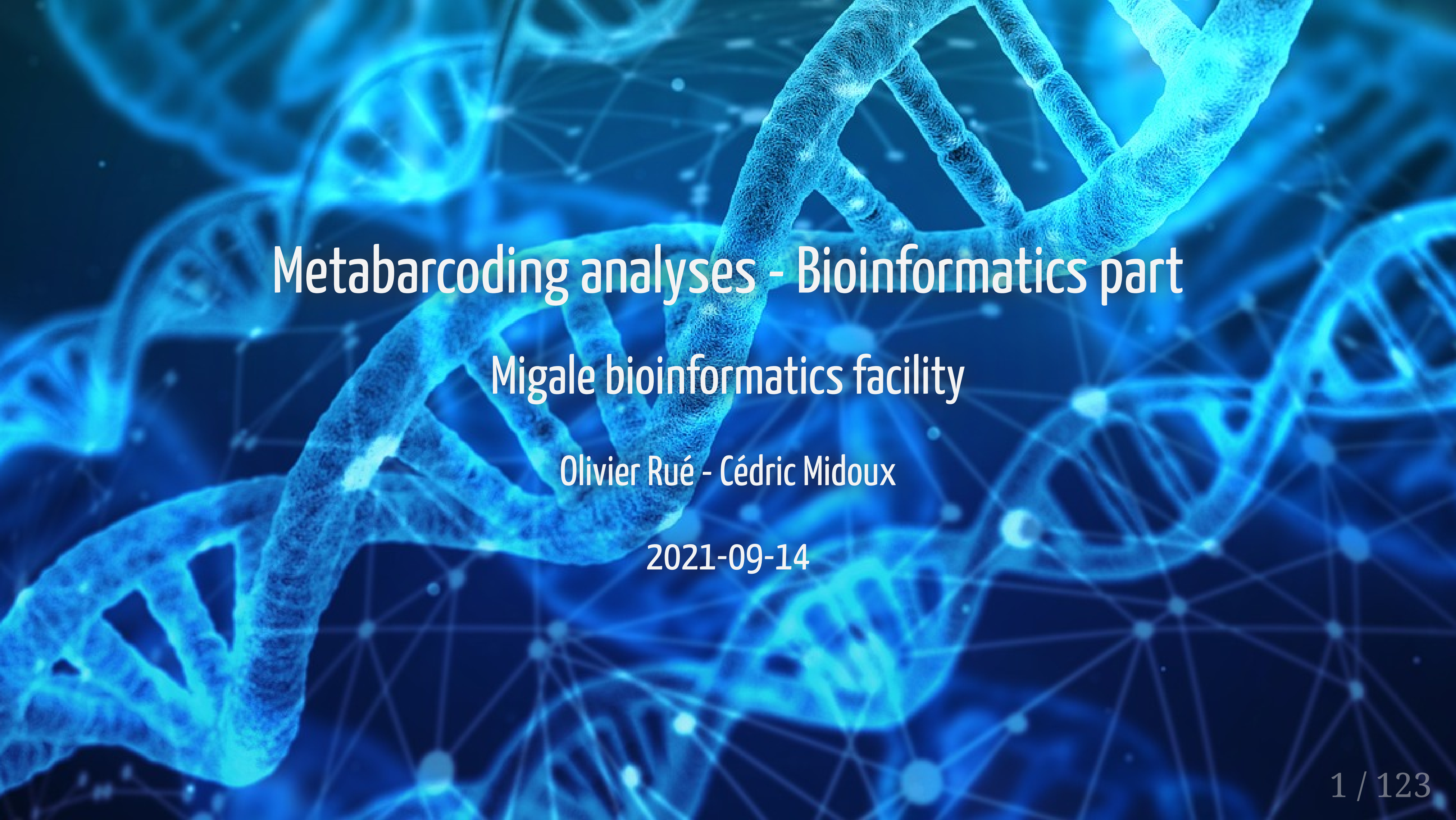


Metabarcoding analyses - Bioinformatics part

Migale bioinformatics facility

Olivier Rué - Cédric Midoux

2021-09-14

Practical informations

- 9h00 - 17h00
- 2 breaks morning and afternoon
- Possibility to have lunch in the INRAE restaurant

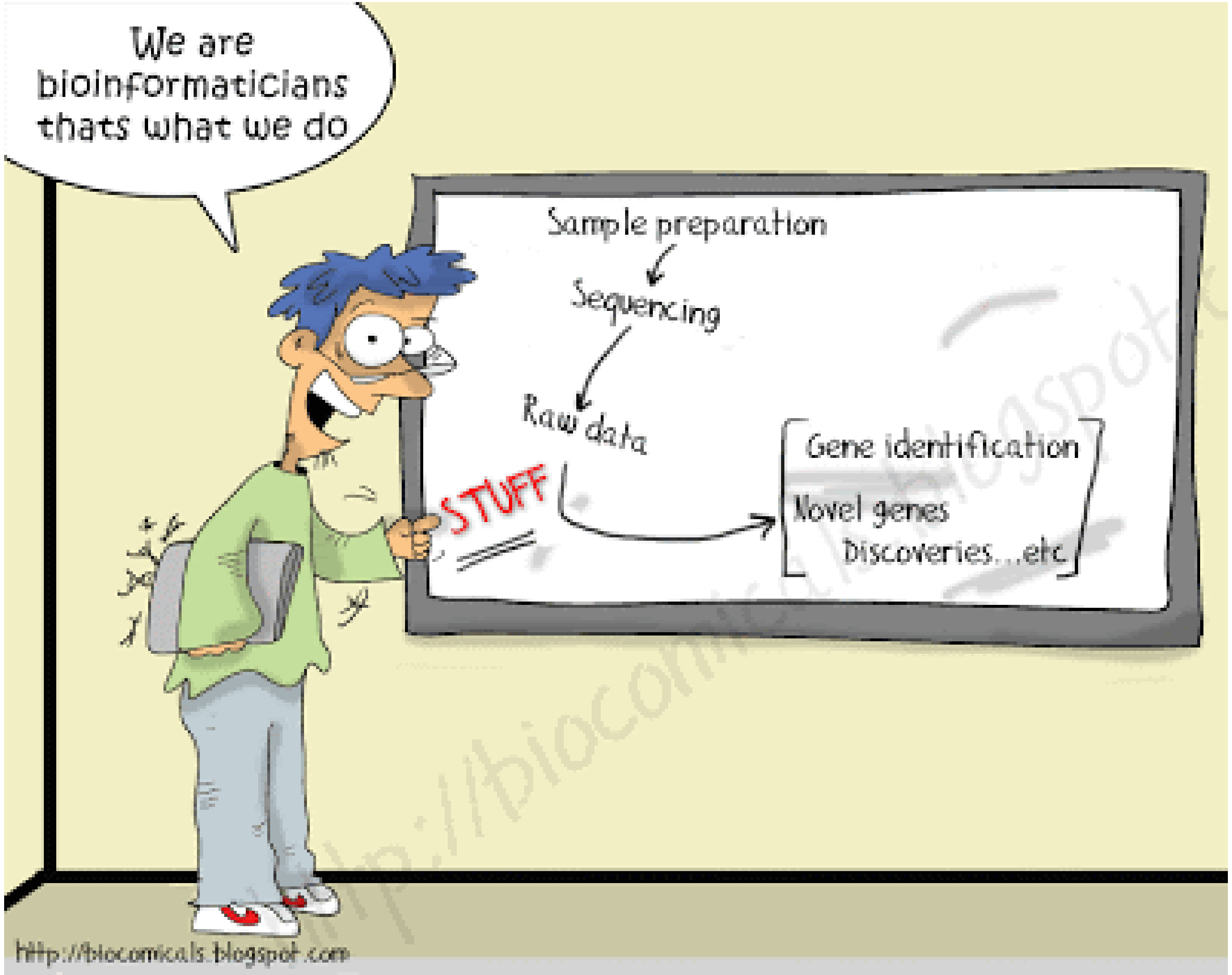
Better know us

- Who are you ?
  - Institution, laboratory, position …
- What are your needs in metagenomics ?
- Do you have already dealed with metagenomics data ?
  - Which kind of data ?
  - Aim of the study ?
- Do you have generated data for a new analysis ?
  - Which design ? How many samples ? Sequencing technology ?

Migale team

- Migale website
- Dedicated service to Data Analysis
  - Specialists in Metagenomics, Genomics, Bacterial genome assembly and annotation
  - Bioinformatics & Statistics
  - 84 projects since 2016
  - Collaboration or Support
- Developments
  - FROGS
  - easy16S, affiliationExplorer

Discover the service offer here


We are bioinformaticians thats what we do
Sample preparation
Sequencing
Raw data
STUFF
Gene identification
Novel genes
Discoveries...etc
http://biocomicals.blogspot.com

Objectives

After this 4 days training, you will:

- Know the outlines, advantages and limits of amplicon sequencing data analysis
- Be able to use **FROGS** (through Galaxy) and **phyloseq** (through easy16S) tools on the training data set
- Be able to identify tools and parameters adapted to your own analyses

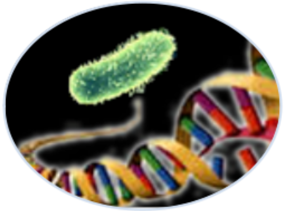
Program

- Day 1 & 2: Bioinformatics
- Day 3 & 4: Statistics

And one time to train with your own data or another dataset.

Introduction to amplicon analyses

Meta-omics using next-genertation sequencing (NGS)



DNA

RNA

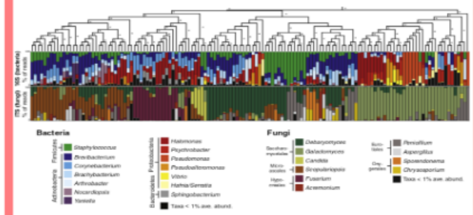
Metagenomics

Metatranscriptomics

Amplicon sequencing

Shotgun sequencing

RNA sequencing

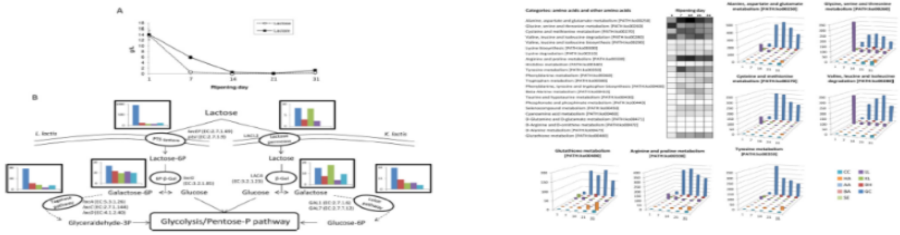


Wolfe *et al.*, 2014

Who is here?

Almeida *et al.*, 2014

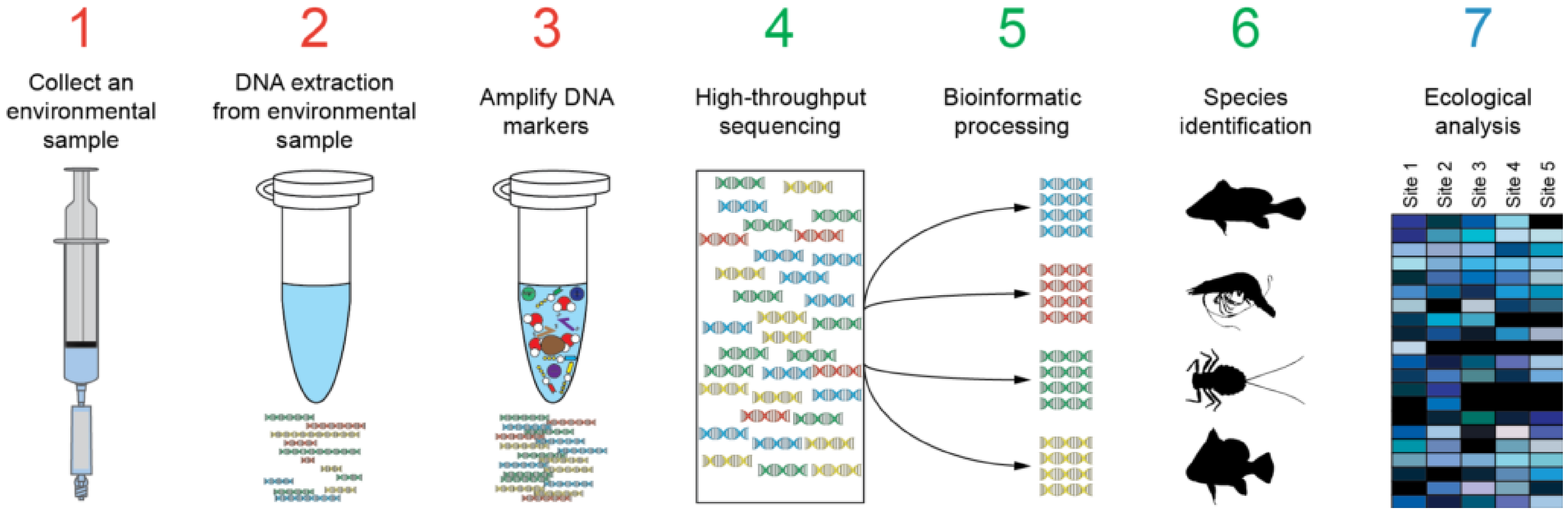
What can they do?



Dugat-Bony *et al.*, 2015

What are they doing?

Meta-omics using next-genertation sequencing (NGS)



1 Collect an environmental sample

2 DNA extraction from environmental sample

3 Amplify DNA markers

4 High-throughput sequencing

5 Bioinformatic processing

6 Species identification

7 Ecological analysis

Strengths and weaknesses of amplicon analyses?

http://scrumblr.ca/strengths_weaknesses

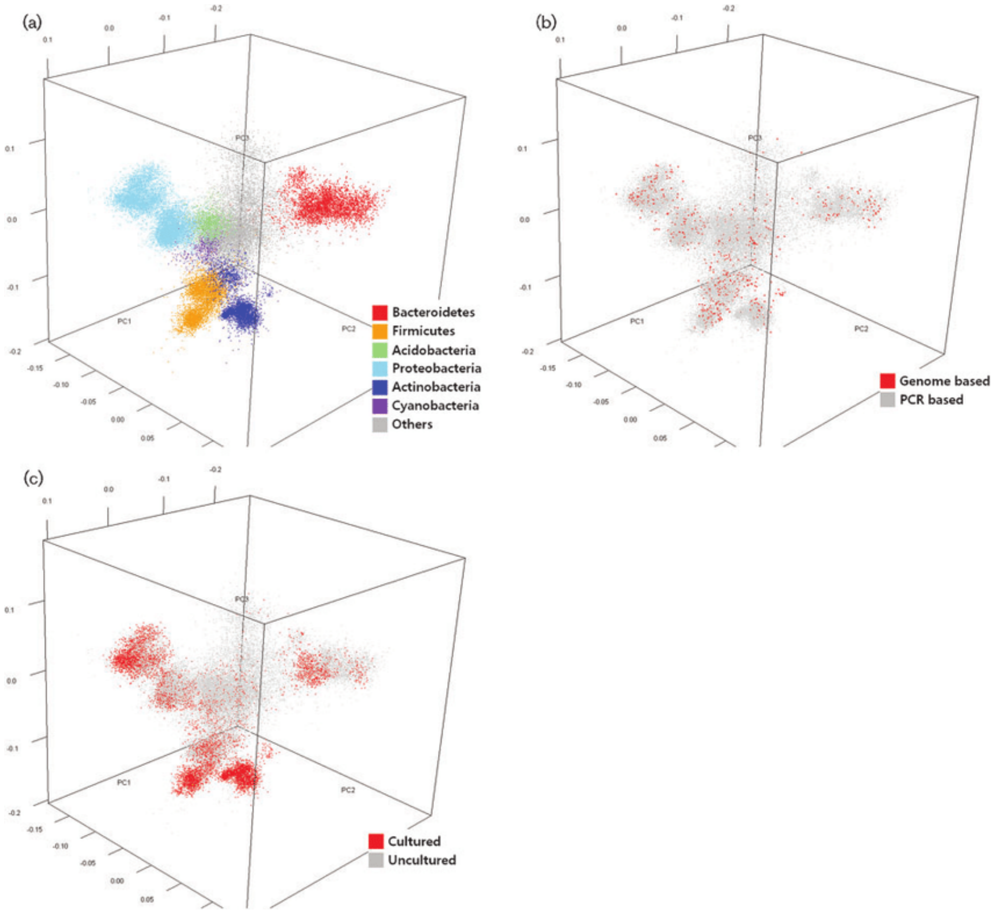
Strengths

- Detect subdominant microorganisms present in complex samples → microbial inventories
- Get (approximate) relative abondances of different taxa in samples
- Analyze and compare many taxa (hundreds) at the same time
- Taxonomic profiles of the communities (usually up to genus level, and sometimes up to species or strain)
- Low cost

Weaknesses

- Compositional data, many biases -> no absolute quantification
- Exact identification of the organisms difficult
- Hard to distinguish live and dead fractions of the communities
- No functional view of the ecosystem

Gene marker power



Kim et al (2012)

Choice of a marker gene

The perfect / ideal gene marker:

- is ubiquist
- is conserved among taxa
- is enough divergent to distinguish stains
- is not submitted to lateral transfer
- has only one copy in genome
- has conserved regions to design *specific* primers
- is enough characterized to be present in databases for taxonomic affiliation

Bacterial targets

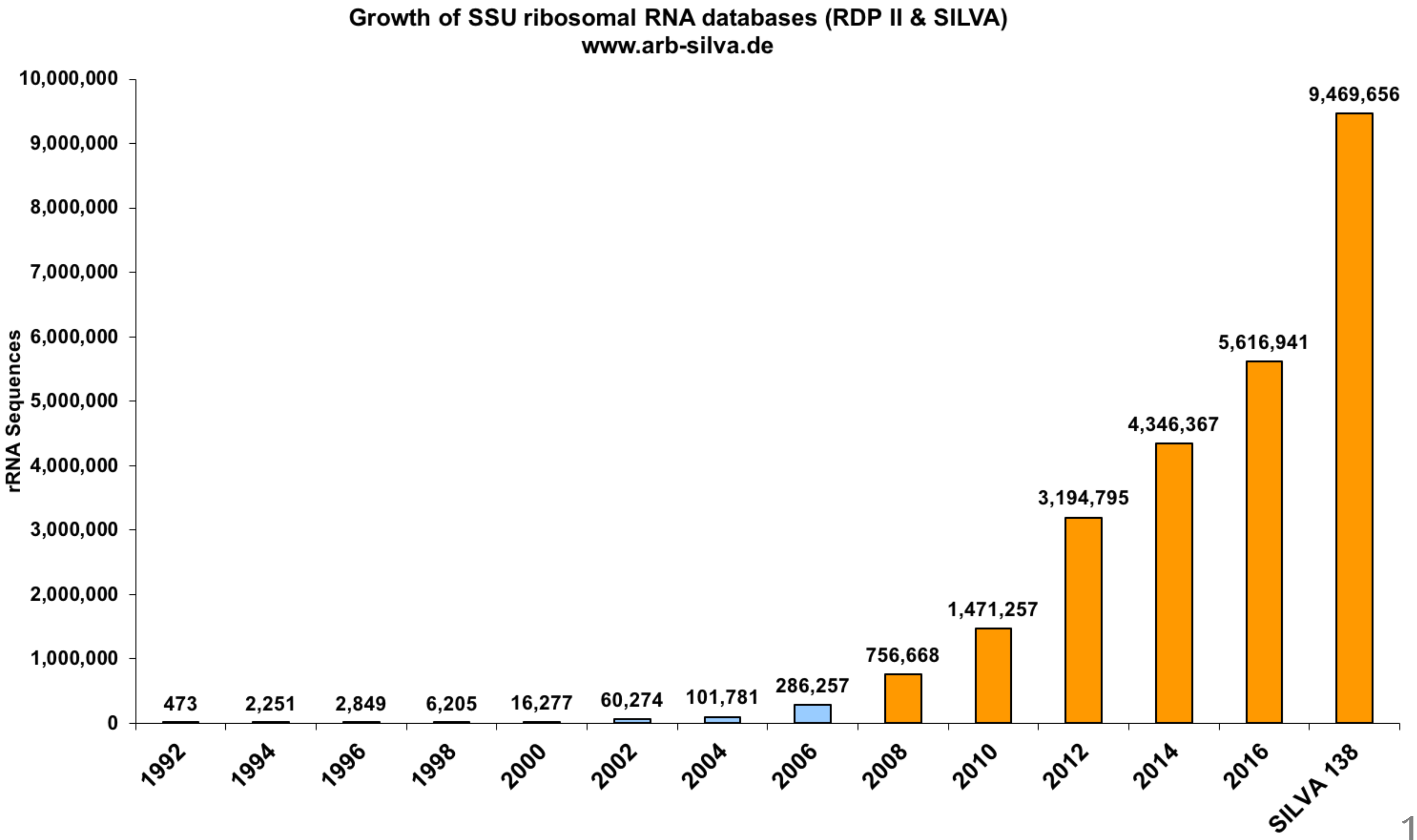
The genes that have been proposed for this task include those encoding :

- 16S / 23S rRNA
- DNA gyrase subunit B (*gyrB*)
- RNA polymerase subunit B (*rpoB*)
- TU elongation factor (*tuf*)
- DNA recombinase protein (*recA*)
- protein synthesis elongation factor-G (*fusA*)
- dinitrogenase protein subunit D (*nifD*) …

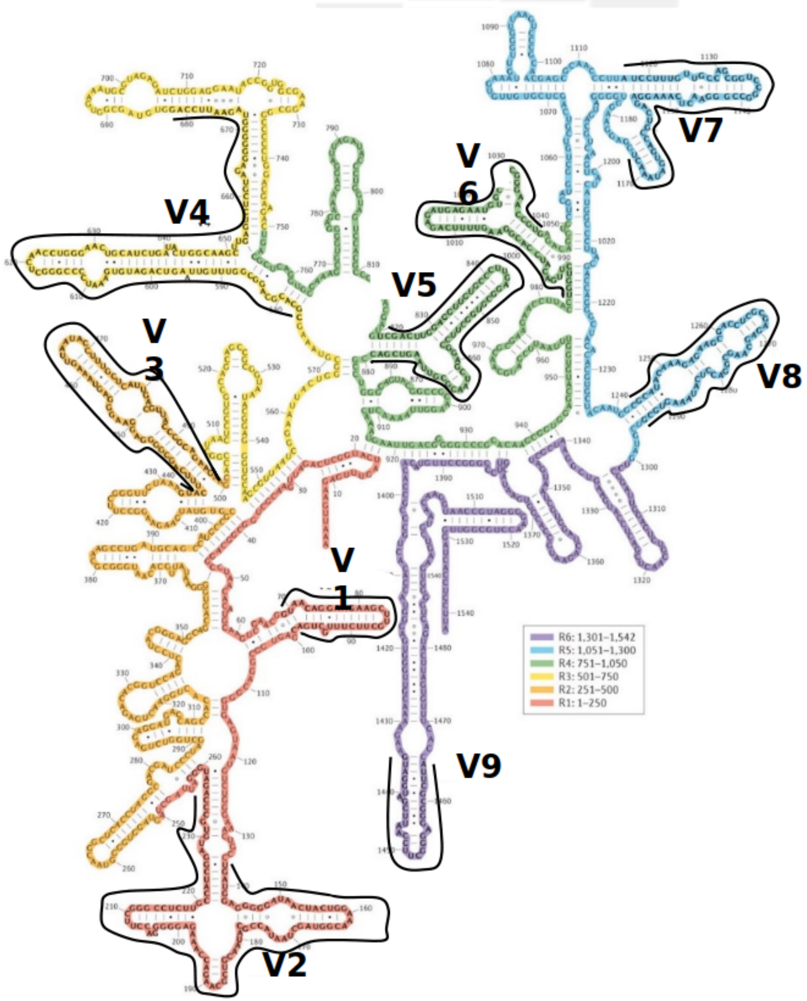
Bacterial lineages vary in their genomic contents, which suggests that different genes might be needed to resolve the diversity within certain taxonomic groups.

