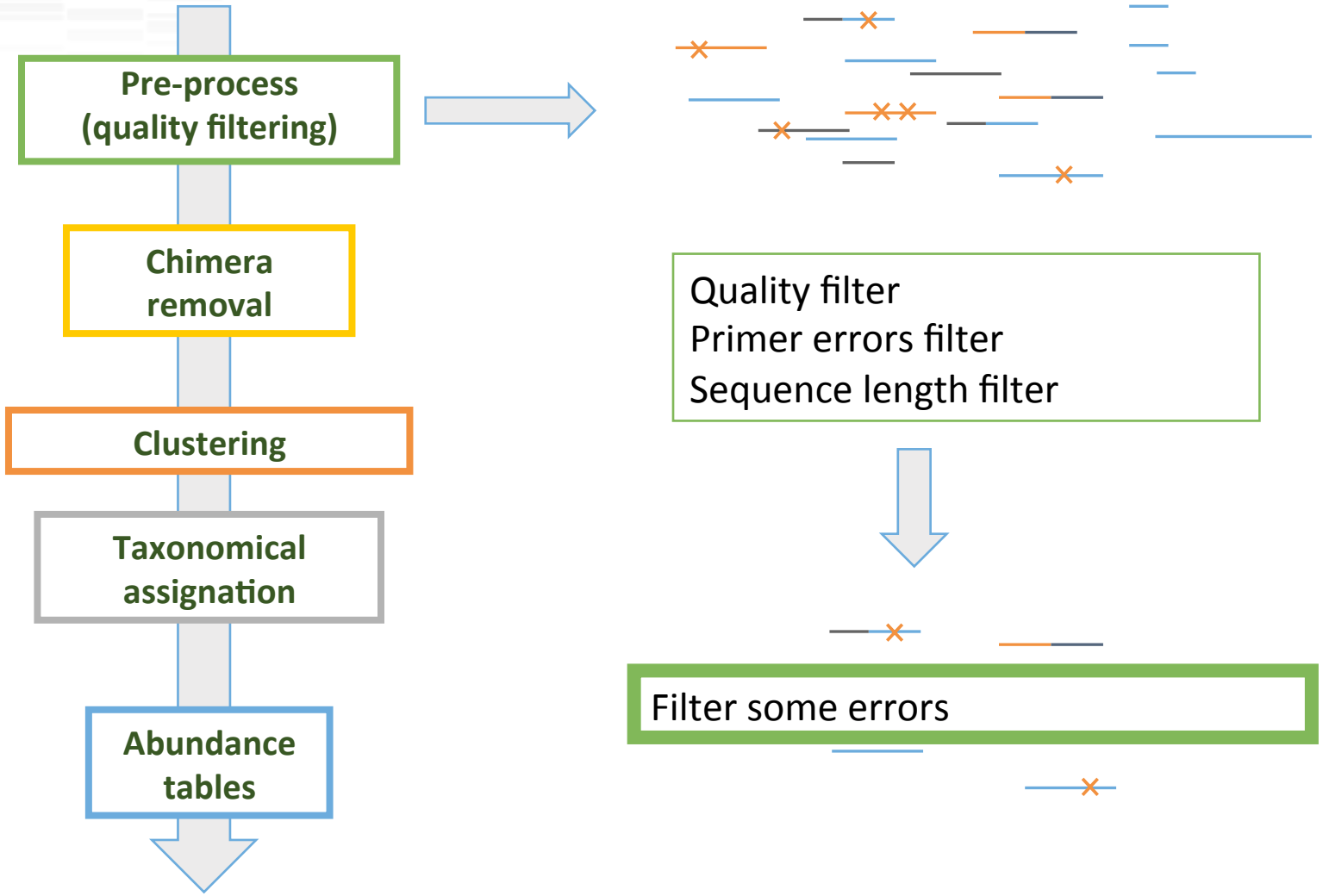
Biases

Conclusion: a lot of biases that the bioinformatics workflow has to take into account !