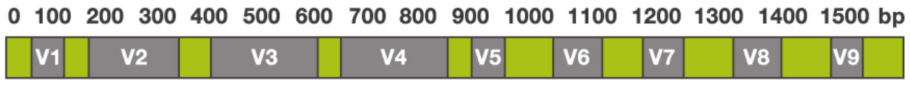
The gene encoding the small subunit of the ribosomal RNA

- 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 transfer
- Availability of databases facilitating comparison

Growth of SSU ribosomal RNA databases (RDP II & SILVA)
www.arb-silva.de
rRNA Sequences
10,000,000
9,000,000
8,000,000
7,000,000
6,000,000
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
473
2,251
2,849
6,205
16,277
60,274
101,781
286,257
756,668
1,471,257
3,194,795
4,346,367
5,616,941
9,469,656
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
SILVA 138

16S rRNA structure



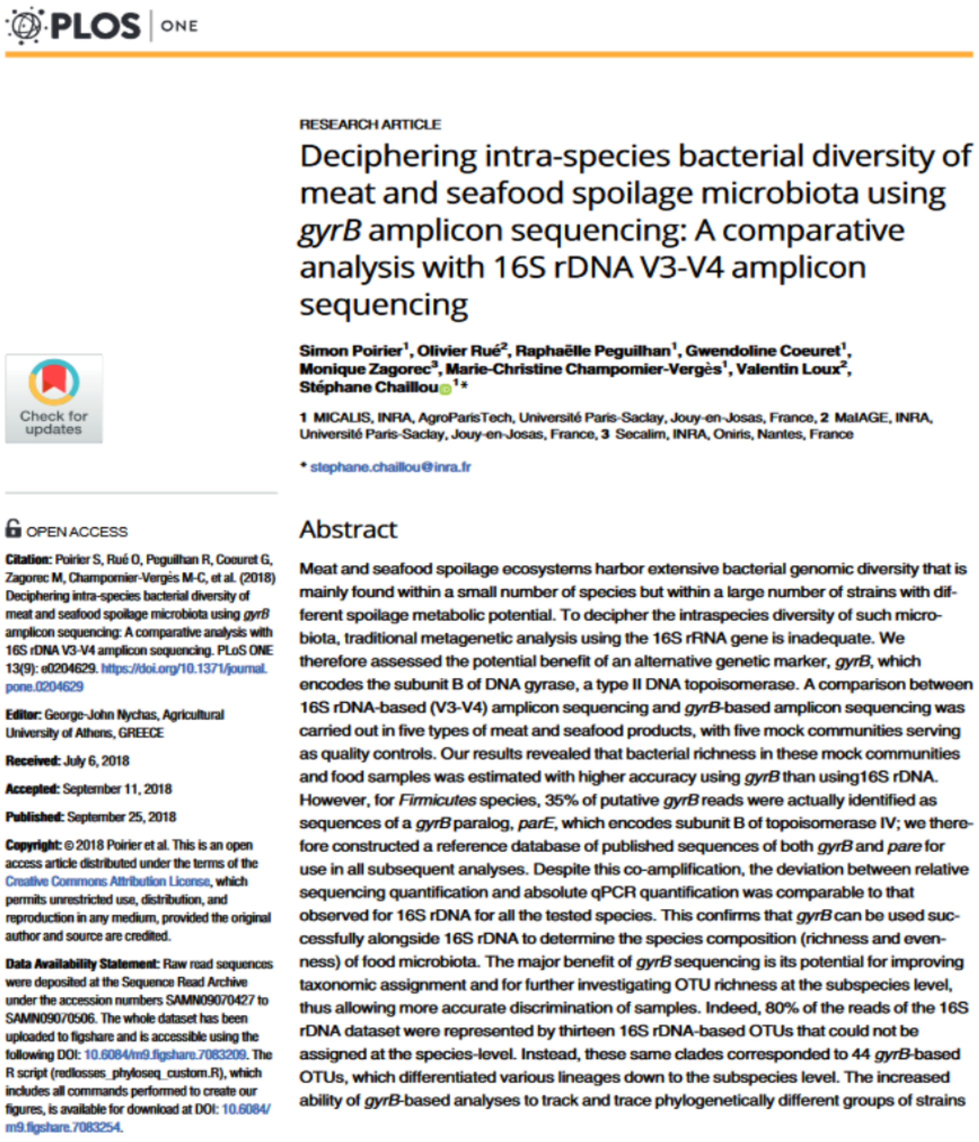


CONSERVED REGIONS: unspecific applications

VARIABLE REGIONS: group or species-specific applications

Example of *gyr*B as interesting marker gene

- A single-copy housekeeping gene that encodes the subunit B of DNA gyrase, a type II DNA topoisomerase, and therefore plays an essential role in DNA replication.
- Essential and ubiquitous in bacteria
- Higher rate of base substitution than 16S rDNA does
- Sufficiently large in size for use in analysis of microbial communities.
- Also present in Eukarya and sometimes in Archaea but it shows enough sequence dissimilarity between the three domains of life to be used selectively for Bacteria.



PLOS | ONE

RESEARCH ARTICLE

Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using *gyrB* amplicon sequencing: A comparative analysis with 16S rDNA V3-V4 amplicon sequencing

Simon Poirier[1], Olivier Rué[2], Raphaëlle Peguilhan[1], Gwendoline Coeuret[1], Monique Zagorec[3], Marie-Christine Champomier-Vergès[1], Valentin Loux[2], Stéphane Chaillou[1]*

1 MICALIS, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France, 2 MaIAGE, INRA, Université Paris-Saclay, Jouy-en-Josas, France, 3 Secalim, INRA, Oniris, Nantes, France

* stephane.chaillou@inra.fr

OPEN ACCESS

Citation: Poirier S, Rué O, Peguilhan R, Coeuret G, Zagorec M, Champomier-Vergès M-C, et al. (2018) Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using *gyrB* amplicon sequencing: A comparative analysis with 16S rDNA V3-V4 amplicon sequencing. PLoS ONE 13(9): e0204629. https://doi.org/10.1371/journal.pone.0204629

Editor: George-John Nychas, Agricultural University of Athens, GREECE

Received: July 6, 2018

Accepted: September 11, 2018

Published: September 25, 2018

Data Availability Statement: Raw read sequences were deposited at the Sequence Read Archive under the accession numbers SAMN09070427 to SAMN09070506. The whole dataset has been uploaded to figshare and is accessible using the following DOI: 10.6084/m9.figshare.7083209. The R script (redlosses_phyloseq_custom.R), which includes all commands performed to create our figures, is available for download at DOI: 10.6084/m9.figshare.7083254.

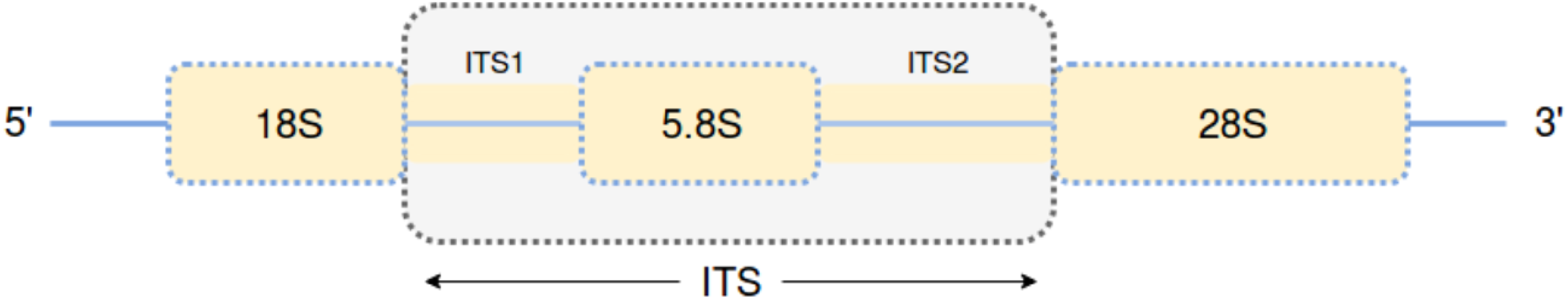
Abstract

Meat and seafood spoilage ecosystems harbor extensive bacterial genomic diversity that is mainly found within a small number of species but within a large number of strains with different spoilage metabolic potential. To decipher the intraspecies diversity of such microbiota, traditional metagenetic analysis using the 16S rRNA gene is inadequate. We therefore assessed the potential benefit of an alternative genetic marker, *gyrB*, which encodes the subunit B of DNA gyrase, a type II DNA topoisomerase. A comparison between 16S rDNA-based (V3-V4) amplicon sequencing and *gyrB*-based amplicon sequencing was carried out in five types of meat and seafood products, with five mock communities serving as quality controls. Our results revealed that bacterial richness in these mock communities and food samples was estimated with higher accuracy using *gyrB* than using16S rDNA. However, for *Firmicutes* species, 35% of putative *gyrB* reads were actually identified as sequences of a *gyrB* paralog, *parE*, which encodes subunit B of topoisomerase IV; we therefore constructed a reference database of published sequences of both *gyrB* and *pare* for use in all subsequent analyses. Despite this co-amplification, the deviation between relative sequencing quantification and absolute qPCR quantification was comparable to that observed for 16S rDNA for all the tested species. This confirms that *gyrB* can be used successfully alongside 16S rDNA to determine the species composition (richness and evenness) of food microbiota. The major benefit of *gyrB* sequencing is its potential for improving taxonomic assignment and for further investigating OTU richness at the subspecies level, thus allowing more accurate discrimination of samples. Indeed, 80% of the reads of the 16S rDNA dataset were represented by thirteen 16S rDNA-based OTUs that could not be assigned at the species-level. Instead, these same clades corresponded to 44 *gyrB*-based OTUs, which differentiated various lineages down to the subspecies level. The increased ability of *gyrB*-based analyses to track and trace phylogenetically different groups of strains

Poirier et al (2018)

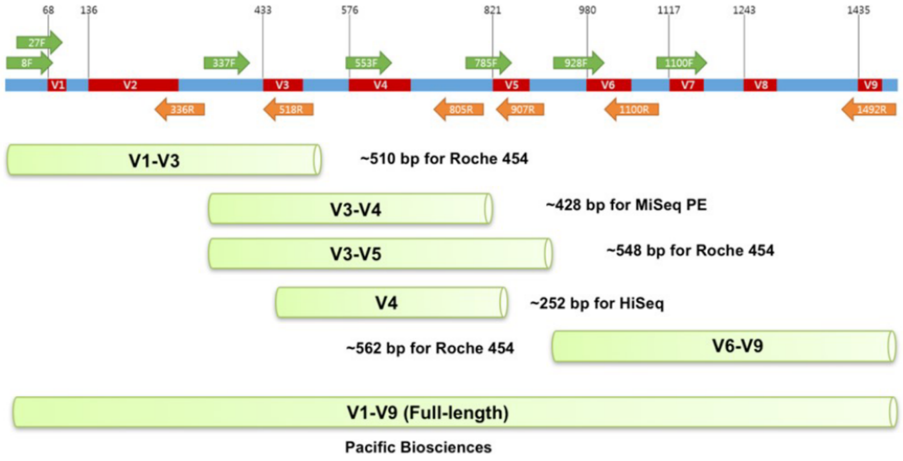
Eukaryotic counterpart

- 18S (small subunit ribosomal RNA)
- ITS (Internal Transcribed Spacers)
    - Length variability (50-1000 nt)
    - Many copies (up to hundreds!)

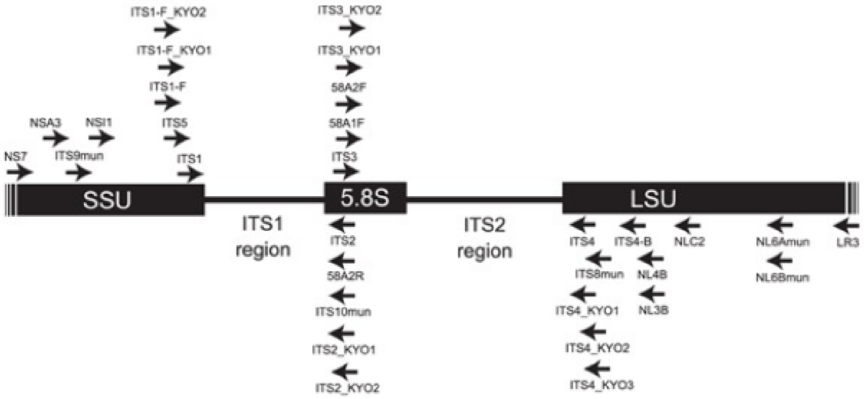


Primers choice

- 16S rRNA gene



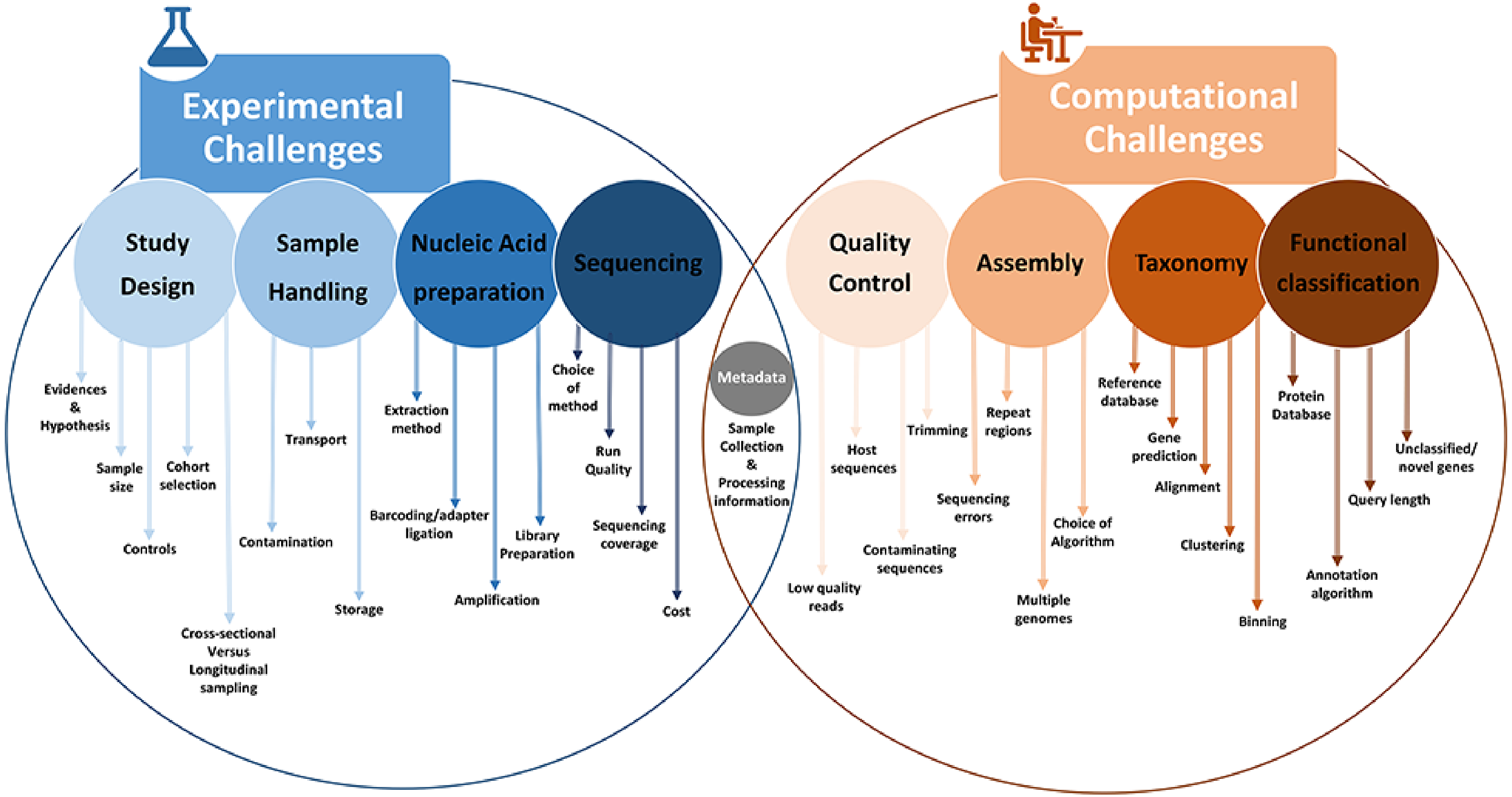
- Internal Transcribed Spacer



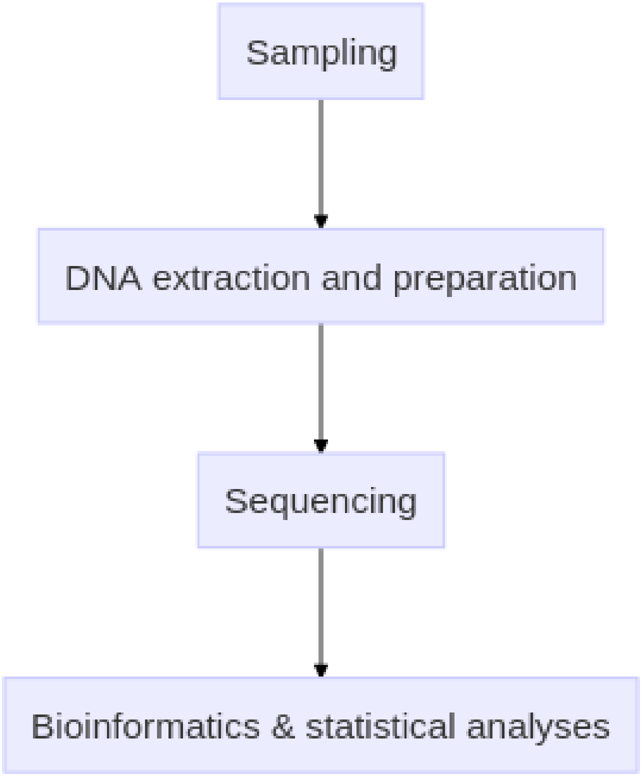
- A lot of others…

Planning an experiment

Planning an experiment



Planning an experiment


Sampling
DNA extraction and preparation
Sequencing
Bioinformatics & statistical analyses

Questions and biases come at each step!

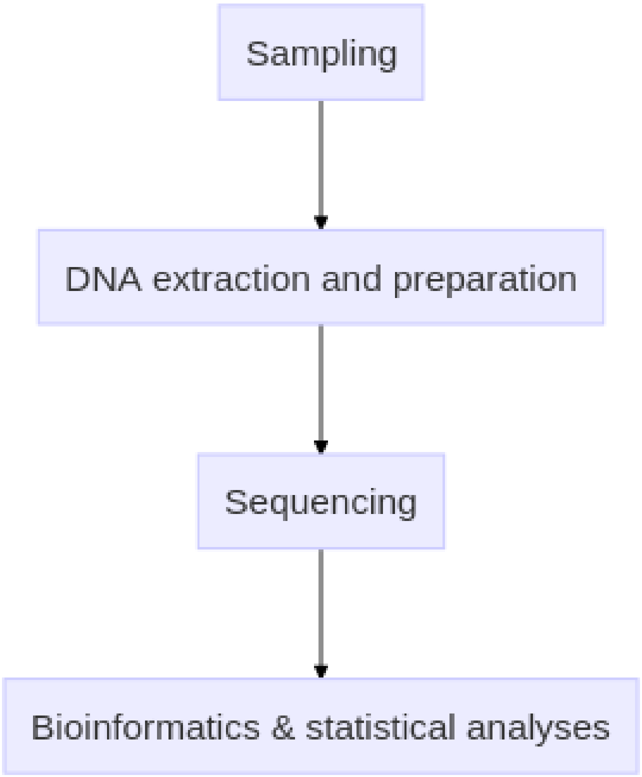
Expected output after bioinformatics

- A matrix table containing "species" and abundances in samples

OTU	Affiliation	Sample1	Sample2	Sample3
OTU1	SpeciesA	0	500	0
OTU2	GenusA	200	41	100
OTU3	SpeciesB	1000	100	1000

Experimental design

Thinking before acting



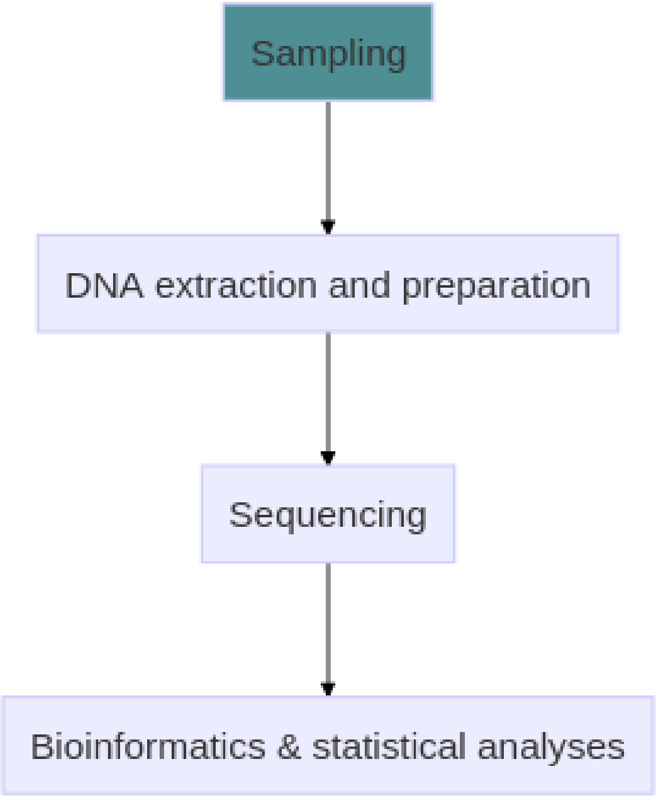
Sampling

DNA extraction and preparation

Sequencing

Bioinformatics & statistical analyses

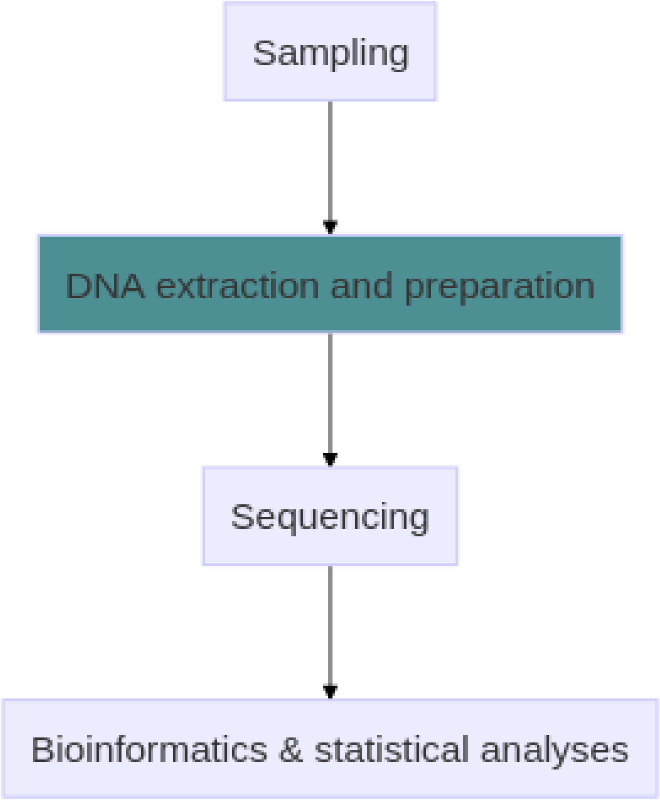
Sampling


Sampling
DNA extraction and preparation
Sequencing
Bioinformatics & statistical analyses

- Number of samples?
- Associated metadata are essential (Too many is better than too few)
- Contamination in lab
- Conservation / Transportation
- Storage

Bharti and Grimm (2019)

DNA extraction and preparation


Sampling
DNA extraction and preparation
Sequencing
Bioinformatics & statistical analyses

- Mechnical or chemical lysis?
- Choice of DNA extraction kit
- PCR amplification biases

Universal primers are not so universal

- *Akkermensia* genus detected (qPCR) but not found in metabarcoding results
- Primers used for amplification
  - F343: TACGGRAGGCAGCAG
  - R784: TACCAGGGTATCTAATCCT
- Mismatches in primers
  - 2 mismatches in Forward
  - 1 mismatch in Reverse
- No amplification…

Alard et al (2016)

Biological biases

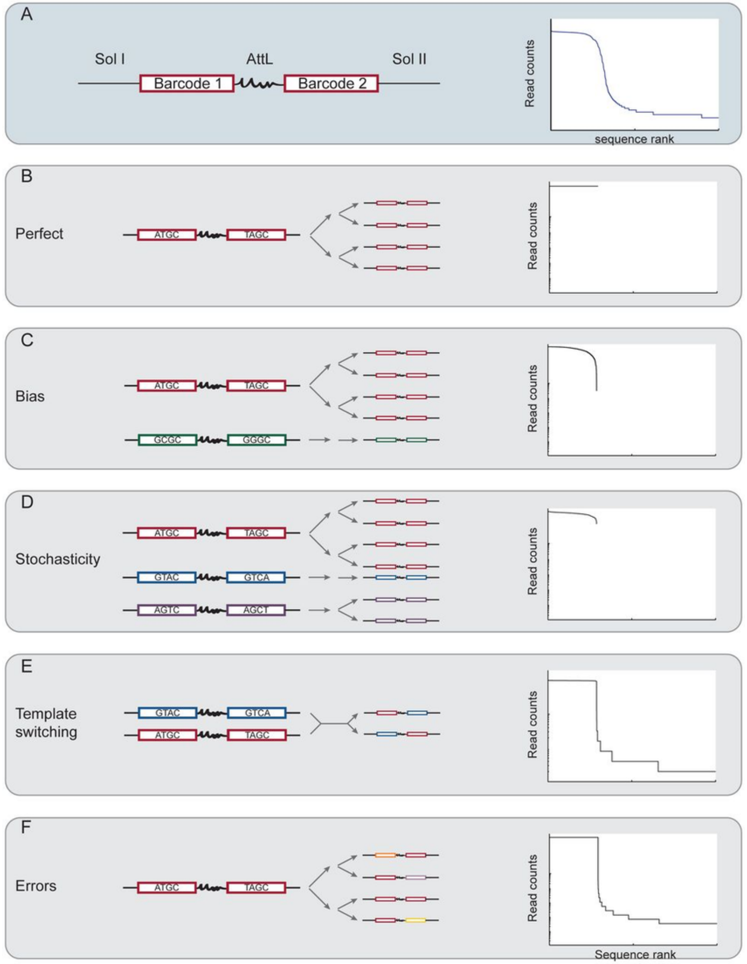
- Gene copy number spans over an order of magnitude, from 1 to up to 15 in Bacteria, but only up to 5 in Archaea
- Only a minority of bacterial genomes harbors identical 16S rRNA gene copies
- Sequence diversity increases with increasing copy numbers.
- While certain taxa harbor dissimilar 16S rRNA genes, others contain sequences common to multiple species.
- Quantification is impossible (in real life)!

Vetrovsky and Baldrian (2013) ; *Angly et al (2014)*

PCR amplification bias

- Amplification by PCR has sequence-dependence efficiency, especially the sequence that binds to primers.
- If one sequence is amplified 10% more than another in one round, it will be 1.130 = 17.4 x more abundant after 30 rounds.
- This effect is most important when the sequence has one or more mismatches with the primer.
- With one mismatch, amplification efficiency is usually significantly less, and with two or more mismatches the sequence may not be amplified to detectable levels.

PCR problems

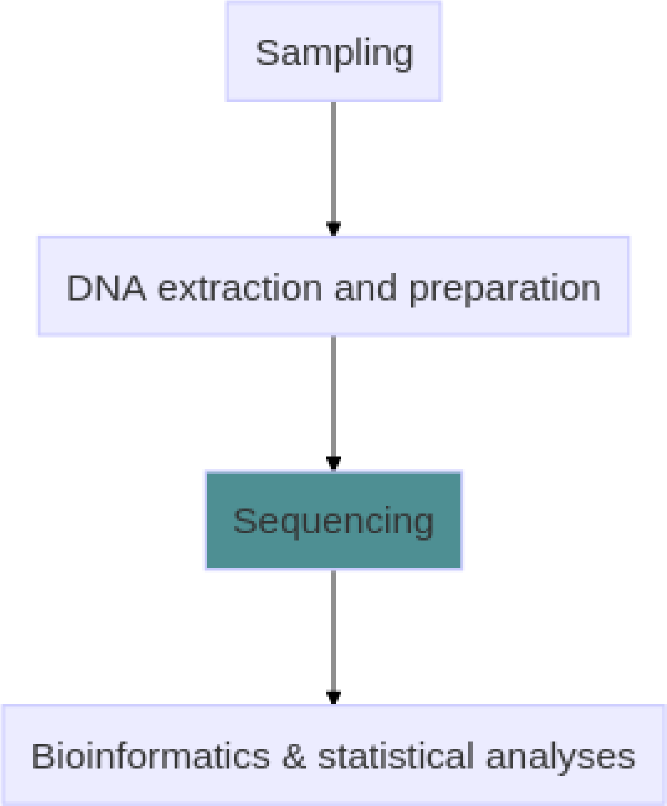


- C and D impact the abundance without adding new sequences
- E and F add new sequences

Kebschull and Zador (2015)

Sequencing

Sequencing



Sampling

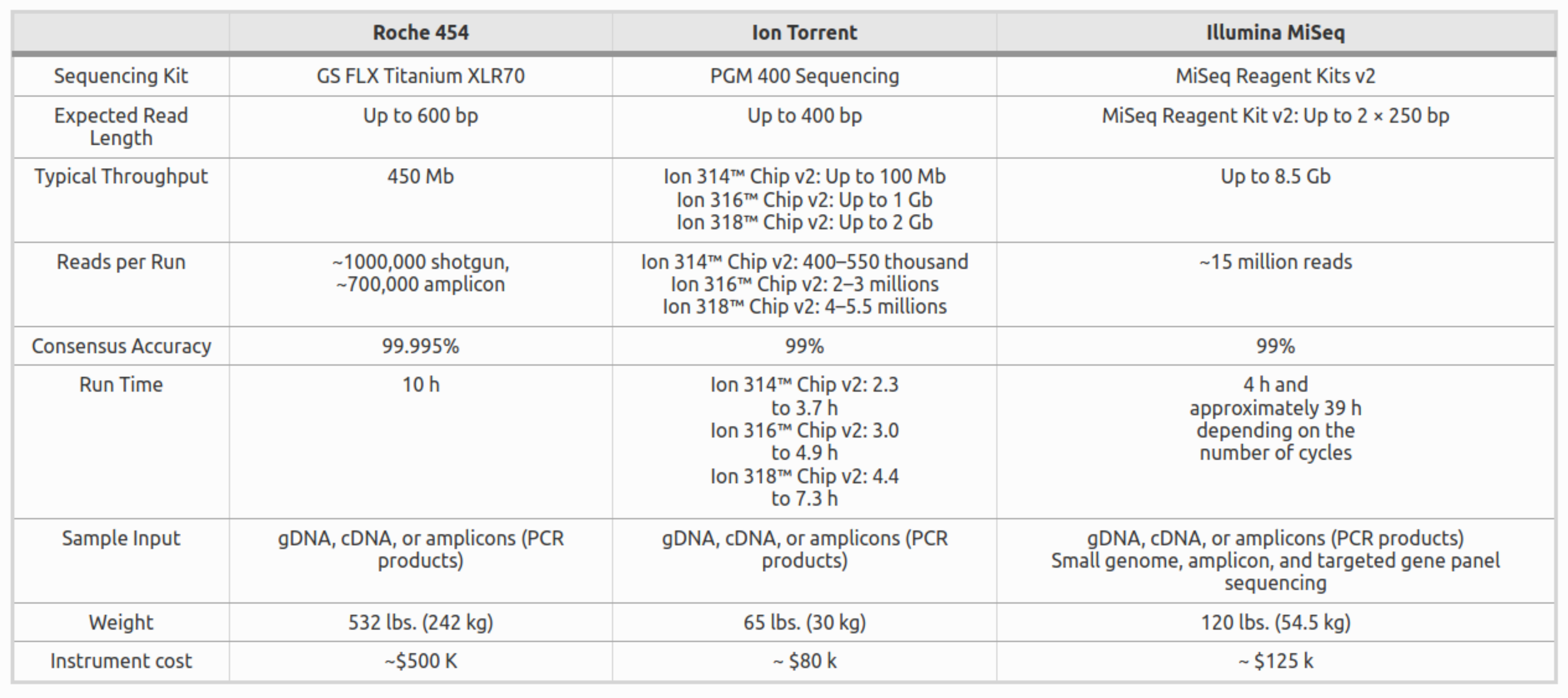
DNA extraction and preparation

Sequencing

Bioinformatics & statistical analyses

Sequencing technologies?

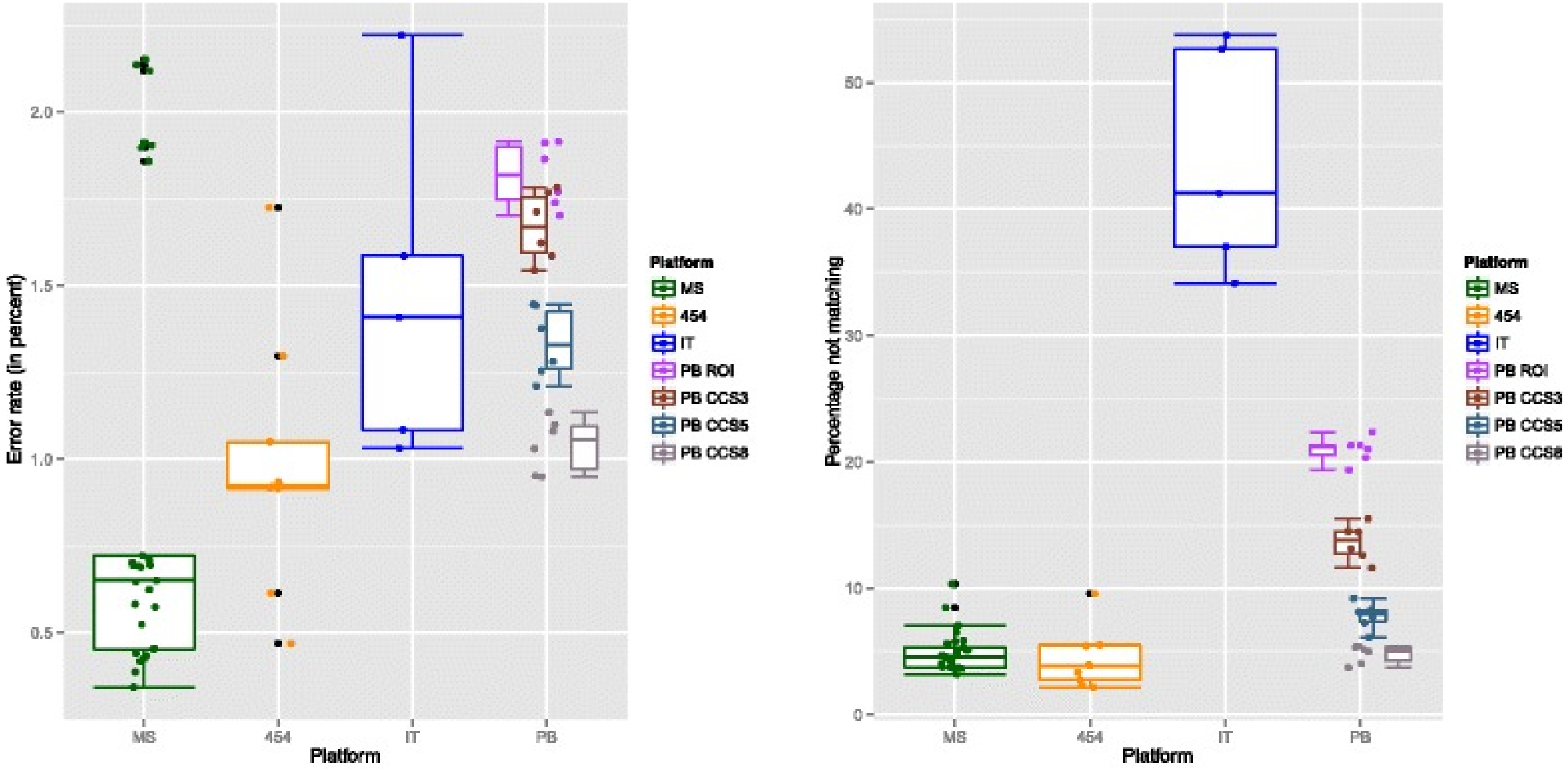
Main sequencing technologies for metabarcoding



	Roche 454	Ion Torrent	Illumina MiSeq
Sequencing Kit	GS FLX Titanium XLR70	PGM 400 Sequencing	MiSeq Reagent Kits v2
Expected Read Length	Up to 600 bp	Up to 400 bp	MiSeq Reagent Kit v2: Up to 2 × 250 bp
Typical Throughput	450 Mb	Ion 314™ Chip v2: Up to 100 Mb Ion 316™ Chip v2: Up to 1 Gb Ion 318™ Chip v2: Up to 2 Gb	Up to 8.5 Gb
Reads per Run	~1000,000 shotgun, ~700,000 amplicon	Ion 314™ Chip v2: 400–550 thousand Ion 316™ Chip v2: 2–3 millions Ion 318™ Chip v2: 4–5.5 millions	~15 million reads
Consensus Accuracy	99.995%	99%	99%
Run Time	10 h	Ion 314™ Chip v2: 2.3 to 3.7 h Ion 316™ Chip v2: 3.0 to 4.9 h Ion 318™ Chip v2: 4.4 to 7.3 h	4 h and approximately 39 h depending on the number of cycles
Sample Input	gDNA, cDNA, or amplicons (PCR products)	gDNA, cDNA, or amplicons (PCR products)	gDNA, cDNA, or amplicons (PCR products) Small genome, amplicon, and targeted gene panel sequencing
Weight	532 lbs. (242 kg)	65 lbs. (30 kg)	120 lbs. (54.5 kg)
Instrument cost	~$500 K	~ $80 k	~ $125 k

Allali et al (2017)

Sequencing errors

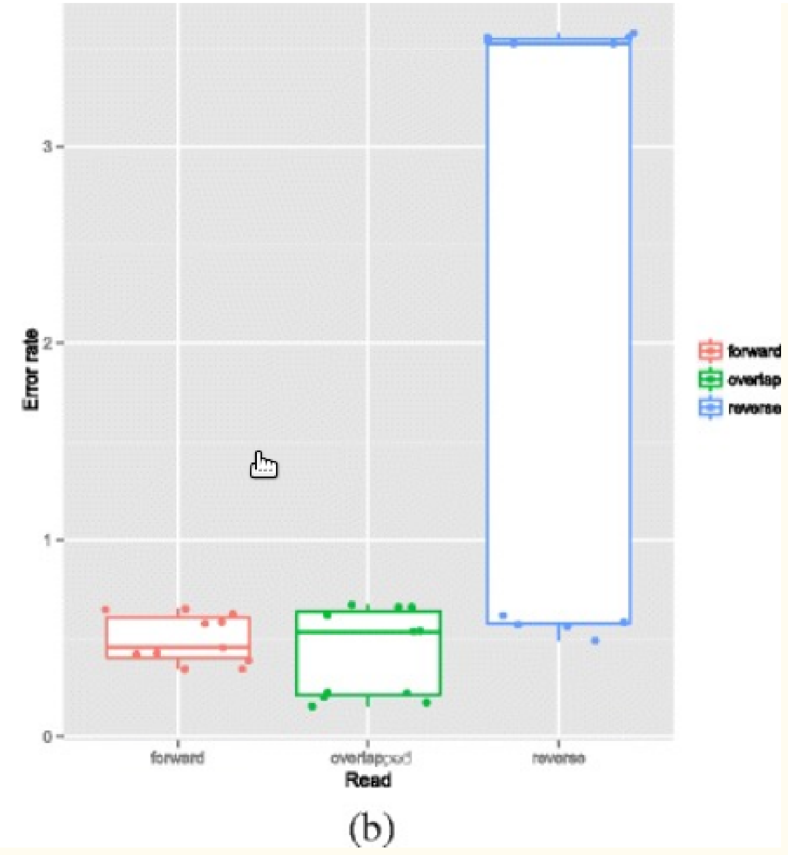


(a)

(b)

D'Amore et al (2016)

Overlapping reads allow to correct some errors

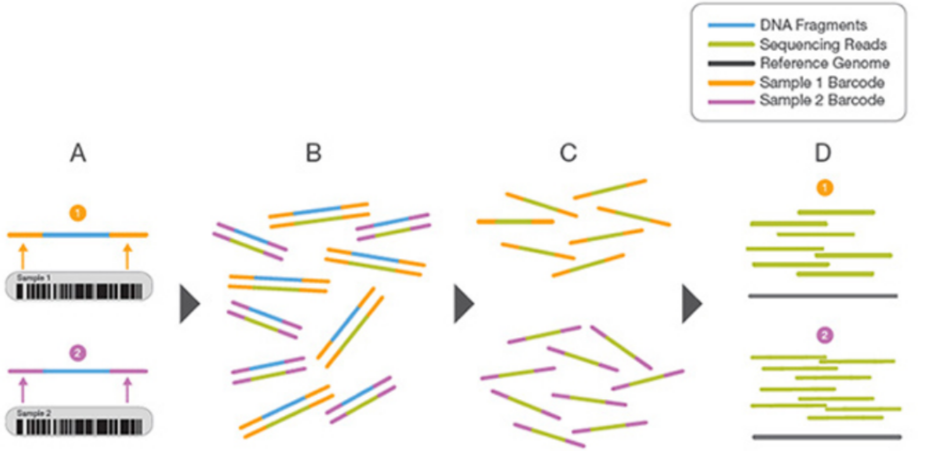


D'Amore et al (2016)

Illumina MiSeq Sequencing

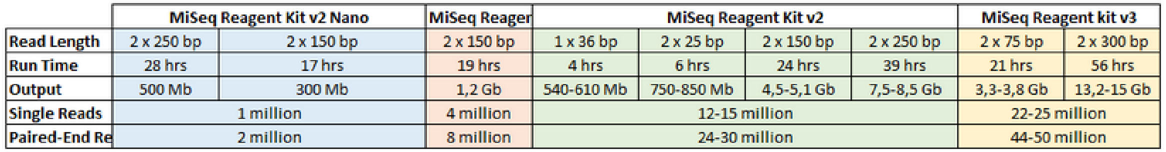
- DNA fragments are bound to the flowcell and sequenced (by synthesis) Video
- Paired-end reads allow to obtain longer fragments than 250 or 300 bp
- Low error-rate
- Substitution type miscalls are the dominant source of errors
- Abordable cost due to *multiplexing*

Multiplexing



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
D. Each set of reads is aligned to the reference sequence.