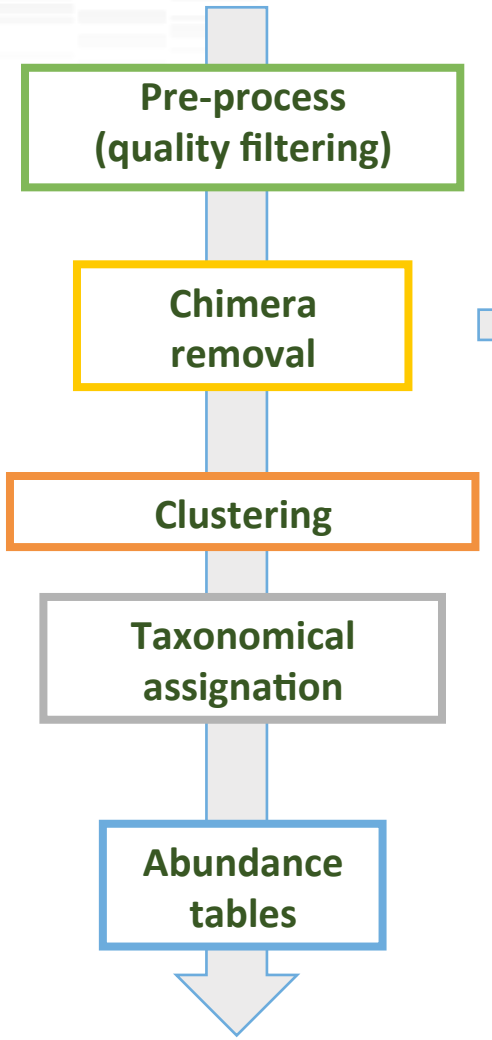
Bioinformatics analysis



Bioinformatics analysis



Method for chimere removal

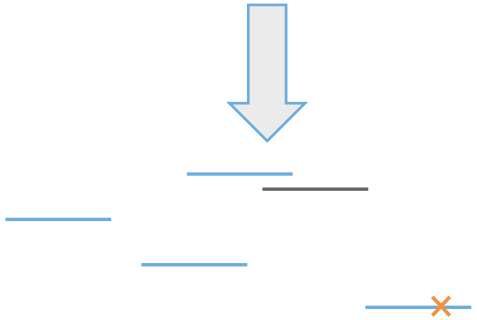


***De novo* methods**

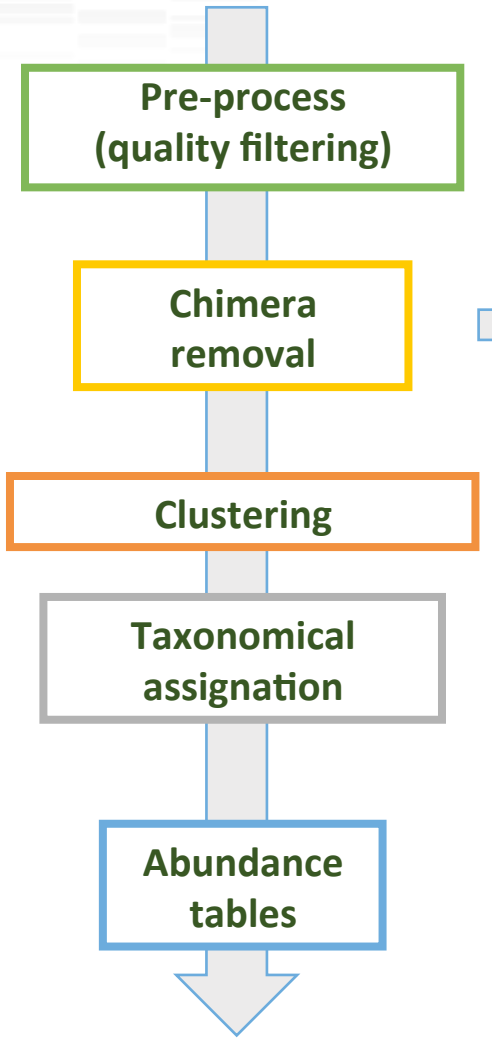
Uchime
ChimeraSlayer
Uparse
CATCh

Reference-based methods

Pintail
Bellerophon
ChimeraSlayer
Uchime
Decipher
CATCh



Method

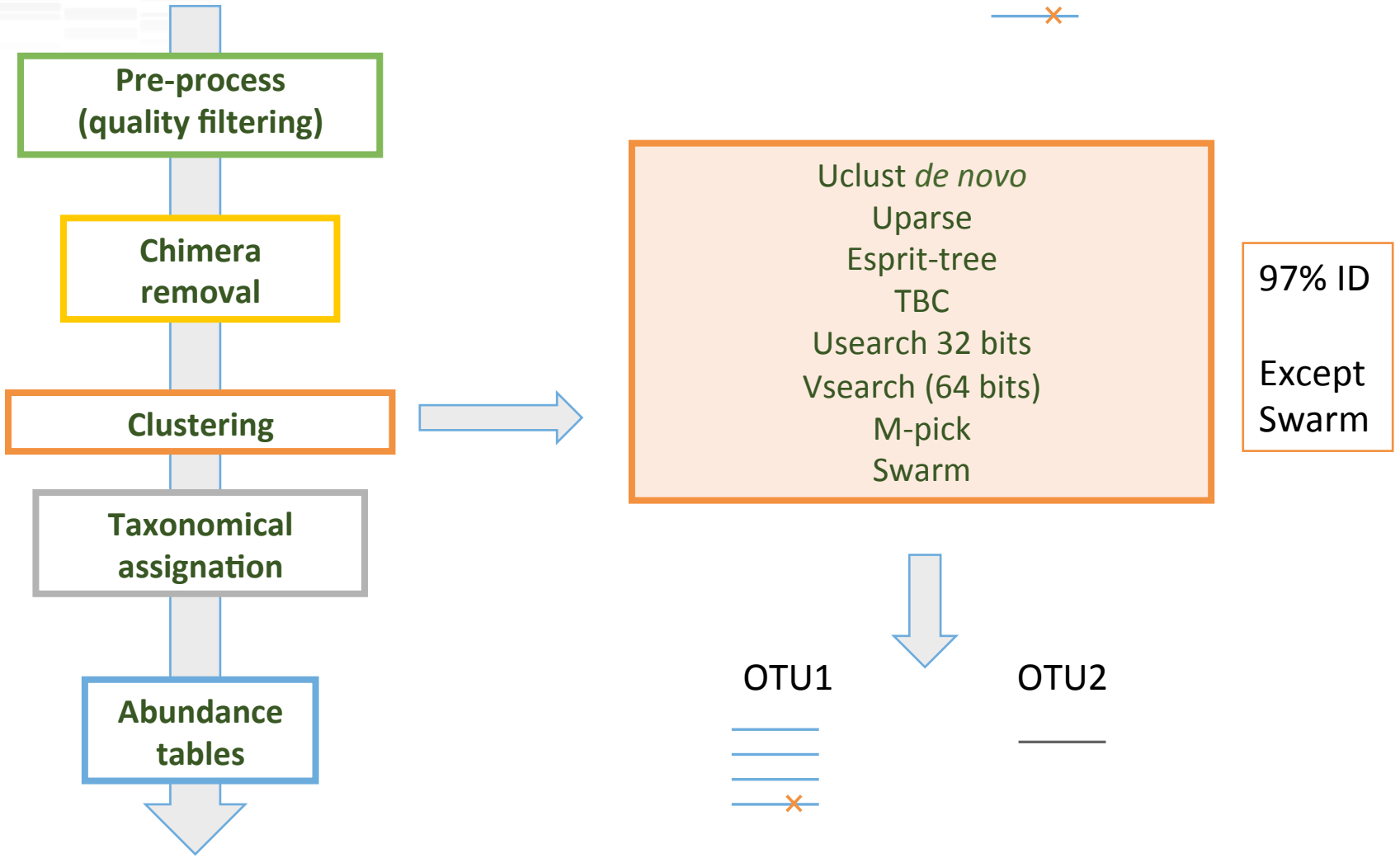


***De novo* methods**
Uchime
ChimeraSlayer
Uparse
CATCh

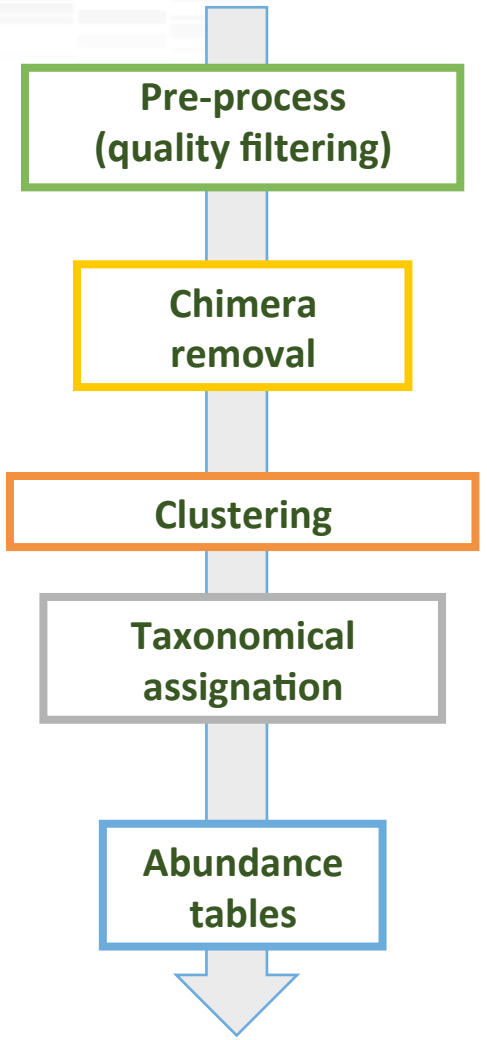
Reference-based methods
Pintail
Bellerophon
ChimeraSlayer
Uchime
Decipher
CATCh

Filter some chimeras