- max 384 indexes by run



	MiSeq Reagent Kit v2 Nano		MiSeq Reager	MiSeq Reagent Kit v2				MiSeq Reagent kit v3	
Read Length	2 x 250 bp	2 x 150 bp	2 x 150 bp	1 x 36 bp	2 x 25 bp	2 x 150 bp	2 x 250 bp	2 x 75 bp	2 x 300 bp
Run Time	28 hrs	17 hrs	19 hrs	4 hrs	6 hrs	24 hrs	39 hrs	21 hrs	56 hrs
Output	500 Mb	300 Mb	1,2 Gb	540-610 Mb	750-850 Mb	4,5-5,1 Gb	7,5-8,5 Gb	3,3-3,8 Gb	13,2-15 Gb
Single Reads	1 million		4 million	12-15 million				22-25 million	
Paired-End Re	2 million		8 million	24-30 million				44-50 million	

PacBio promises

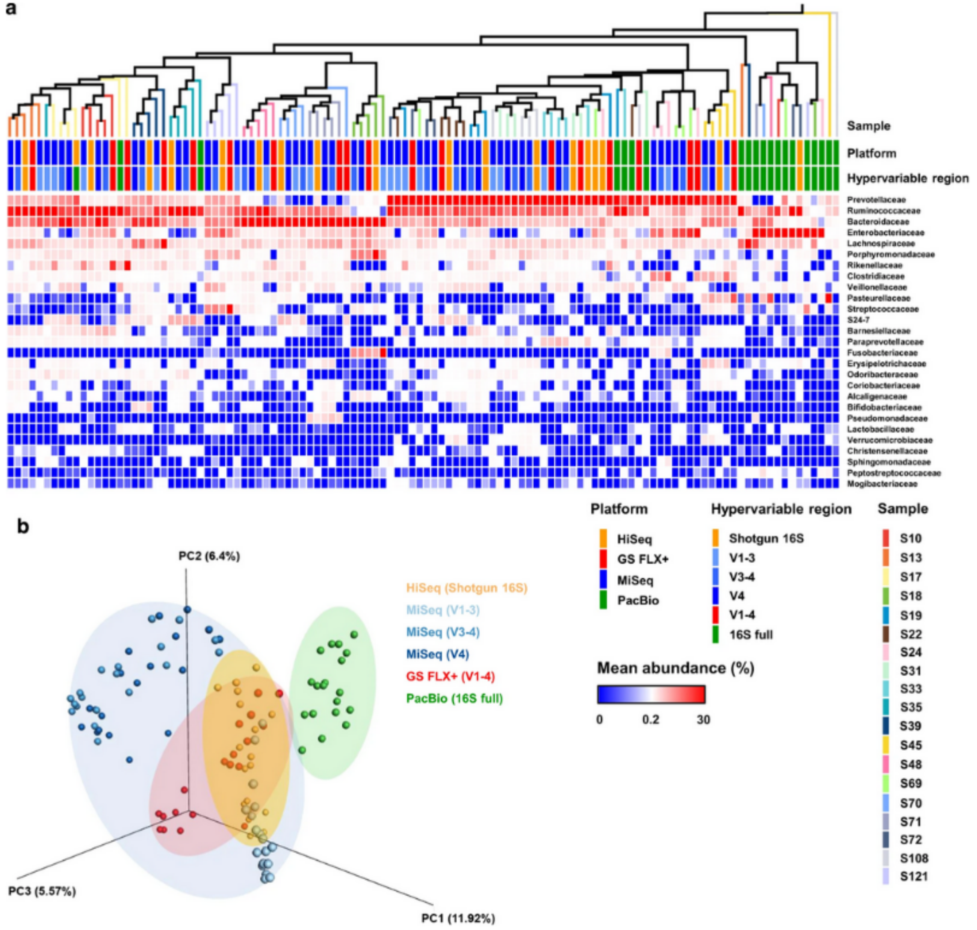
- Get the full 16S sequence!
- *We further demonstrate that full-length sequencing platforms are sufficiently accurate to resolve subtle nucleotide substitutions (but not insertions/deletions) that exist between intragenomic copies of the 16S gene.*

Johnson et al (2019)

PacBio caveats

- *Low sequencing accuracy and low coverage of terminal regions in public 16S rRNA databases deteriorate the advantages of long read length, resulting in low taxonomic resolution in amplicon sequencing of human gut microbiota*

Whon et al (2018)



Fecal samples collected from 19 human subjects were sequenced using the indicated platforms: GS FLX+ (V1–4, red), Illumina MiSeq (V1–3, light blue; V3–4, blue; V4, dark blue), and PacBio CCS (V1–9, green). Whole-genome shotgun sequences generated by Illumina HiSeq (Shotgun 16 S, orange) were included as a reference for community structure without amplification bias. (**a**) The sequence data were clustered using a UPGMA dendrogram based on the Bray-Curtis dissimilarity matrix, and samples from the same individual are shown in the same color. The relative abundances of bacterial taxa are displayed as a heatmap over 27 families (>1% relative abundance). (**b**) The sequence data were clustered by principal component analysis.

Whon et al (2018)

Sequencing biases

- Contamination between samples during the same run
- Contamination between samples during different runs (residual contaminants)
- Variability between runs: take into account for experimental plan
- Variability inside run: add some controls

Salter et al (2014)

Negative controls are important!

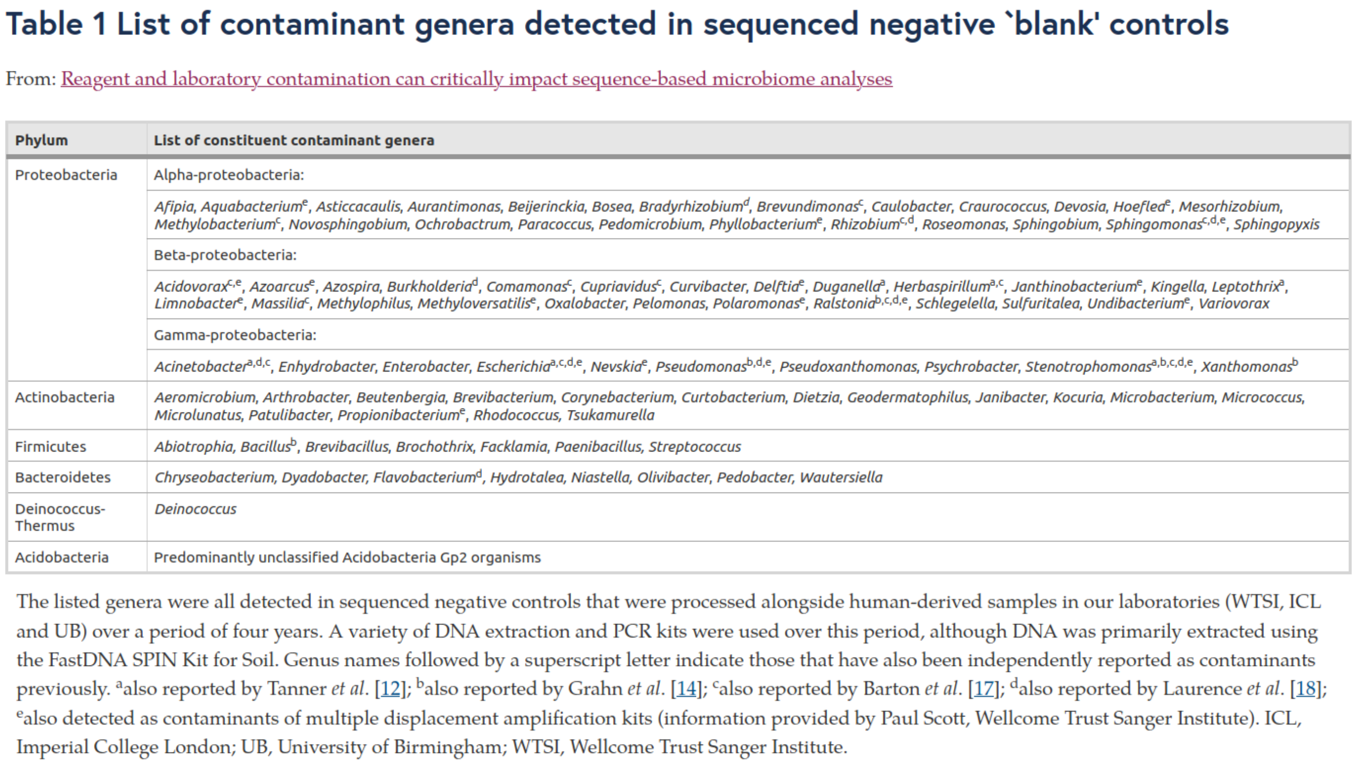


Table 1 List of contaminant genera detected in sequenced negative `blank' controls

From: Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Phylum	List of constituent contaminant genera
Proteobacteria	Alpha-proteobacteria:
	Afipia, Aquabacterium[e], *Asticcacaulis, Aurantimonas, Beijerinckia, Bosea, Bradyrhizobium*[d], *Brevundimonas*[c], *Caulobacter, Craurococcus, Devosia, Hoefled*[e], *Mesorhizobium, Methylobacterium*[c], *Novosphingobium, Ochrobactrum, Paracoccus, Pedomicrobium, Phyllobacterium*[e], *Rhizobium*[c,d], *Roseomonas, Sphingobium, Sphingomonas*[c,d,e], *Sphingopyxis*
	Beta-proteobacteria:
	Acidovorax[c,e], *Azoarcus*[e], *Azospira, Burkholderia*[d], *Comamonas*[c], *Cupriavidus*[c], *Curvibacter, Delftia*[e], *Duganella*[a], *Herbaspirillum*[a,c], *Janthinobacterium*[e], *Kingella, Leptothrix*[a], *Limnobacter*[e], *Massilia*[c], *Methylophilus, Methyloversatilis*[e], *Oxalobacter, Pelomonas, Polaromonas*[e], *Ralstonia*[b,c,d,e], *Schlegelella, Sulfuritalea, Undibacterium*[e], *Variovorax*
	Gamma-proteobacteria:
	Acinetobacter[a,d,c], *Enhydrobacter, Enterobacter, Escherichia*[a,c,d,e], *Nevskia*[e], *Pseudomonas*[b,d,e], *Pseudoxanthomonas, Psychrobacter, Stenotrophomonas*[a,b,c,d,e], *Xanthomonas*[b]
Actinobacteria	*Aeromicrobium, Arthrobacter, Beutenbergia, Brevibacterium, Corynebacterium, Curtobacterium, Dietzia, Geodermatophilus, Janibacter, Kocuria, Microbacterium, Micrococcus, Microlunatus, Patulibacter, Propionibacterium*[e], *Rhodococcus, Tsukamurella*
Firmicutes	*Abiotrophia, Bacillus*[b], *Brevibacillus, Brochothrix, Facklamia, Paenibacillus, Streptococcus*
Bacteroidetes	*Chryseobacterium, Dyadobacter, Flavobacterium*[d], *Hydrotalea, Niastella, Olivibacter, Pedobacter, Wautersiella*
Deinococcus-Thermus	*Deinococcus*
Acidobacteria	Predominantly unclassified Acidobacteria Gp2 organisms

The listed genera were all detected in sequenced negative controls that were processed alongside human-derived samples in our laboratories (WTSI, ICL and UB) over a period of four years. A variety of DNA extraction and PCR kits were used over this period, although DNA was primarily extracted using the FastDNA SPIN Kit for Soil. Genus names followed by a superscript letter indicate those that have also been independently reported as contaminants previously. [a]also reported by Tanner *et al*. [12]; [b]also reported by Grahn *et al*. [14]; [c]also reported by Barton *et al*. [17]; [d]also reported by Laurence *et al*. [18]; [e]also detected as contaminants of multiple displacement amplification kits (information provided by Paul Scott, Wellcome Trust Sanger Institute). ICL, Imperial College London; UB, University of Birmingham; WTSI, Wellcome Trust Sanger Institute.

Salter et al (2014)

Negative controls are important!

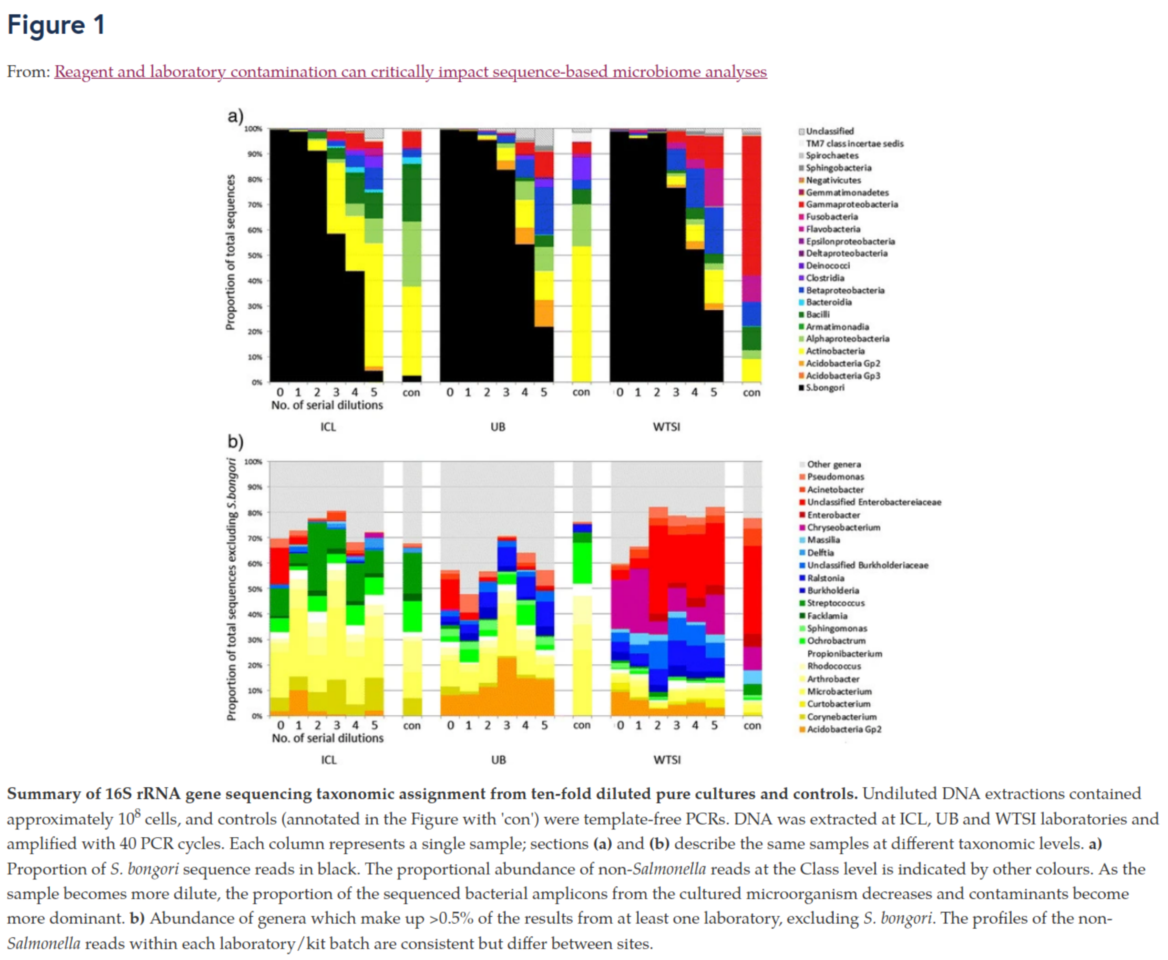


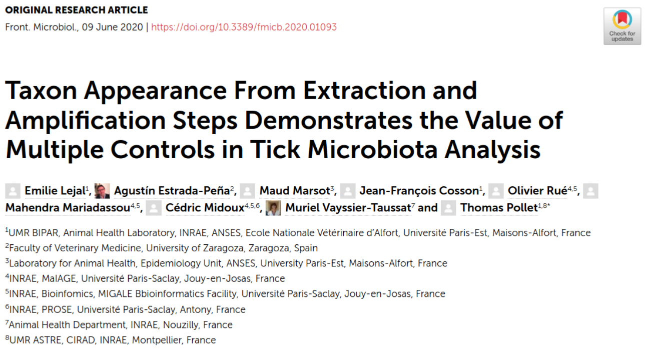
Figure 1

From: Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Summary of 16S rRNA gene sequencing taxonomic assignment from ten-fold diluted pure cultures and controls. Undiluted DNA extractions contained approximately 10^8 cells, and controls (annotated in the Figure with 'con') were template-free PCRs. DNA was extracted at ICL, UB and WTSI laboratories and amplified with 40 PCR cycles. Each column represents a single sample; sections **(a)** and **(b)** describe the same samples at different taxonomic levels. **a)** Proportion of *S. bongori* sequence reads in black. The proportional abundance of non-*Salmonella* reads at the Class level is indicated by other colours. As the sample becomes more dilute, the proportion of the sequenced bacterial amplicons from the cultured microorganism decreases and contaminants become more dominant. **b)** Abundance of genera which make up >0.5% of the results from at least one laboratory, excluding *S. bongori*. The profiles of the non-*Salmonella* reads within each laboratory/kit batch are consistent but differ between sites.

Salter et al (2014)

Illustration



ORIGINAL RESEARCH ARTICLE

Front. Microbiol., 09 June 2020 | https://doi.org/10.3389/fmicb.2020.01093

Taxon Appearance From Extraction and Amplification Steps Demonstrates the Value of Multiple Controls in Tick Microbiota Analysis

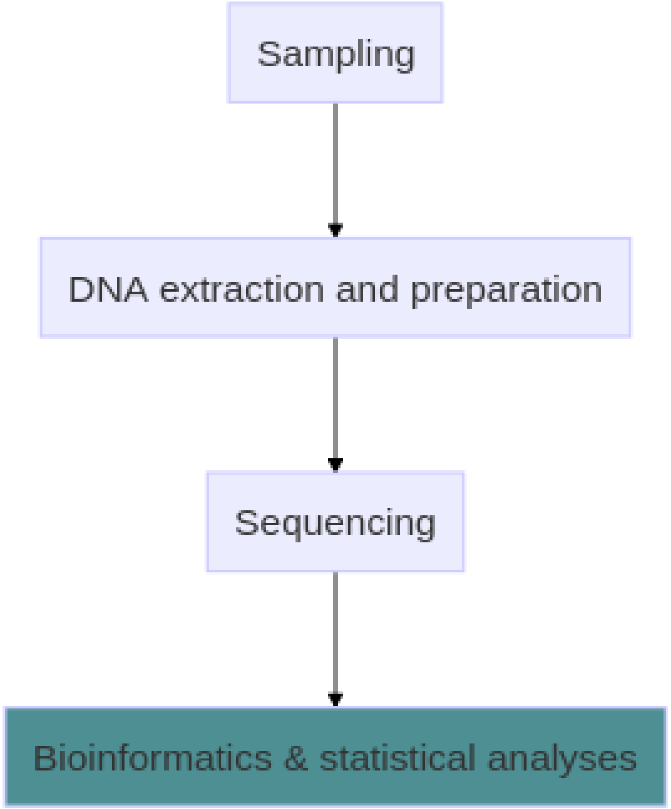
Emilie Lejal[1], Agustín Estrada-Peña[2], Maud Marsot[3], Jean-François Cosson[1], Olivier Rué[4,5], Mahendra Mariadassou[4,5], Cédric Midoux[4,5,6], Muriel Vayssier-Taussat[7] and Thomas Pollet[1,8*]

[1]UMR BIPAR, Animal Health Laboratory, INRAE, ANSES, Ecole Nationale Vétérinaire d'Alfort, Université Paris-Est, Maisons-Alfort, France
[2]Faculty of Veterinary Medicine, University of Zaragoza, Zaragoza, Spain
[3]Laboratory for Animal Health, Epidemiology Unit, ANSES, University Paris-Est, Maisons-Alfort, France
[4]INRAE, MaIAGE, Université Paris-Saclay, Jouy-en-Josas, France
[5]INRAE, Bioinfomics, MIGALE Bbioinformatics Facility, Université Paris-Saclay, Jouy-en-Josas, France
[6]INRAE, PROSE, Université Paris-Saclay, Antony, France
[7]Animal Health Department, INRAE, Nouzilly, France
[8]UMR ASTRE, CIRAD, INRAE, Montpellier, France

Here, we showed that contaminant OTUs from extraction and amplification steps can represent more than half the total sequence yield in sequencing runs, and lead to unreliable results when characterizing tick microbial communities. We thus strongly advise the routine use of negative controls in tick microbiota studies, and more generally in studies involving low biomass samples

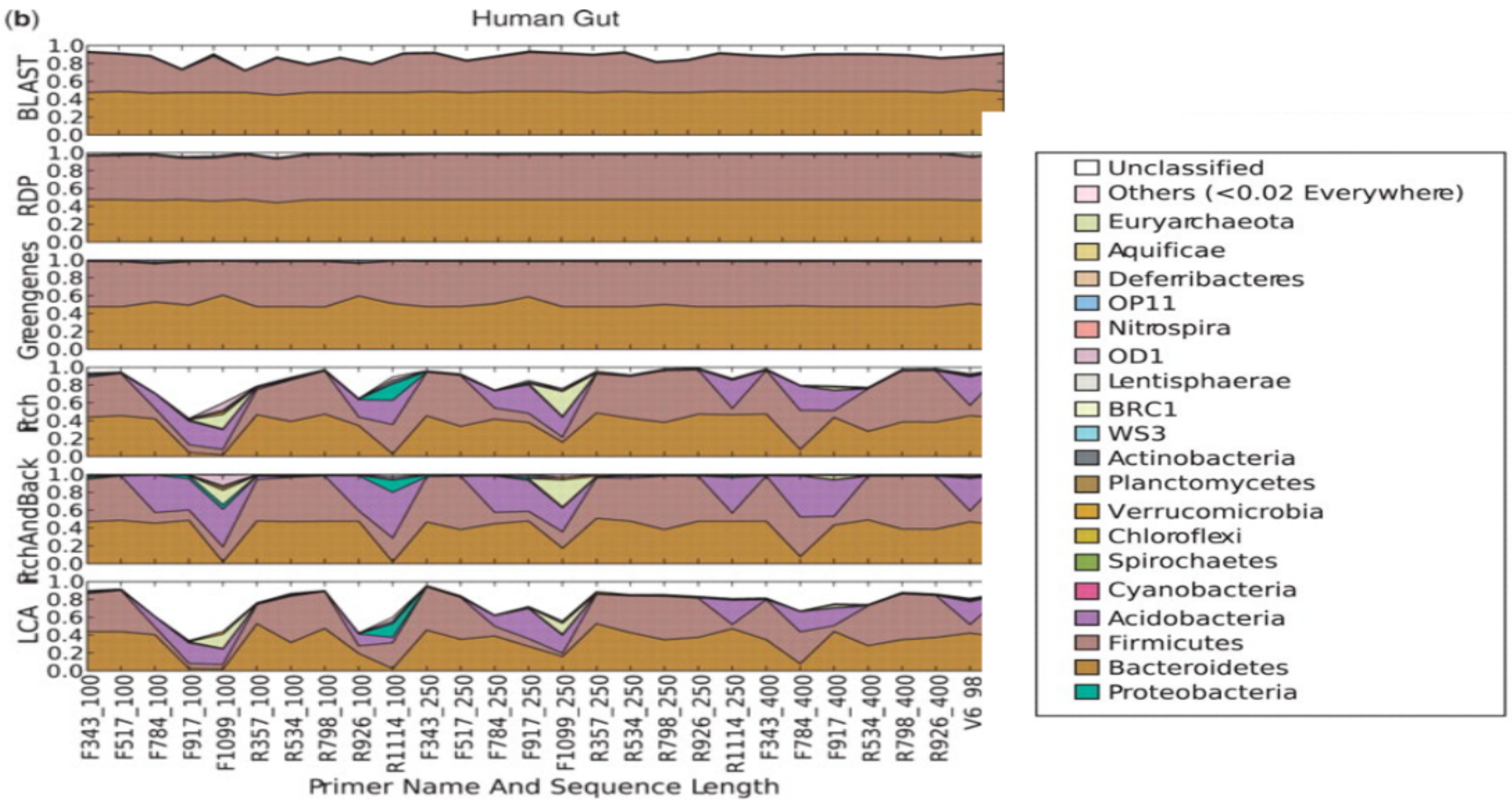
Lejal et al (2019)

Bioinformatics & Biostatistics


Sampling
DNA extraction and preparation
Sequencing
Bioinformatics & statistical analyses