Method


Pre-process
(quality filtering)
Chimera
removal
Clustering
Taxonomical
assignation
Abundance
tables
Uclust *de novo*
Uparse
Esprit-tree
TBC
Usearch 32 bits
Vsearch (64 bits)
M-pick
Swarm
97% ID
Except
Swarm
OTU1
OTU2

Method



Uclust *de novo*
Uparse
Esprit-tree
TBC
Usearch 32 bits
Vsearch (64 bits)
M-pick
Swarm

97% ID

Except
Swarm

Cluster in the same OTU:
- Sequences of the same species
  (variability inside species)
- Sequences with sequencing errors

Preprocessing & removal of chimera

Table 3.
Comparison of the number of clusters formed for each clustering algorithm and pre-processing method

Pre-processing	Clustering algorithms							
	QIIME BLAST	CD-HIT	ESPRIT-Tree	Mothur furthest	Mothur average	UCLUST	UCLUST ref	UCLUST ref optimal
No cleaning	562	485	328	3286	2236	379	651	
Chimera checked	288	131	66	2107	1373	77	323	
Denoised	95	112	104	155	154	104	124	
Denoised + chimera	30	25	25	38	38	25	31	31

Mock dataset: 15 species

- In several cases, the inferred number of OTUs largely exceeded the total number of cells in the samples.
- Such inflation of the OTU numbers corresponded to '**rare biosphere' taxa, composed largely of artifacts.**