- Tools
- Databases
- Normalization
- Diversity indices

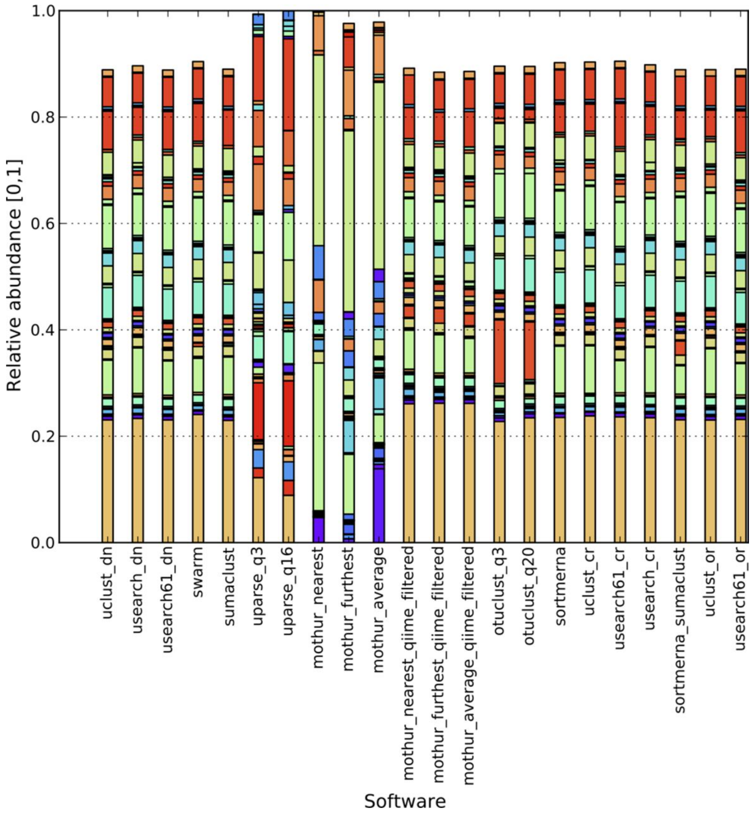
Impact of method and targeted region



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

Liu et al (2008)

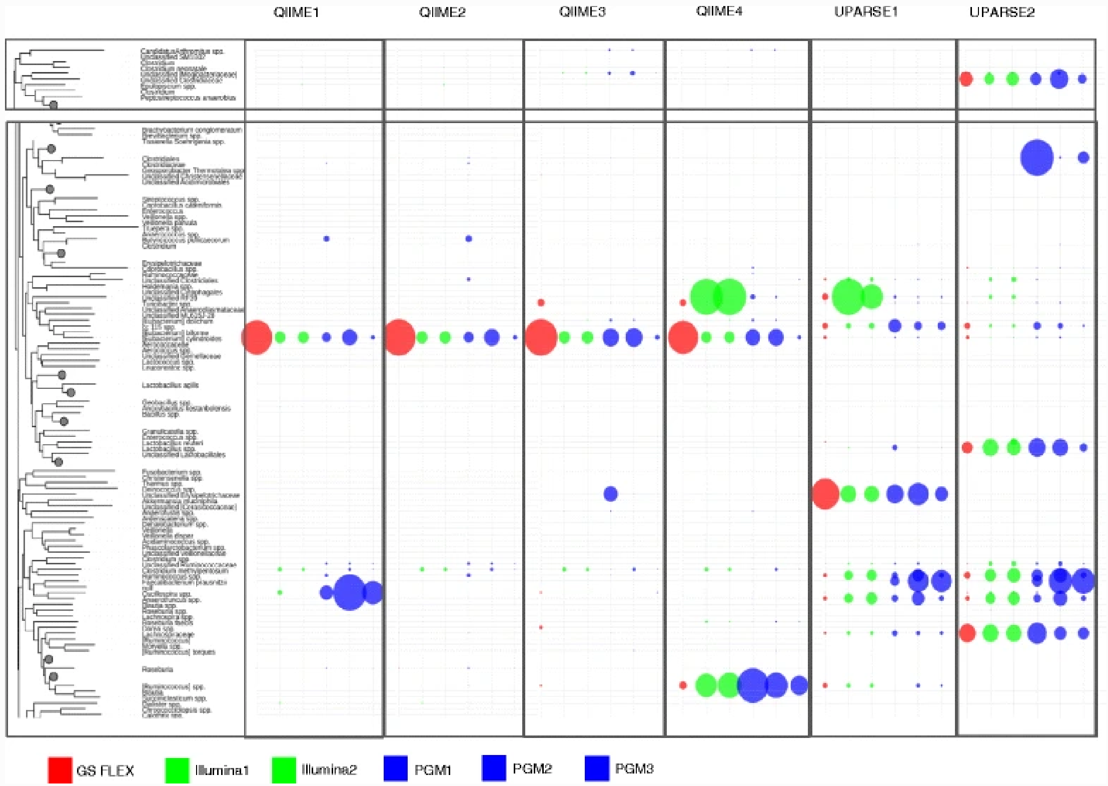
Benchmarks



- Be cautious with benchmarks!
- Input data are never identical -> results are never exactly the same

Kopylova et al (2016)

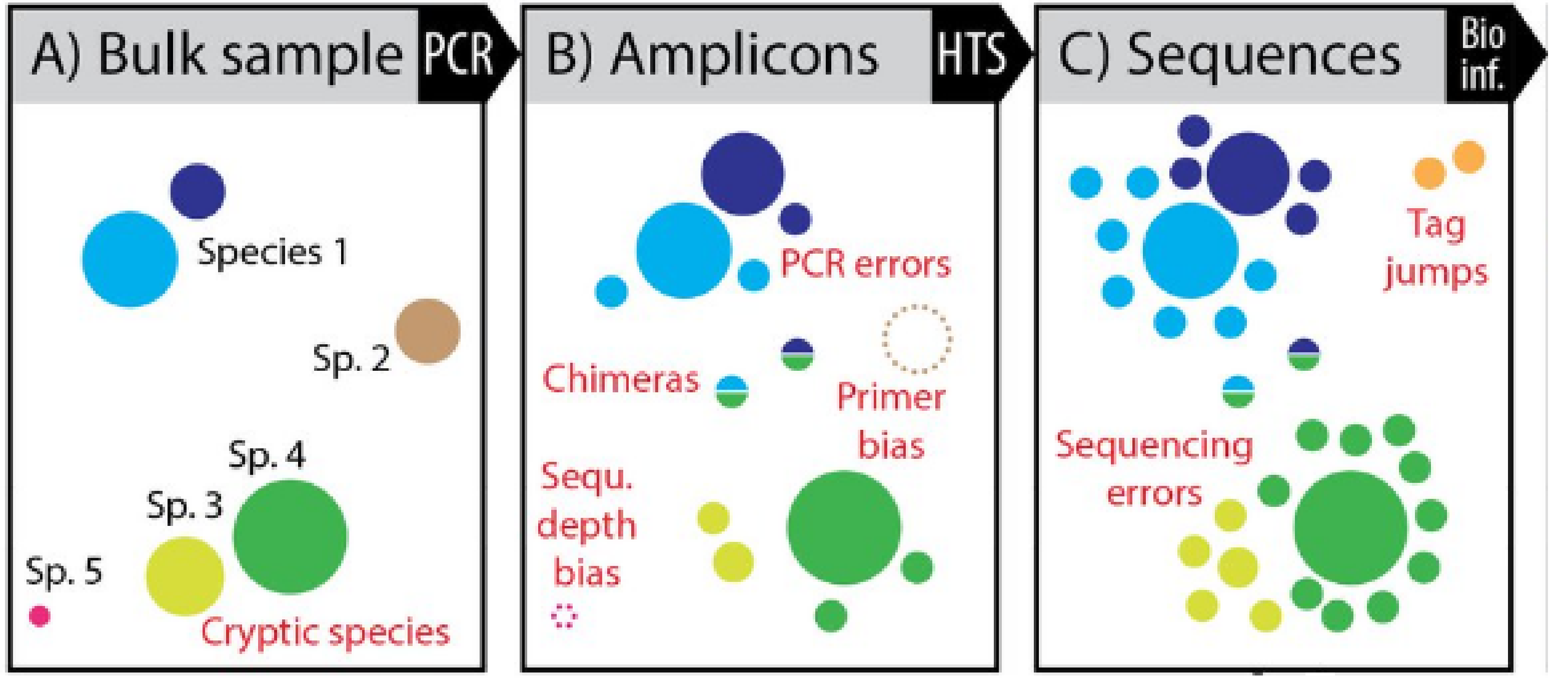
Benchmarks

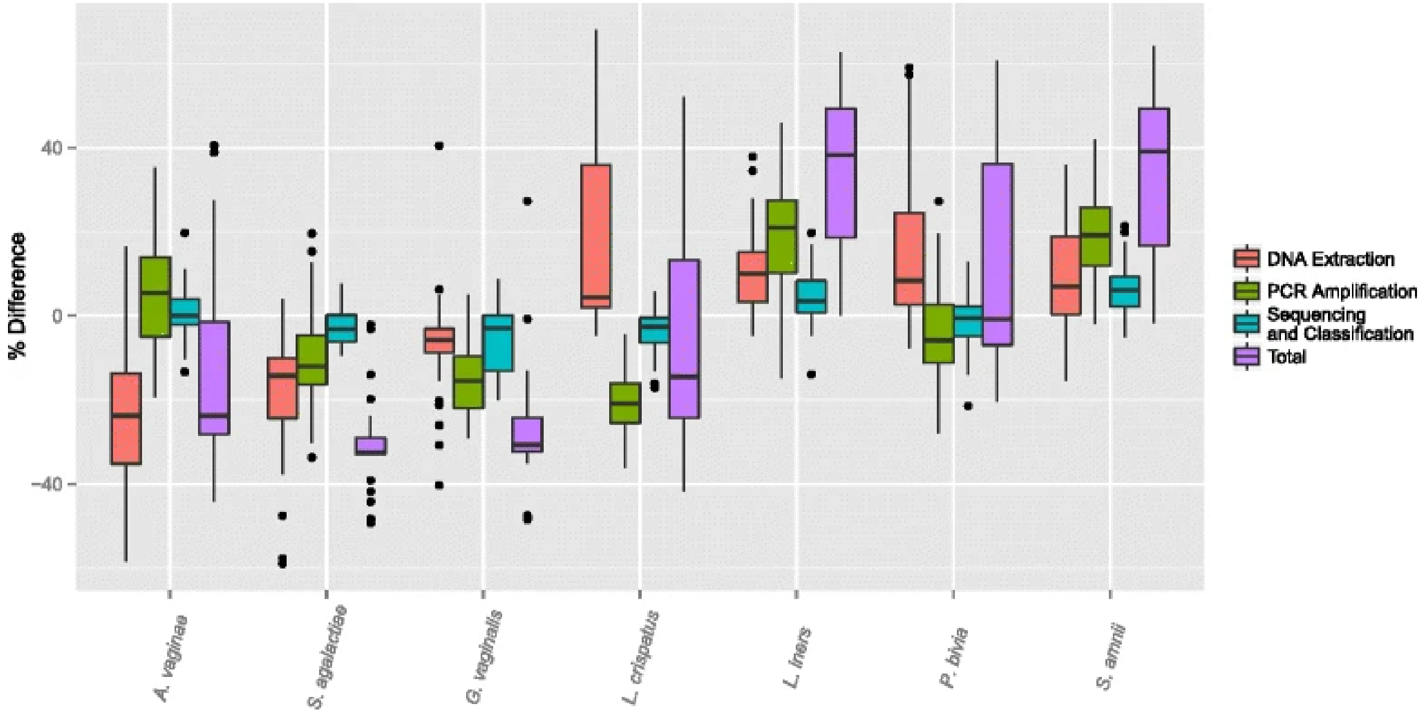


Selected differences in relative abundances of the most impacted taxa according to data generated by different platforms (indicated by different colors) and bioniformatic analysis pipelines (indicated across the top). The full figure can be seen in Additional file 2: Figure S2

Allali et al (2017)

Conclusion 1: sequencing data do not contain exactly what you sampled...


A) Bulk sample
PCR
Species 1
Sp. 2
Sp. 4
Sp. 3
Sp. 5
Cryptic species
B) Amplicons
HTS
PCR errors
Chimeras
Primer bias
Sequ. depth bias
C) Sequences
Bio inf.
Tag jumps
Sequencing errors

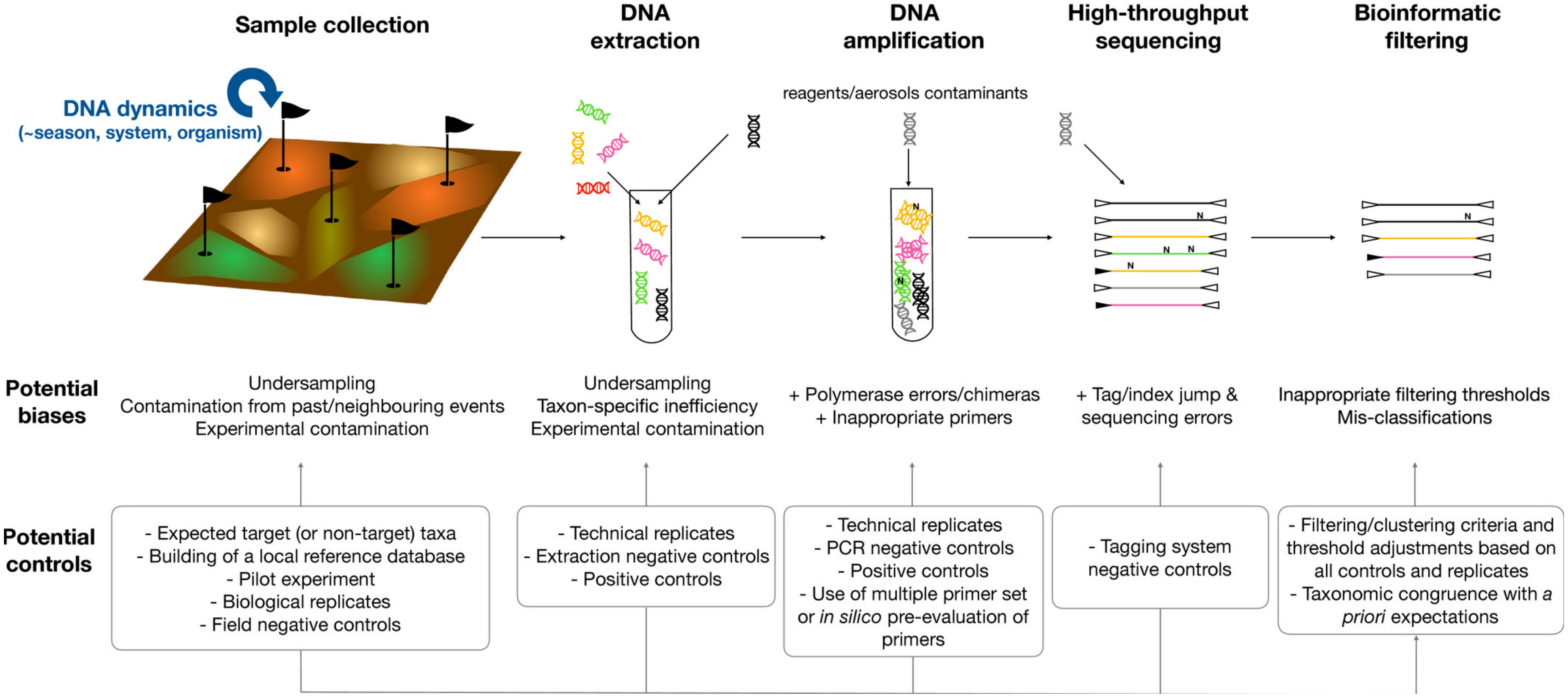


Observed bias by bacterium. The observed bias (the observed minus the actual proportions) for each bacterium in the experimental design due to the different effects of our DNA Extraction, PCR amplification, and sequencing and taxonomic classification protocols. The total bias is also plotted for each bacterium. For each box and whisker plot, only the samples including the bacterium were included.

Brooks et al (2015)

Conclusion 2: but keep biases in mind for analyzing your data!

Summary



Sample collection → **DNA extraction** → **DNA amplification** → **High-throughput sequencing** → **Bioinformatic filtering**

Potential biases

- Sample collection: Undersampling; Contamination from past/neighbouring events; Experimental contamination
- DNA extraction: Undersampling; Taxon-specific inefficiency; Experimental contamination
- DNA amplification: + Polymerase errors/chimeras; + Inappropriate primers
- High-throughput sequencing: + Tag/index jump & sequencing errors
- Bioinformatic filtering: Inappropriate filtering thresholds; Mis-classifications

Potential controls

- Sample collection:
  - Expected target (or non-target) taxa
  - Building of a local reference database
  - Pilot experiment
  - Biological replicates
  - Field negative controls
- DNA extraction:
  - Technical replicates
  - Extraction negative controls
  - Positive controls
- DNA amplification:
  - Technical replicates
  - PCR negative controls
  - Positive controls
  - Use of multiple primer set or *in silico* pre-evaluation of primers
- High-throughput sequencing:
  - Tagging system negative controls
- Bioinformatic filtering:
  - Filtering/clustering criteria and threshold adjustments based on all controls and replicates
  - Taxonomic congruence with *a priori* expectations

Key advice

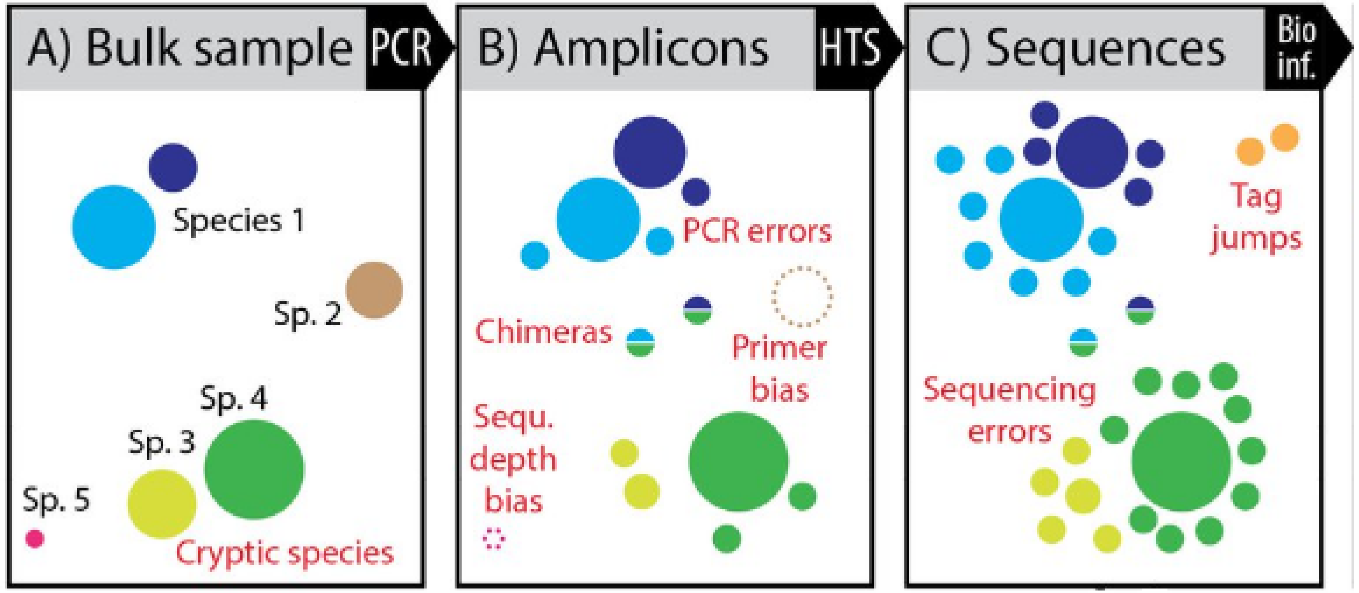
- Discuss with everyone involved in the experiment, from the field technician to the statistician
- **Each choice affects the following steps!**

Bioinformatics

The aim is to find the correct amplicon sequences and their abundances for each sample

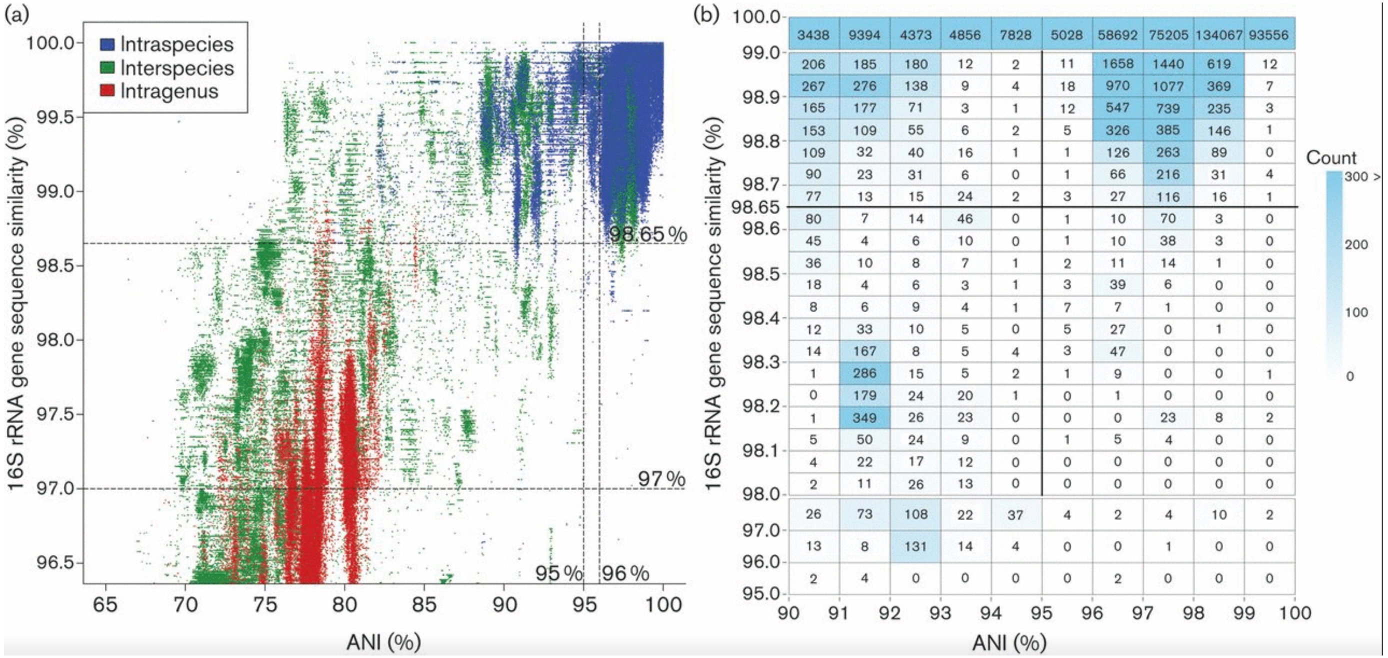
OTU	Affiliation	Sample1	Sample2	Sample3
OTU1	SpeciesA	0	500	0
OTU2	GenusA	200	41	100
OTU3	SpeciesB	1000	100	1000

Step 1: construct real amplicon sequences



Step 2: Assign a taxonomy to sequences

- Is that easy?