Bonder *et al.*, Bioinformatics (2012)

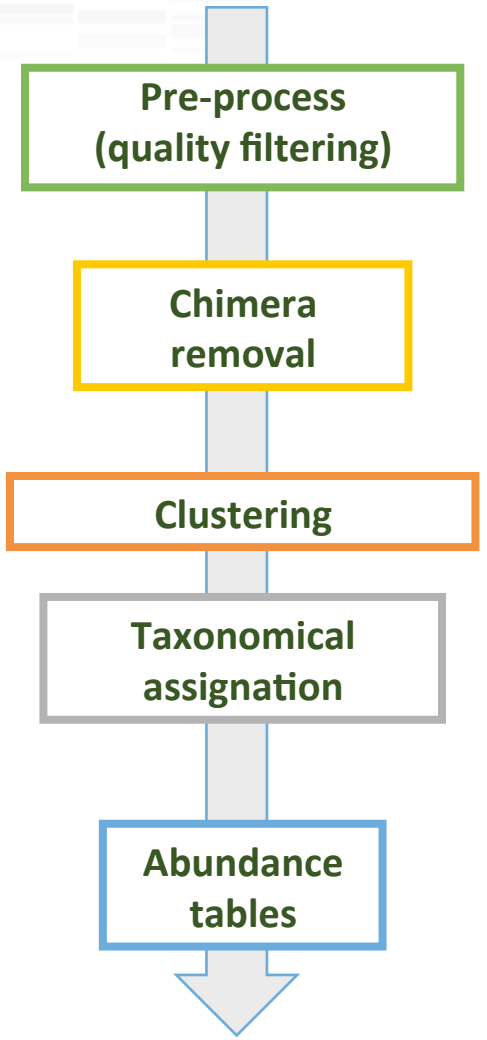
Filters

Filter the result with different criteria, for example

- Suppress singletons (= OTUs with 1 read)
- Abundance of OTUs (> 0.0005% of reads)
- Number of reads by OTU (>100 reads)
- OTU shared between samples (for example OTU in at least 3 samples, if triplicates)
- Most abundant OTUs (first 100 OTUs)

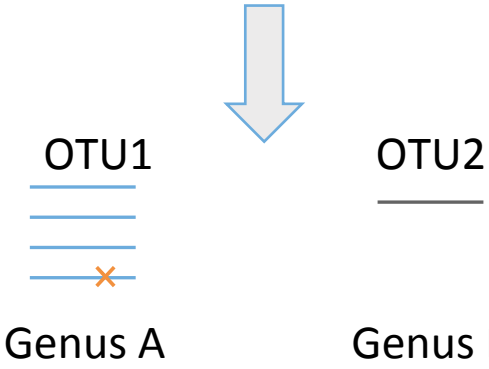
Remove some sequencing errors, chimera, etc