3438	9394	4373	4856	7828	5028	58692	75205	134067	93556
206	185	180	12	2	11	1658	1440	619	12
267	276	138	9	4	18	970	1077	369	7
165	177	71	3	1	12	547	739	235	3
153	109	55	6	2	5	326	385	146	1
109	32	40	16	1	1	126	263	89	0
90	23	31	6	0	1	66	216	31	4
77	13	15	24	2	3	27	116	16	1
80	7	14	46	0	1	10	70	3	0
45	4	6	10	0	1	10	38	3	0
36	10	8	7	1	2	11	14	1	0
18	4	6	3	1	3	39	6	0	0
8	6	9	4	1	7	7	1	0	0
12	33	10	5	0	5	27	0	1	0
14	167	8	5	4	3	47	0	0	0
1	286	15	5	2	1	9	0	0	1
0	179	24	20	1	0	1	0	0	0
1	349	26	23	0	0	0	23	8	2
5	50	24	9	0	1	5	4	0	0
4	22	17	12	0	0	0	0	0	0
2	11	26	13	0	0	0	0	0	0
26	73	108	22	37	4	2	4	10	2
13	8	131	14	4	0	0	1	0	0
2	4	0	0	0	0	2	0	0	0

Bioinformatics solutions

- MG-RAST (2008)
- Mothur (2009)
- Qiime (2010)
- UPARSE (2013)
- FROGS (2014)
- DADA2 (2016)
- Qiime2 (2019)
- …

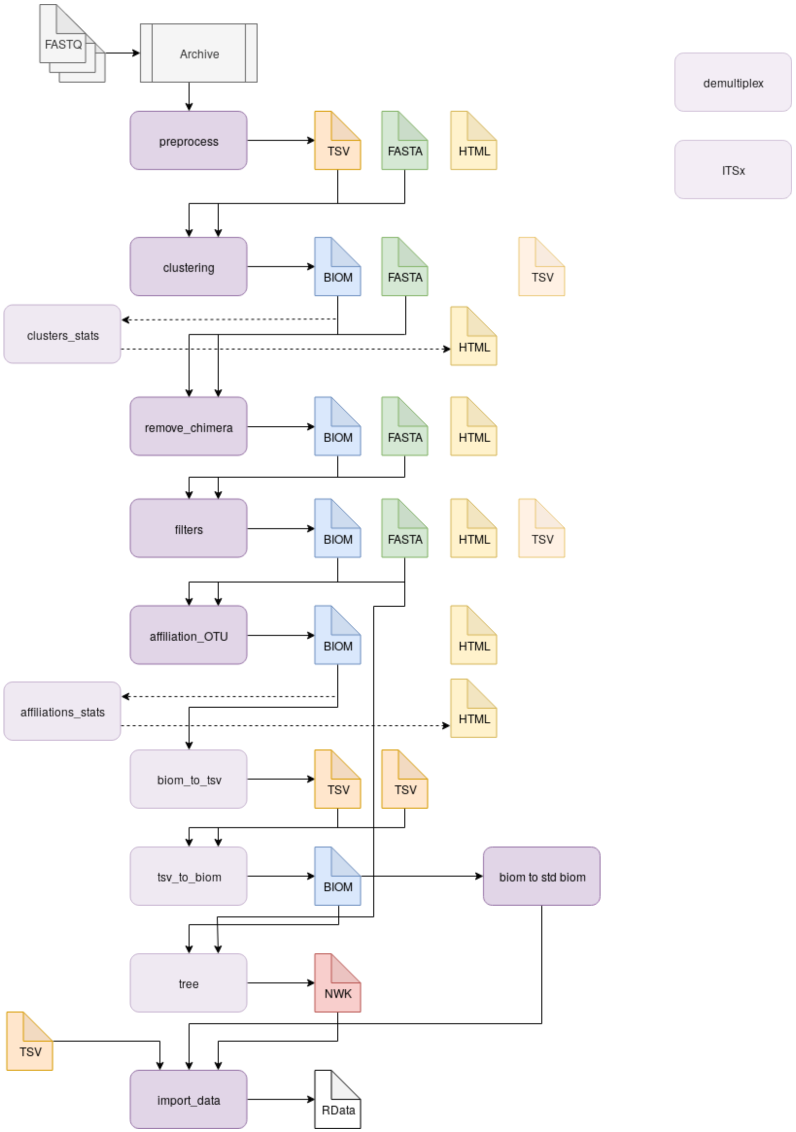
Main differences

- Ease of use
  - command line vs graphical interfaces
  - fitting complexity
- Scalling
- Paradigm: Clustering or denoising
- Chimera detection
- Taxonomic affiliation method
  - with training set
  - blast alignment

FROGS: Find Rapidly OTUs with Galaxy Solution

- Easy to use for biologists
- Last updated and adapted tools
- Innovative affiliation tag to highlight databases conflicts and uncertainties
- Designed by a group of experts of metabarcoding analyses
- Better accuracy than other tools from 16S and ITS simulated and real data
- Complete informations

Escudié et al (2007)


FASTQ
Archive
preprocess
TSV
FASTA
HTML
demultiplex
ITSx
clustering
BIOM
FASTA
TSV
clusters_stats
HTML
remove_chimera
BIOM
FASTA
HTML
filters
BIOM
FASTA
HTML
TSV
affiliation_OTU
BIOM
HTML
affiliations_stats
HTML
biom_to_tsv
TSV
TSV
tsv_to_biom
BIOM
biom to std biom
tree
NWK
TSV
import_data
RData

Switch to TP: Galaxy

Sequencing data

Content of sequenced fragments

- The expected amplicon sequence

ACTGGGTGTAAGAGCT

- The primers are sequenced too

ACTGACTGGGTGTAAGAGCTCTTA

- With two fragments:
  - R1 **ACTGACTGGGTGTAAG**
  - R2 **TAAGAGCTCTTACACC**

Content of sequenced fragments in multiplexed file

- Barcodes are added to each extremity

TTTTACTGACTGGGTGTAAGAGCTCTTACCCC

- With two fragments:
  - R1 **TTTTACTGACTGGGTG**
  - R2 **GGGGTAAGAGCTCTTA**

Demultiplexing

- Assign each read to FASTQ files depending on barcode found
- BARCODE FILE is expected to be tabular:
    - first column corresponds to the sample name (unique, without space)
    - second to the forward sequence barcode used (None if only reverse barcode)
    - optional third is the reverse sequence barcode (optional)

Switch to TP: FROGS Demultiplex

FASTQ

FASTQ syntax

The FASTQ format consists of 4 sections:

1. A FASTA-like header, but instead of the `>` symbol it uses the `@` symbol. This is followed by an ID and more optional text, similar to the FASTA headers.
2. The second section contains the measured sequence (typically on a single line), but it may be wrapped until the `+` sign starts the next section.
3. The third section is marked by the `+` sign and may be optionally followed by the same sequence id and header as the first section
4. The last line encodes the quality values for the sequence in section 2, and must be of the same length as section 2.

FASTQ syntax

Example

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

FASTQ quality

The weird characters in the 4th section are the so called “encoded” numerical values. Each character represents a numerical value: a so-called *Phred score*, encoded via a single letter encoding.

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|    |    |    |    |    |    |    |    |
0....5...10...15...20...25...30...35...40
|    |    |    |    |    |    |    |    |
worst................................best
```

FASTQ quality

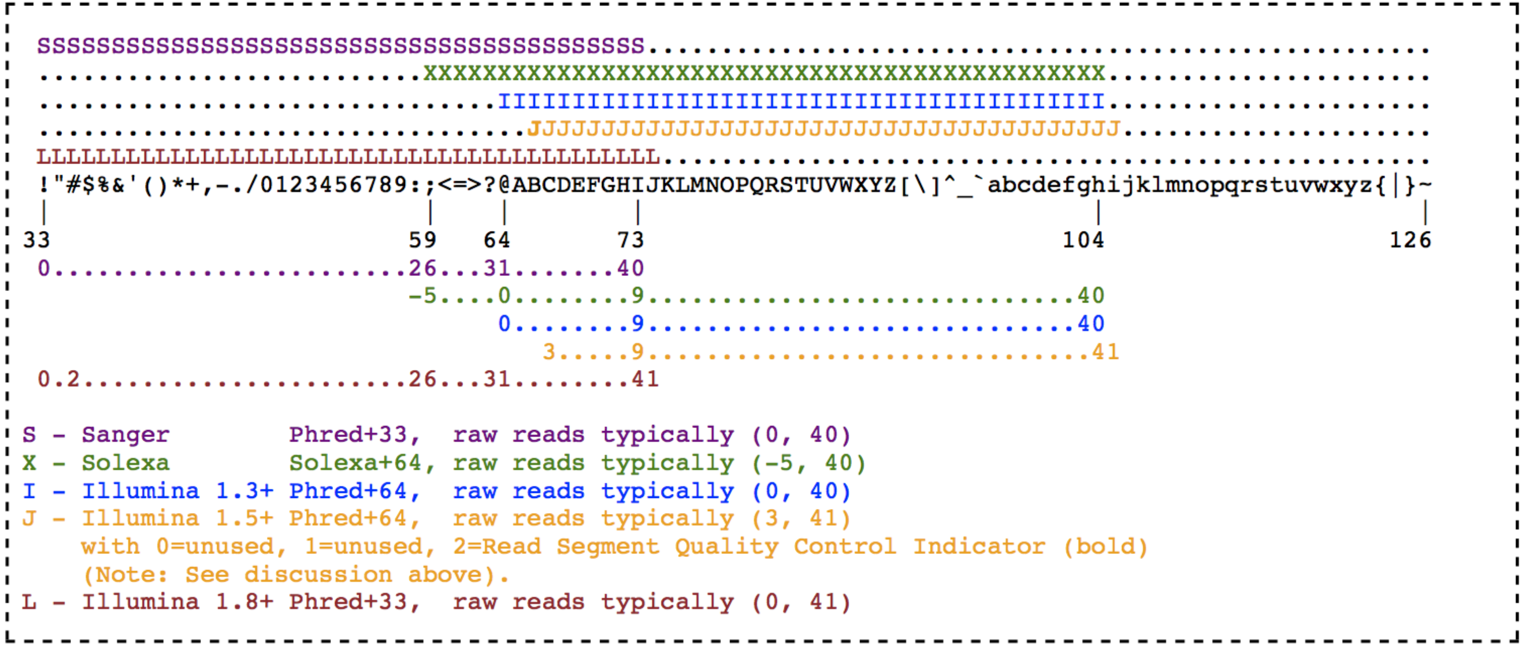
The quality values of the FASTQ files are on top. The numbers in the middle of the scale from 0 to 40 are called Phred scores. The numbers represent the error probabilities via the formula:

Error=10^(-P/10) It is basically summarized as:

- P=0 means 1/1 (100% probability of error)
- P=10 means 1/10 (10% probability of error)
- P=20 means 1/100 (1% probability of error)
- P=30 means 1/1000 (0.1% probability of error)
- P=40 means 1/10000 (0.01% probability of error)

FASTQ quality encoding specificities

There was a time when instrumentation makers could not decide at what character to start the scale. The **current standard** shown above is the so-called Sanger (+33) format where the ASCII codes are shifted by 33. There is the so-called +64 format that starts close to where the other scale ends.



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0.........................26...31.......40
                          -5....0........9..............................40
                                0........9..............................40
                                   3.....9...............................41
 0.2.......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

FASTQ files

- R1

```
@id1:1
ACTGACTGGGTGTAAG
+
EF!![!:;;;;;::;A
```

- R2

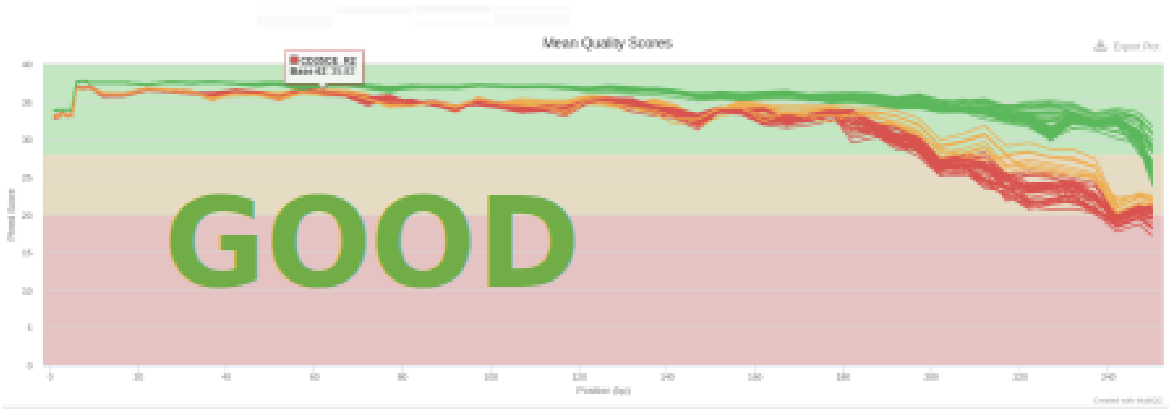
```
@id1:2
TAAGAGCTCTTACACC
+
;:,??!!???;..FFF
```

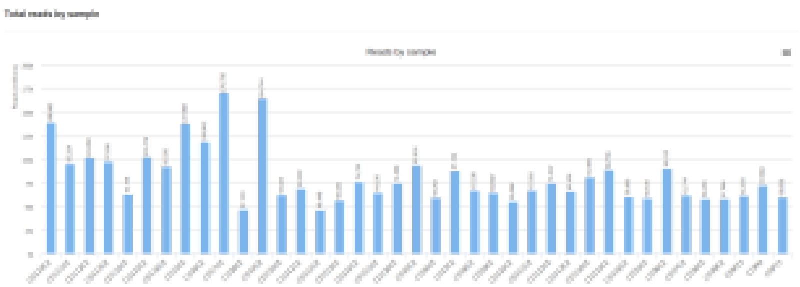
FASTQ files

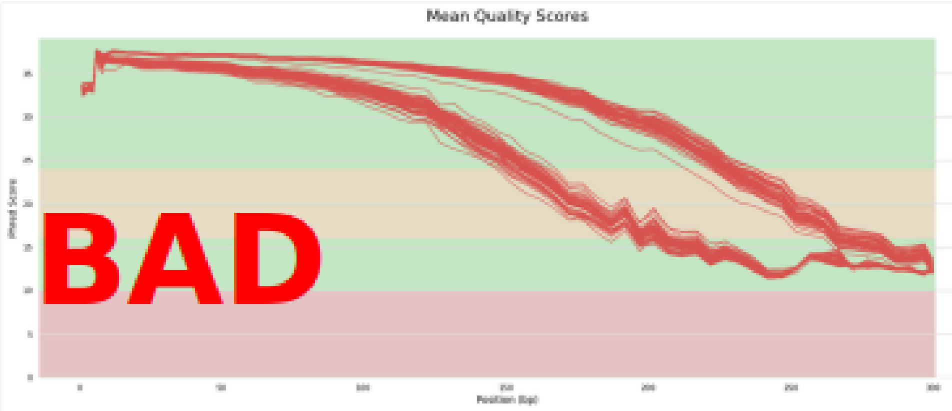
- It is crucial to check the quality of the raw data
    - expected number of files?
    - expected number of reads per file?
    - data quality?
- Do not start an analysis if something wrong
    - Unusefull if some data are missing
    - You can (have to) discuss with the sequencing platform to understand

Switch to TP: Quality control

Quality profiles





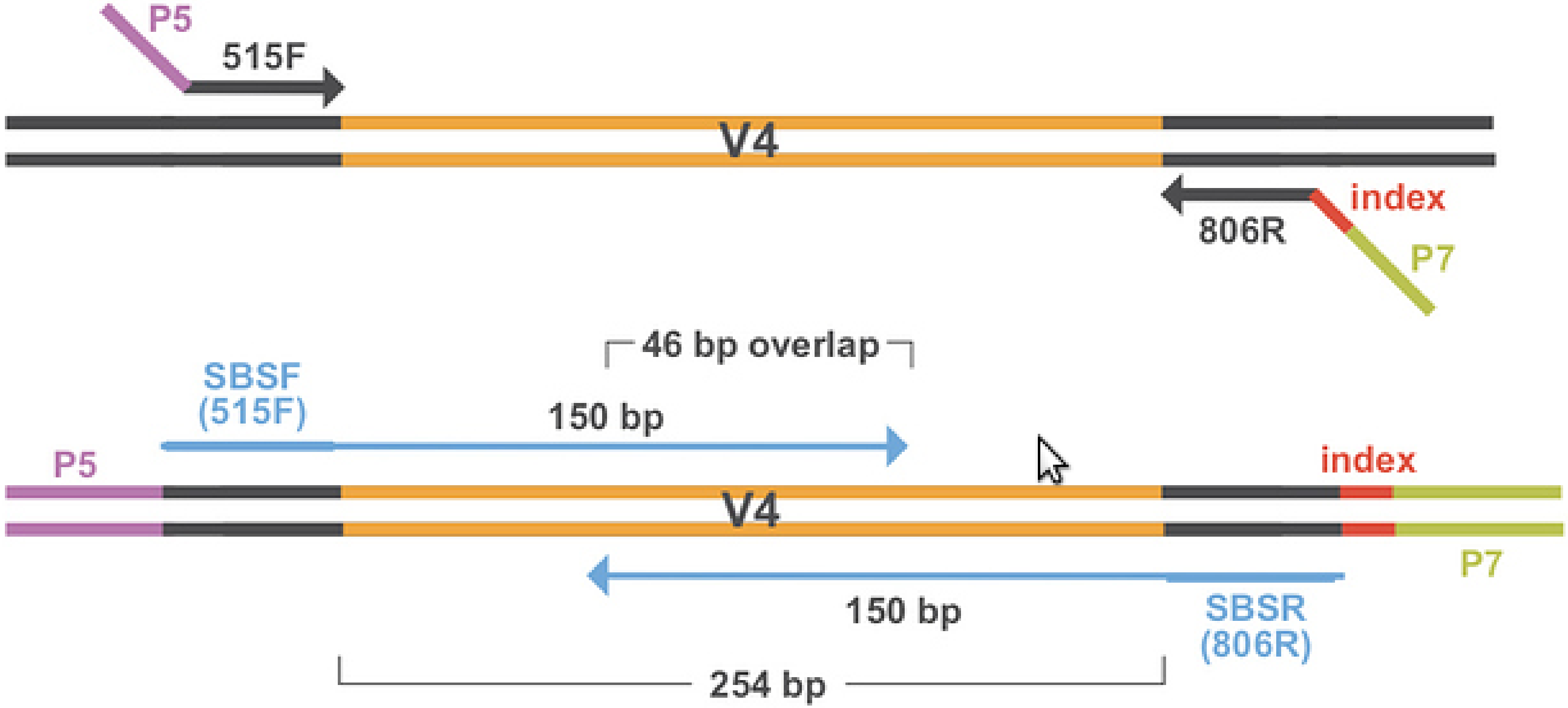


Preprocess

Preprocess

- Remove non-biological informations
    - primers, barcodes, remaining sequencing primers…
- Filter on length
- Filter on nucleotide content
- Overlap reads if possible

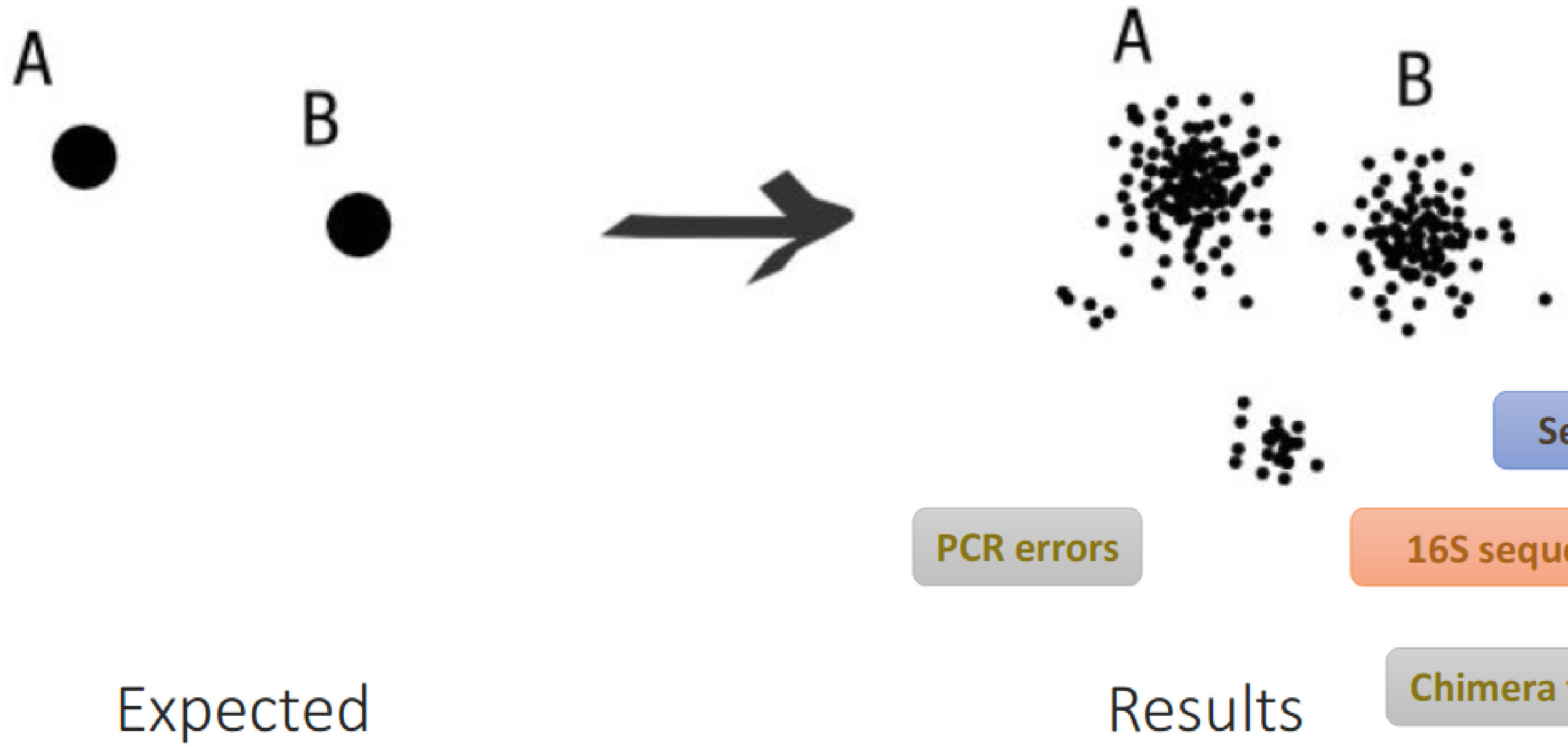
Overlap


P5
515F
V4
index
806R
P7
46 bp overlap
SBSF
(515F)
150 bp
P5
index
V4
P7
150 bp
SBSR
(806R)
254 bp

Switch to TP: FROGS Preprocess

Clustering

Sequencing data are noised



Sequencing errors

PCR errors

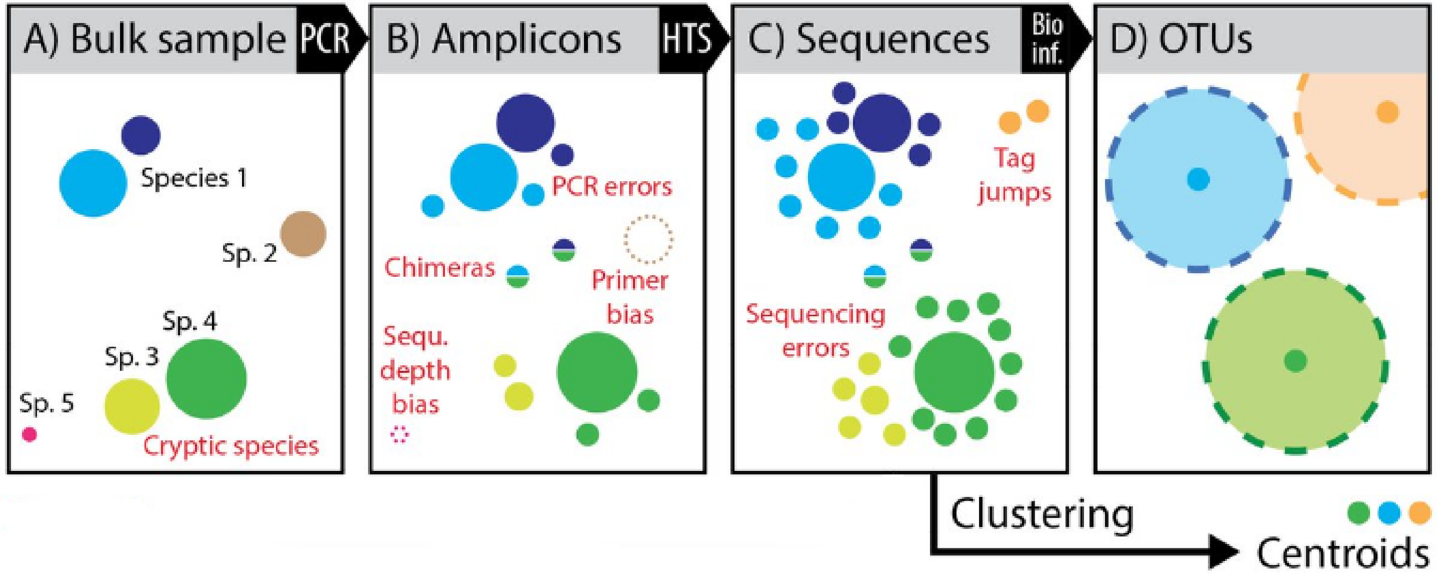
16S sequence diversity

Chimera formation

Contamination

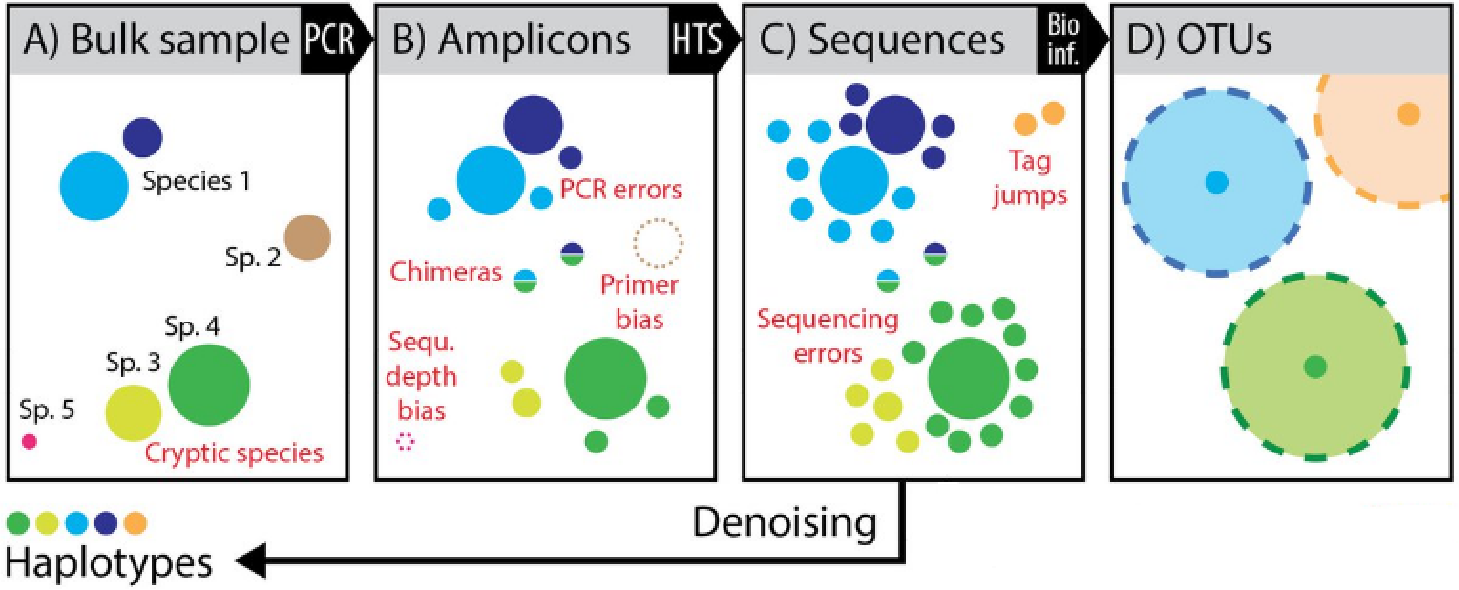
OTU paradigm

- Operational Taxonomic Unit



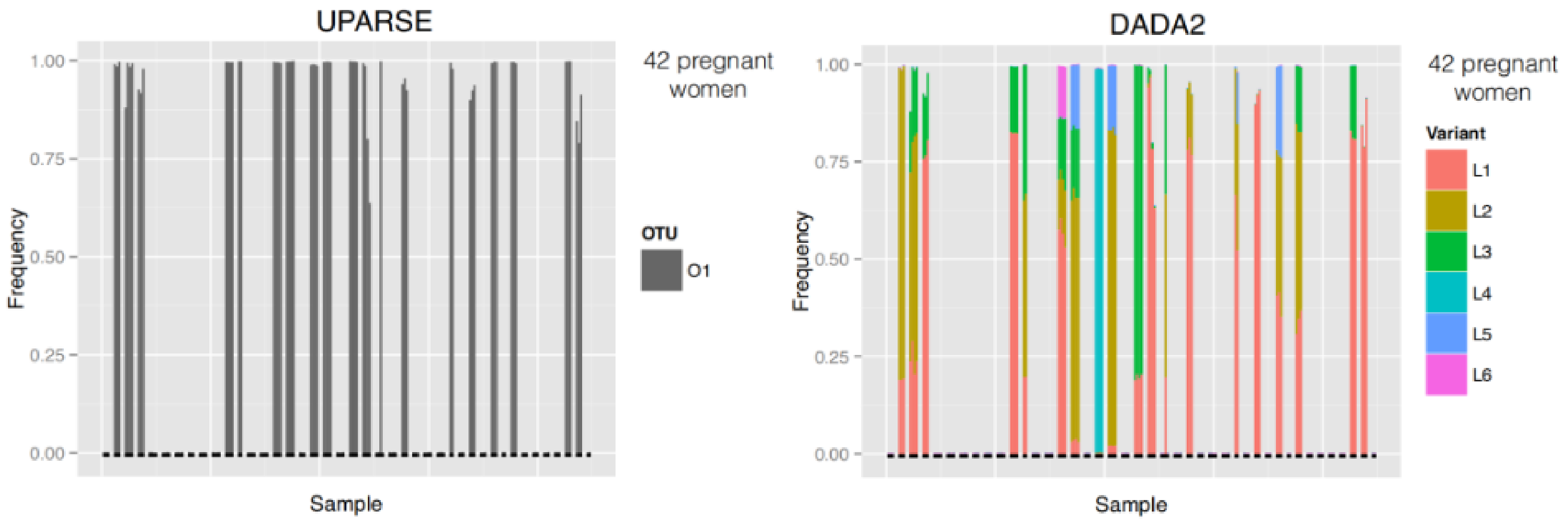
ASV paradigm

- Amplicon Sequence Variants

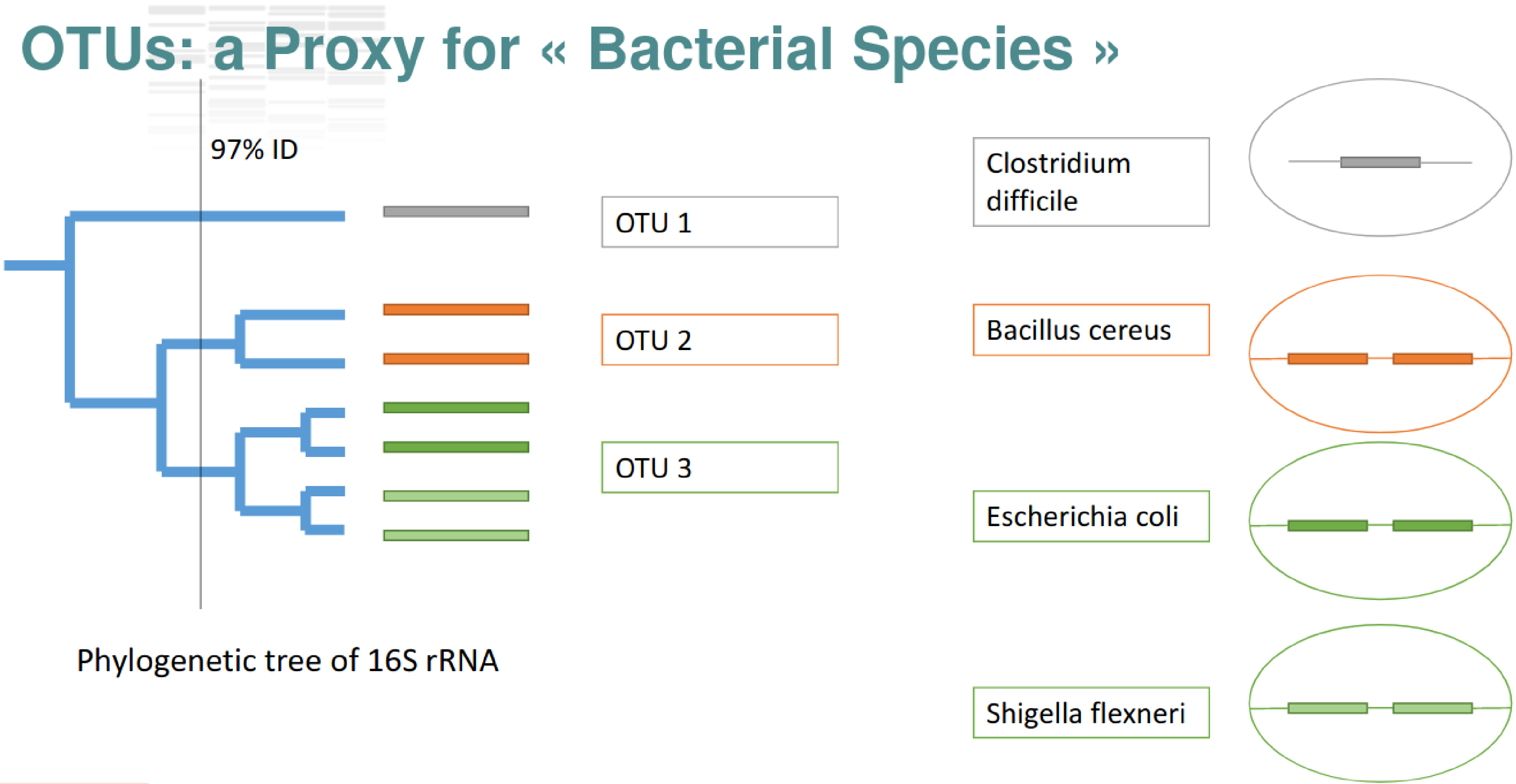


ASV are inferred by a de novo process in which biological sequences are discriminated from errors on the basis of the expectation that biological sequences are more likely to be repeatedly observed than are error-containing sequences.

ASV promises



Operational Taxonomic Units



OTUs: a Proxy for « Bacterial Species »

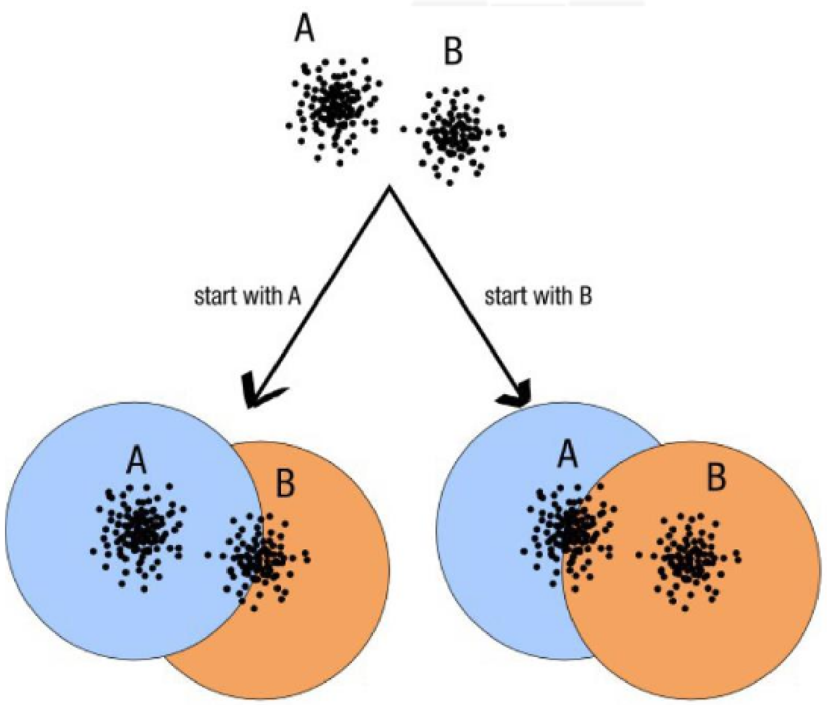
Phylogenetic tree of 16S rRNA

OTUs construction strategies

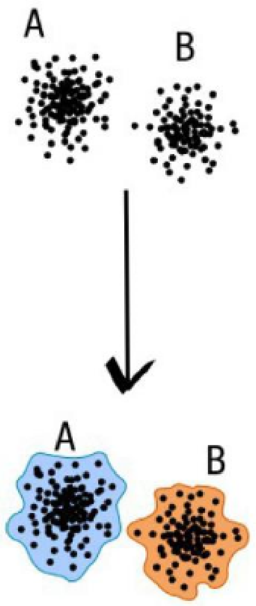
- De novo OTU picking
  - by choosing a fixed sequenced dissimilarity
  - by relying on a small local linking threshold, representing the maximum number of differences between two amplicons
- Closed-reference OTU picking
  - by using a reference databank
  - discards all reads not similar to the reference databank
- Open-reference OTU picking
  - by using a reference databank
  - de novo clusters remaining reads

Fixed sequence dissimilarity: the traditional 97%...

… is input order dependent

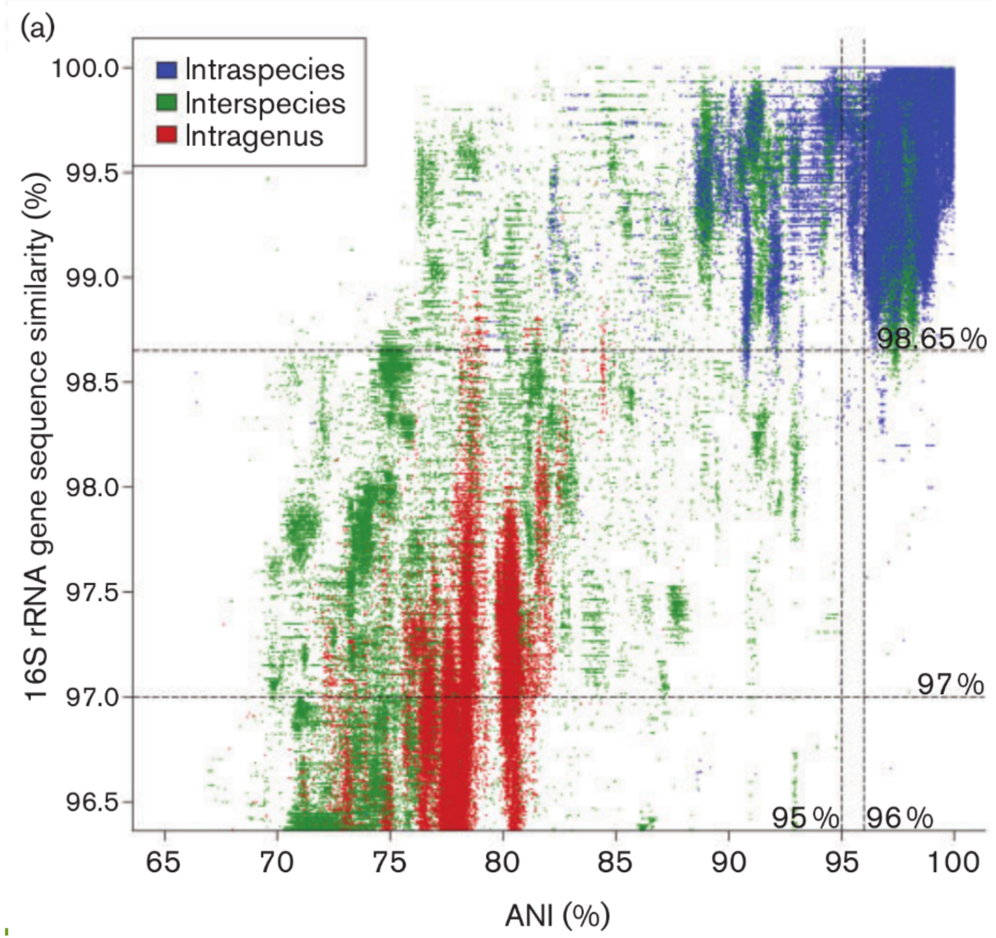


decreasing length,
decreasing abundance,
external references



natural limits of clusters

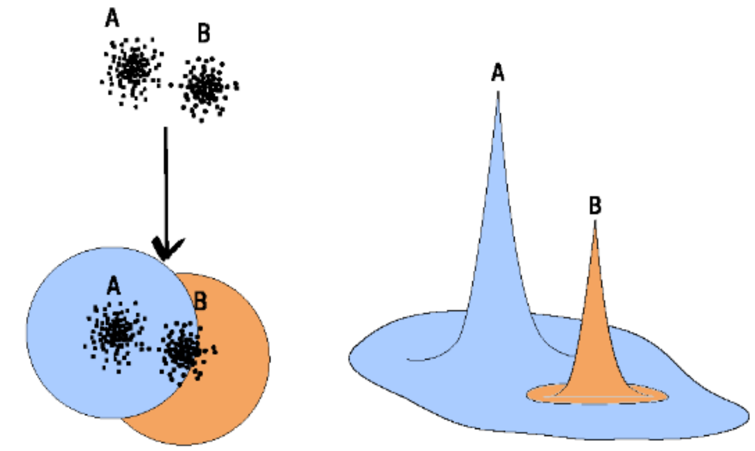
97%?


(a)
🟦 Intraspecies
🟩 Interspecies
🟥 Intragenus
16S rRNA gene sequence similarity (%)
100.0
99.5
99.0
98.5
98.0
97.5
97.0
96.5
98.65 %
97 %
95 %
96 %
65
70
75
80
85
90
95
100
ANI (%)

Swarm: A smart idea

swarm

large-scale clustering



clustering threshold (often 97%) is most of the time unadapted and can mask diversity.

swarm uses abundance values and a new clustering strategy to delineate natural high-quality OTUs.

agglomeration rather than division

Torbjørn Rognes

Oslo University

PeerJ

Swarm v2: highly-scalable and high-resolution amplicon clustering

Frédéric Mahé[1], Torbjørn Rognes[2,3], Christopher Quince[4], Colomban de Vargas[5,6] and Micah Dunthorn[1]

[1] Department of Ecology, Technische Universität Kaiserslautern, Kaiserslautern, Germany
[2] Department of Informatics, University of Oslo, Oslo, Norway
[3] Department of Microbiology, Oslo University Hospital, Rikshospitalet, Oslo, Norway
[4] Warwick Medical School, University of Warwick, Warwick, United Kingdom
[5] UMR 7144, EPEP–Évolution des Protistes et des Écosystèmes Pélagiques, Station Biologique de Roscoff, CNRS, Roscoff, France
[6] UMR7144 Station Biologique de Roscoff, Sorbonne Universités, UPMC Univ Paris 06, Roscoff, France

Submitted 4 August 2015
Accepted 31 October 2015
Published 10 December 2015

Corresponding authors
Frédéric Mahé,
mahe@rhrk.uni-kl.de
Torbjørn Rognes,
torognes@ifi.uio.no

Academic editor
Gilles van Wezel

Additional Information and Declarations can be found on page 10

DOI 10.7717/peerj.1420

ABSTRACT

Previously we presented Swarm v1, a novel and open source amplicon clustering program that produced fine-scale molecular operational taxonomic units (OTUs), free of arbitrary global clustering thresholds and input-order dependency. Swarm v1 worked with an initial phase that used iterative single-linkage with a local clustering threshold (*d*), followed by a phase that used the internal abundance structures of clusters to break chained OTUs. Here we present Swarm v2, which has two important novel features: (1) a new algorithm for $d = 1$ that allows the computation time of the program to scale linearly with increasing amounts of data; and (2) the new fastidious option that reduces under-grouping by grafting low abundant OTUs (e.g., singletons and doubletons) onto larger ones. Swarm v2 also directly integrates the clustering and breaking phases, dereplicates sequencing reads with $d = 0$, outputs OTU representatives in fasta format, and plots individual OTUs as two-dimensional networks.

Subjects Biodiversity, Bioinformatics, Environmental Sciences, Microbiology, Molecular Biology
Keywords Environmental diversity, Barcoding, Molecular operational taxonomic units

INTRODUCTION