Method for taxonomical assignation

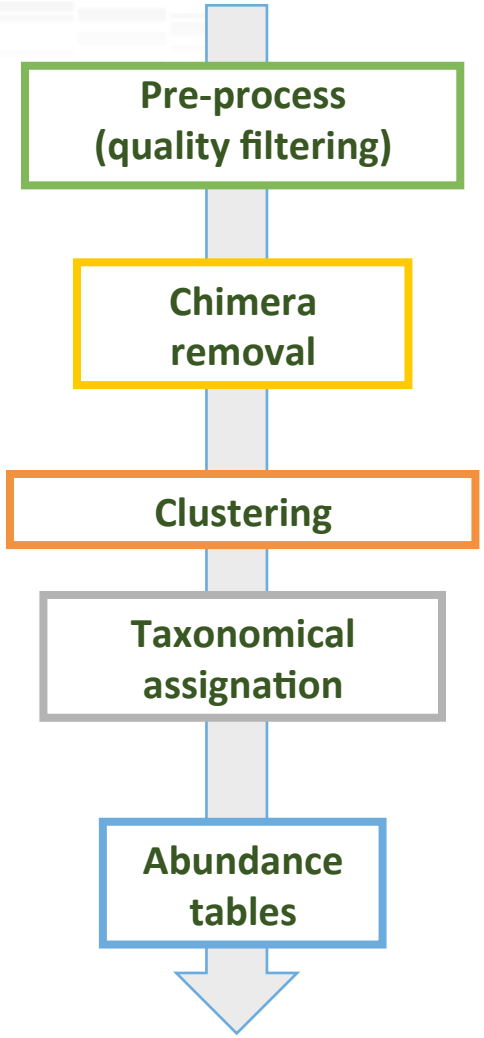


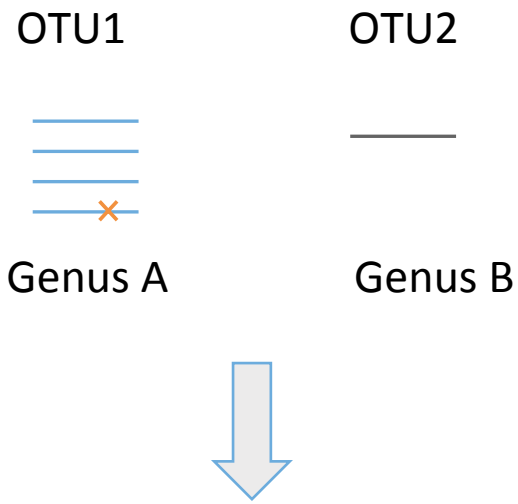
Tools:
RDP Classifier
Megablast
Blast
Rtax

Databases:
Silva
Greengenes
LTP
Dedicated databases



Method





OTU1 OTU2

Genus A Genus B

4 1

	Affiliation	Sample 1	Sample 2	Sample 3
OTU1	Species A	0	100	0
OTU2	Species B	741	0	456
OTU3	Species C	12786	45	3

Which bioinformatics solutions ?

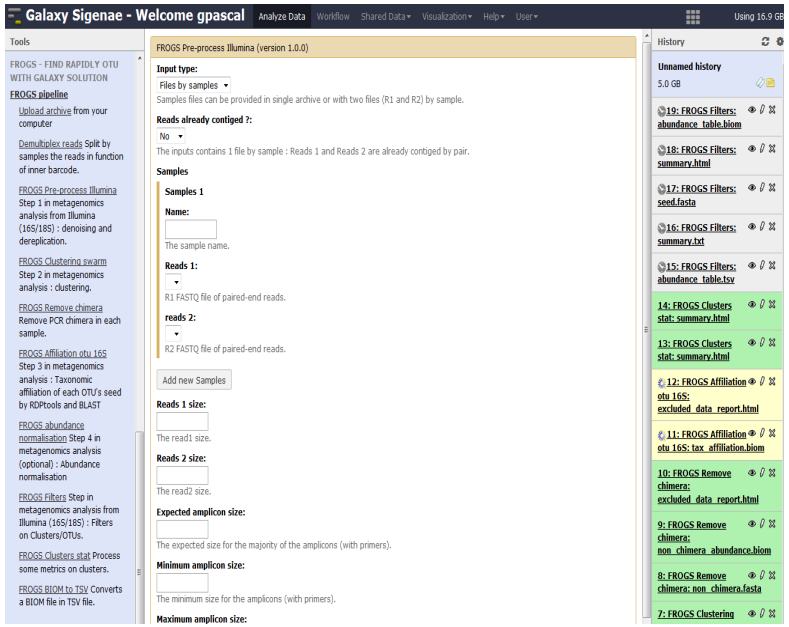
Name	Features
QIIME	http://qiime.org
UPARSE	https://www.drive5.com/uparse/
MOTHUR	https://www.mothur.org
MG-RAST	http://metagenomics.anl.gov
EBI-Metagenomics	https://www.ebi.ac.uk/metagenomics/
FROGS	http://sigenae-workbench.toulouse.inra.fr/

QIIME allows analysis of high-throughput community sequencing data
J Gregory Caporaso et al, Nature Methods, 2010; doi:10.1038/nmeth.f.303
Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities.
Schloss, P.D., et al., Appl Environ Microbiol, 2009, doi: 10.1128/AEM.01541-09

UPARSE: Highly accurate OTU sequences from microbial amplicon reads
Edgar, R.C. et al, *Nature Methods*, 2013, dx.doi.org/10.1038/nmeth.2604
The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes
F Meyer et al, BMC Bioinformatics, 2008, doi:10.1186/1471-2105-9-386

FROGS

- Use platform Galaxy
- Set of modules = Tools to analyze your “big” data
- Independent modules
- Run on Illumina/454 data 16S, 18S, and 23S
- New clustering method
- Many graphics for interpretation
- User friendly, hiding bioinformatics infrastructure/complexity



Pipeline FROGS on http://sigenae-workbench.toulouse.inra.fr/
Poster FROGS: Escudie F., Auer L., Bernard M., Cauquil L., Vidal K., Maman S., Mariadassou M., Hernadez-Raquet G., Pascal G., 2015. FROGS: Find Rapidly OTU with Galaxy Solution. In: Environmental Genomics 2015, Montpellier, France, http://bioinfo.genotoul.fr/fileadmin/user_upload/FROGS_2015_GE_Montpellier_poster.pdf

What one can say or not when using amplicon sequencing?

Can do:

- **detecting microorganisms** present in complex samples with an **unprecedent scale** (detecting sub-dominant taxa can be achieved by sequencing tens of thousand reads per sample) → micobial inventories
- detecting the **relative abondance** of the different taxa in the samples (OTUs more or less abundant)
- analyzing **several samples** at the same time (>300), producing and comparing community profiles (same protocols for every samples)
- **sequence affiliation down to the genus level** in most cases, sometimes down to the species-level (low complexity and well-described ecosystems)

Can't do:

- **exact quantification of the different taxa** detected in the samples (relative abondance, several bias)
- **exact identification** of the microorganisms. It is impossible to distinguish strains belonging to the same species, sometimes even two species belonging to the same genus
- distinguishing between **live and dead** microorganisms
- speculating about the **functional role** of the detected taxa

What's next?

Use OTU tables and statistical tools to analyze community composition and perform biodiversity analysis to evaluate:

- The **richness** to number of OTUs or functional groups present in communities. It caracterises the **composition**.
- The **diversity** takes into account the relative abundancy of species. It caracterises the **structure**

Biodiversity analysis: definitions

- Compute and compare diversity indices. 3 levels of diversity
  - **Alpha diversity**

    Species richness (number of taxa) within a single microbial environment.

    How many different microbial species could be detected in a specific sample?

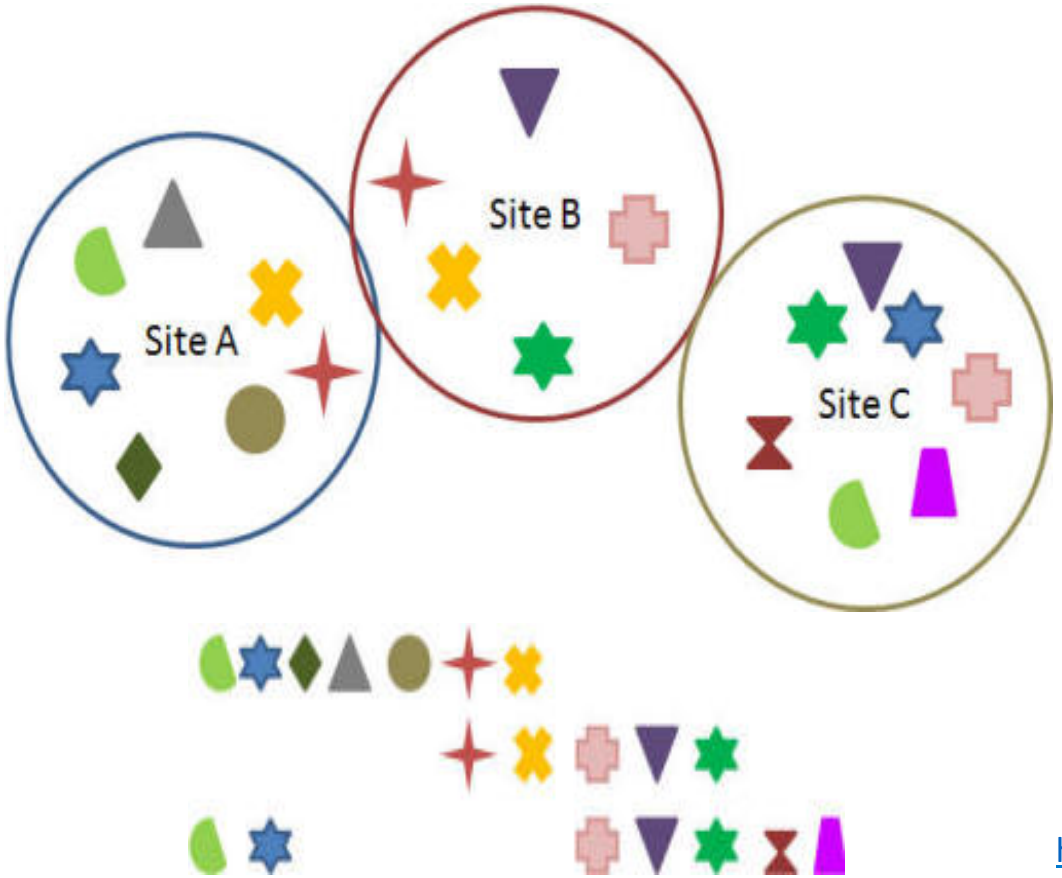
  - **Beta diversity**

    Diversity in microbial community between different environments (difference in taxonomic abundance profiles from different samples).

    How different is the microbial composition in one environment compared to another ?

  - **Gamma-diversity:** a measure of the overall diversity within a large region.

Biodiversity analysis: example



Alpha Diversity
Site A = 7 species, Site B = 5 species, Site C = 7 species

Beta Diversity
A vs B = 8 species
B vs C = 4 species
A vs C = 10 species

Gamma diversity is 3 habitats with 12 species total diversity.

http://www.webpages.uidaho.edu/veg_measure/Modules/Lessons/Module%209%28Composition&Diversity%29/9_2_Biodiversity.htm

Biblio

- Angly, 2014, CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction
- Bachy, 2013, Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study
- Bonder, 2012, Comparing clustering and pre-processing in taxonomy analysis
- Cruaud, 2014, Influence of DNA Extraction Method, 16S rRNA Targeted Hypervariable Regions, and Sample Origin on Microbial Diversity Detected by 454 Pyrosequencing in Marine Chemosynthetic Ecosystems
- Kumar, 2011, Target Region Selection Is a Critical Determinant of Community Fingerprints Generated by 16S Pyrosequencing
- Liu, 2008, Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers
- Mysara, 2014, CATCh, an Ensemble Classifier for Chimera Detection in 16S rRNA Sequencing Studies
- Salter, 2014, Reagent and laboratory contamination can critically impact sequence-based microbiome analyses
- Schloss, 2011, Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies
- Shin, 2014, Regional effects on chimera formation in 454 pyrosequenced amplicons from a mock community
- Smets, 2015, A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing
- Větrovský , 2013, The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses
- Phyloseq package: https://joey711.github.io/phyloseq/index.html

Remerciements

Eric Dugat Bony
Stéphane Chaillou
Lucas Auer
+pole 16S