Traditional *de novo* amplicon clustering methods that can handle large high-throughput sequencing datasets (e.g., *Edgar, 2010*; *Ghodsi, Liu & Pop, 2011*; *Fu et al., 2012*) suffer from two fundamental problems. First, they rely on an arbitrary fixed global clustering threshold to group amplicons into molecular operational taxonomic units (OTUs). Global clustering thresholds have rarely been justified and are not applicable to all taxa and marker lengths (e.g., *Caron et al., 2009*; *Nebel et al., 2011*; *Dunthorn et al., 2012*; *Brown et al., 2015*). Second, there is variability in the clustering results due to amplicon input order (*Koeppel & Wu, 2013*; *Mahé et al., 2014*).

To solve these problems, we previously introduced the open source Swarm v1 program that implemented an initial clustering phase written in C++, then a breaking phase written in Python (*Mahé et al., 2014*). Swarm's clustering phase (Fig. 1A) was novel in its approach to single linkage clustering in that, instead of using a global clustering (e.g., *Hartmann et al., 2012*; *Huse et al., 2010*), amplicons were iteratively added together using a

How to cite this article Mahé et al. (2015), Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ 3:e1420; DOI 10.7717/peerj.1420

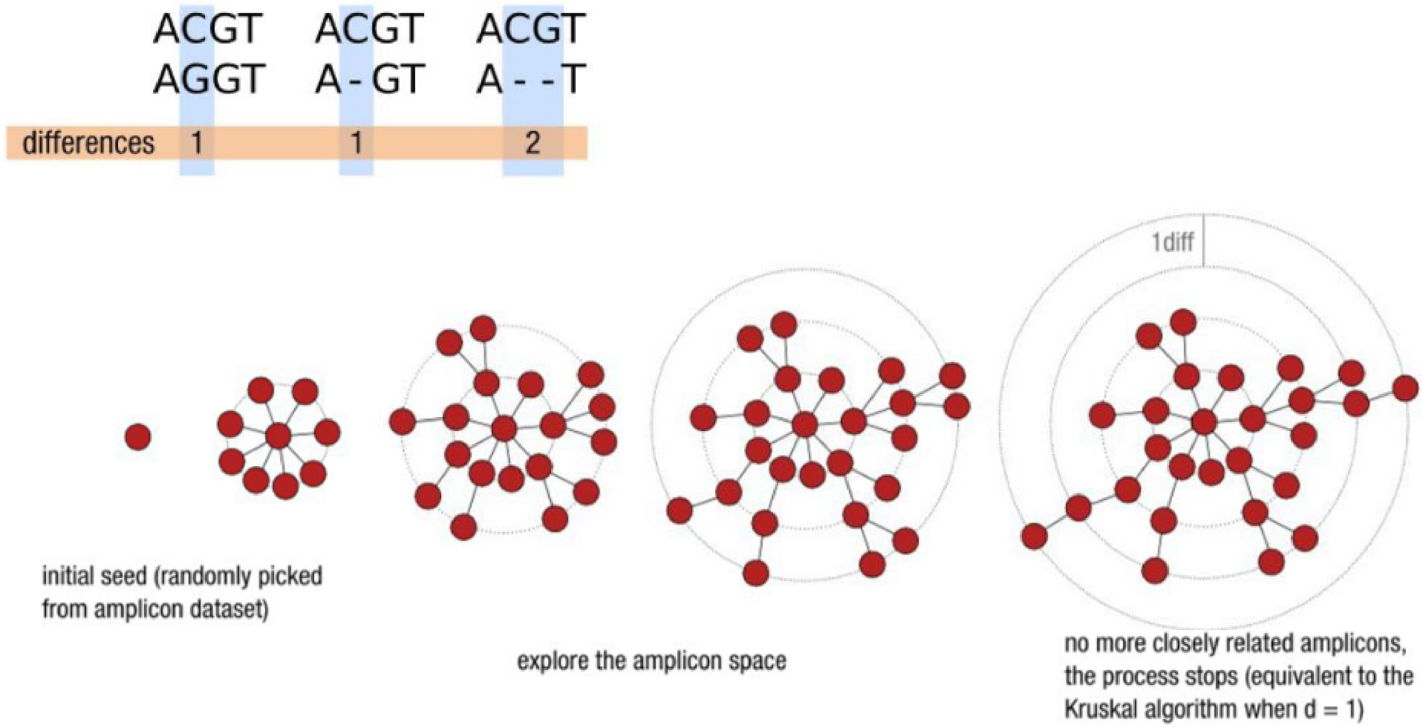
Mahé et al (2015) ; *Mahé et al (2014)*

Swarm

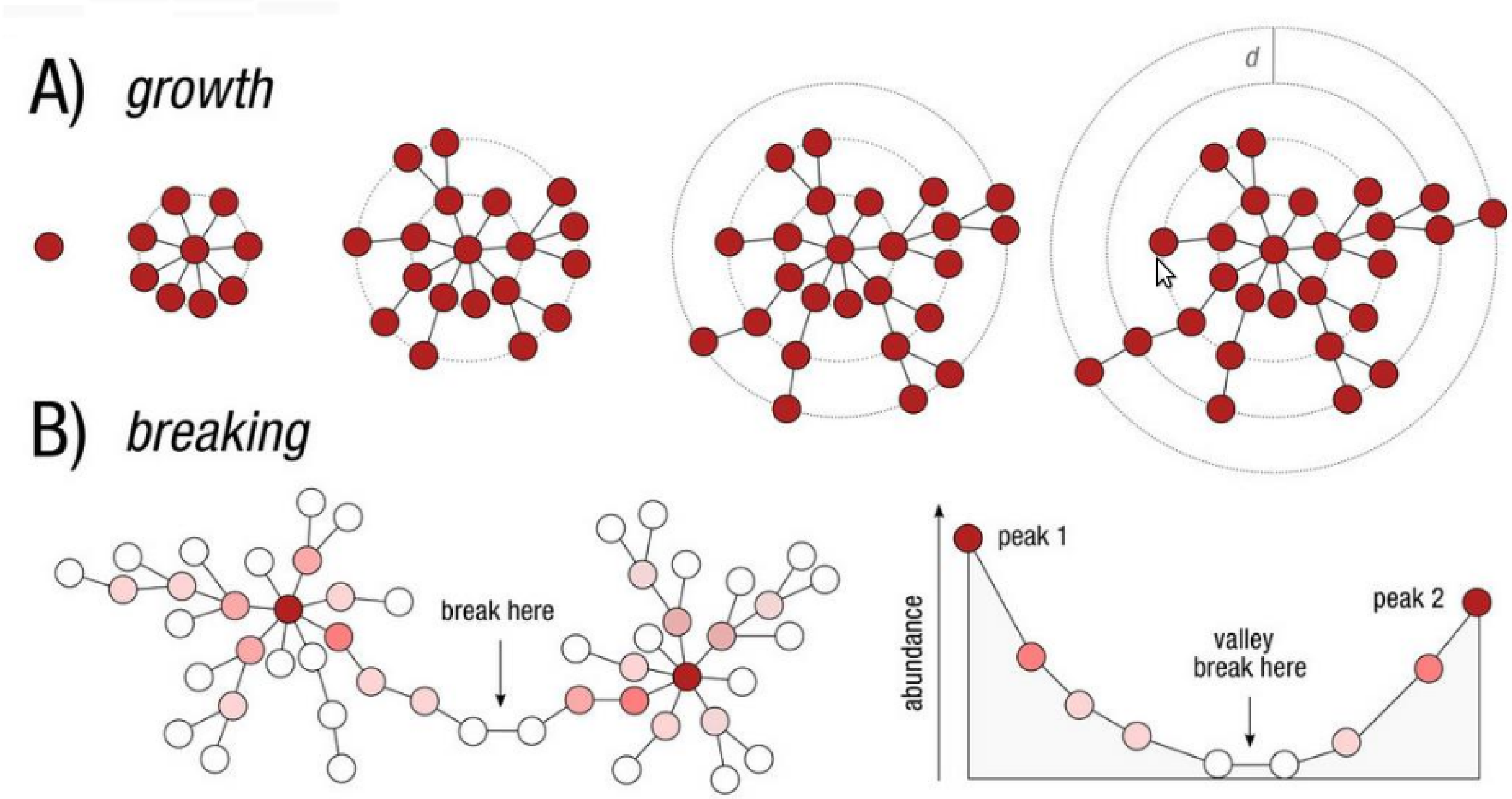
- A robust and fast clustering method for amplicon-based studies
- The purpose of swarmis to provide a novel clustering algorithm to handle large sets of amplicons
- swarm results are resilient to input-order changes and rely on a small local linking threshold d, the maximum number of differences between two amplicons
- swarm forms stable high-resolution clusters, with a high yield of biological information
- Default: forms a lot of low-abundant OTUs that are in fact artifacts and need to be removed

Mahé et al (2015) ; *Mahé et al (2014)*

d: the small local linking threshold


ACGT ACGT ACGT
AGGT A - GT A - - T
differences 1 1 2
1 diff
initial seed (randomly picked from amplicon dataset)
explore the amplicon space
no more closely related amplicons, the process stops (equivalent to the Kruskal algorithm when d = 1)

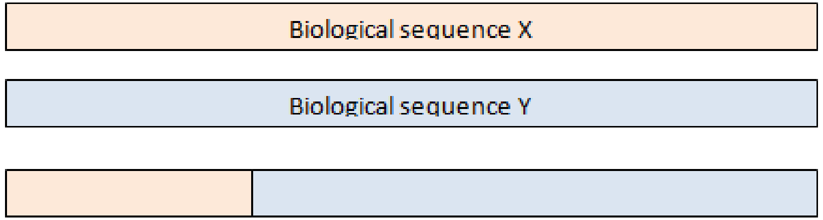
Swarm steps


A) growth
d
B) breaking
break here
peak 1
peak 2
abundance
valley
break here

Switch to TP: FROGS clustering

Chimera removal

Chimera



Biological sequence X

Biological sequence Y

Chimera formed from X and Y

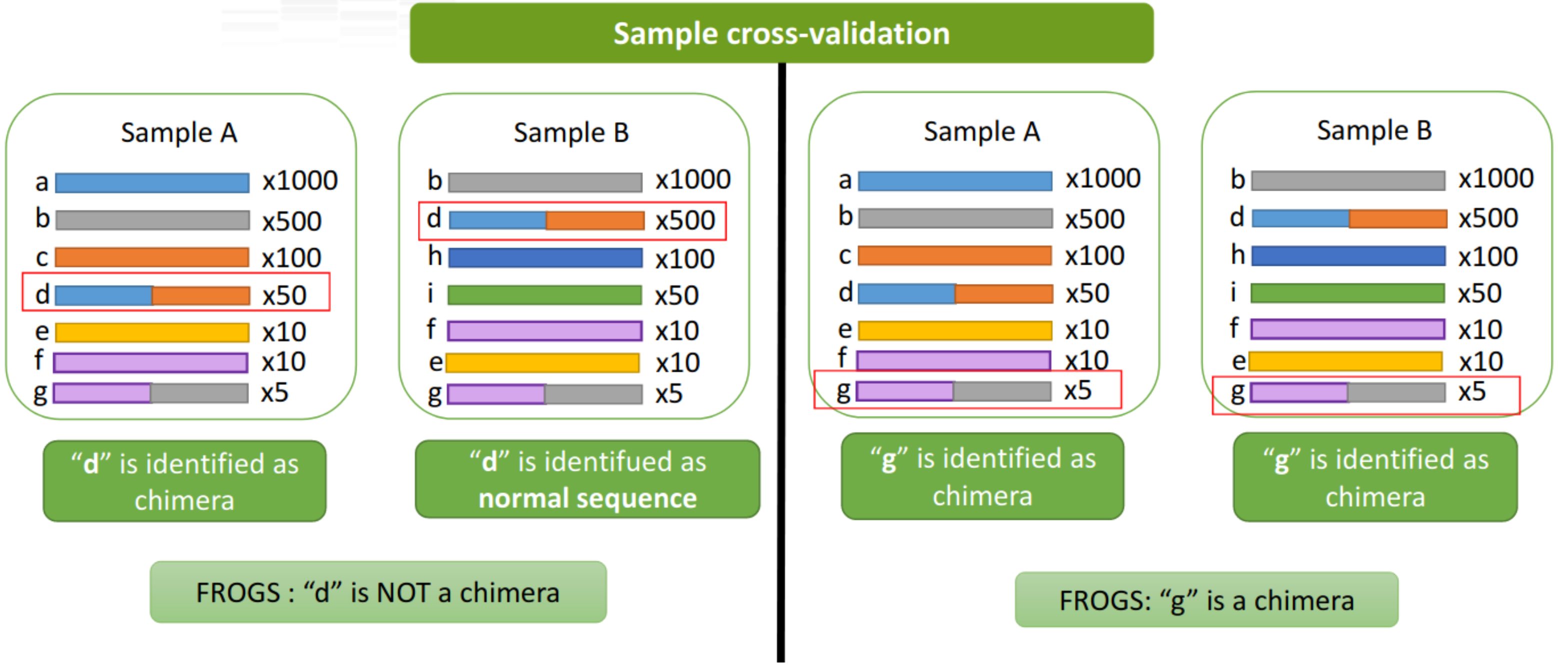
Chimera detection strategies

- Reference based: against a database of «genuine» sequences
- De novo: against abundant sequences in the samples
- FROGS uses vsearch as chimera removal tool

Rognes et al (2016)

Sample-cross validation

- FROGS adds a sample-cross validation



Sample cross-validation

“d” is identified as chimera

“d” is identified as **normal sequence**

FROGS : “d” is NOT a chimera

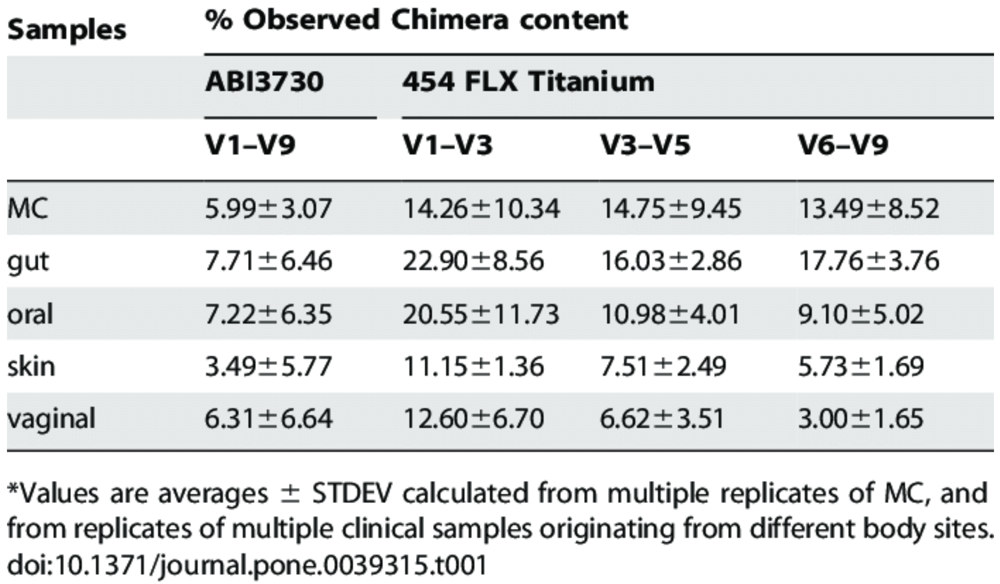
“g” is identified as chimera

“g” is identified as chimera

FROGS: “g” is a chimera

Chimera rates in samples

- From 5 to 40% in 16S data



Samples	% Observed Chimera content			
	ABI3730	454 FLX Titanium		
	V1–V9	V1–V3	V3–V5	V6–V9
MC	5.99±3.07	14.26±10.34	14.75±9.45	13.49±8.52
gut	7.71±6.46	22.90±8.56	16.03±2.86	17.76±3.76
oral	7.22±6.35	20.55±11.73	10.98±4.01	9.10±5.02
skin	3.49±5.77	11.15±1.36	7.51±2.49	5.73±1.69
vaginal	6.31±6.64	12.60±6.70	6.62±3.51	3.00±1.65

*Values are averages ± STDEV calculated from multiple replicates of MC, and from replicates of multiple clinical samples originating from different body sites.

doi:10.1371/journal.pone.0039315.t001

- Few with ITS (<10%)

Ward et al (2012)

Switch to TP: FROGS remove chimera

Abundance filters

Filters

- Scientific considerations :
    - Low abundant sequences are often chimeric
    - Impossible to distinguish rare biosphere and artefacts
    - Better accuracy after removing singletons
    - Smart to use replicates to keep good OTUs
    - Contaminations?

Switch to TP: FROGS filters

Taxonomic affiliation

Taxonomic affiliation

- Blast
- RDP-classifier
- IDTAXA
- QIIME
- SINTAX
- …

RDP classifier caveats

- Bootstrap confidence

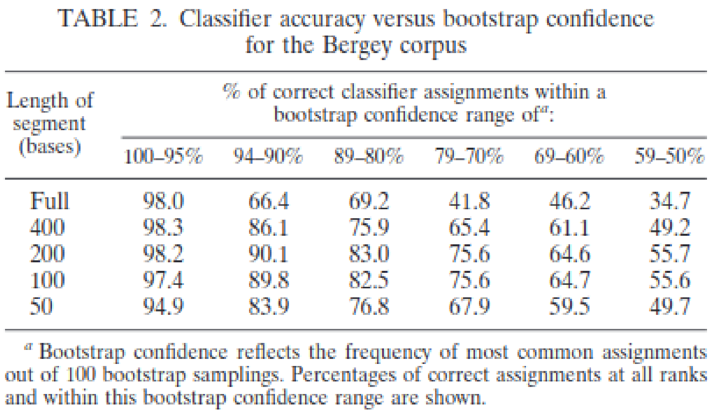


TABLE 2. Classifier accuracy versus bootstrap confidence for the Bergey corpus

Length of segment (bases)	% of correct classifier assignments within a bootstrap confidence range of[a]:					
	100–95%	94–90%	89–80%	79–70%	69–60%	59–50%
Full	98.0	66.4	69.2	41.8	46.2	34.7
400	98.3	86.1	75.9	65.4	61.1	49.2
200	98.2	90.1	83.0	75.6	64.6	55.7
100	97.4	89.8	82.5	75.6	64.7	55.6
50	94.9	83.9	76.8	67.9	59.5	49.7

[a] Bootstrap confidence reflects the frequency of most common assignments out of 100 bootstrap samplings. Percentages of correct assignments at all ranks and within this bootstrap confidence range are shown.

- Precision

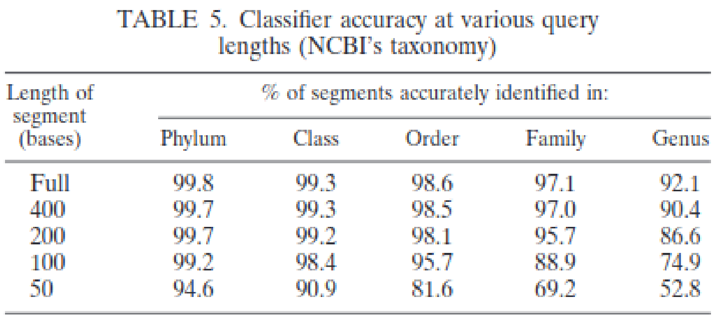


TABLE 5. Classifier accuracy at various query lengths (NCBI's taxonomy)

Length of segment (bases)	% of segments accurately identified in:				
	Phylum	Class	Order	Family	Genus
Full	99.8	99.3	98.6	97.1	92.1
400	99.7	99.3	98.5	97.0	90.4
200	99.7	99.2	98.1	95.7	86.6
100	99.2	98.4	95.7	88.9	74.9
50	94.6	90.9	81.6	69.2	52.8

Wang et al (2007)

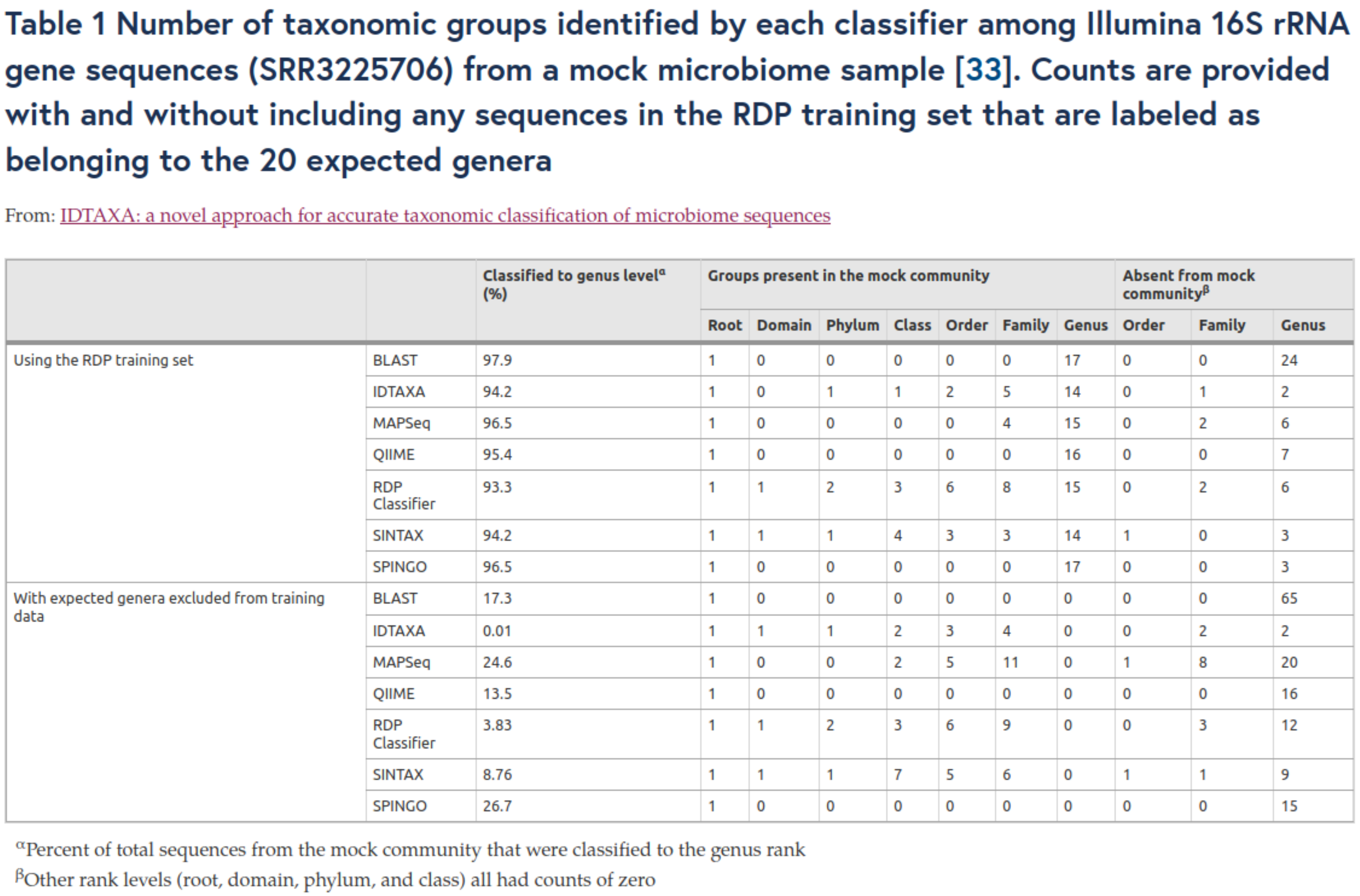


Table 1 Number of taxonomic groups identified by each classifier among Illumina 16S rRNA gene sequences (SRR3225706) from a mock microbiome sample [33]. Counts are provided with and without including any sequences in the RDP training set that are labeled as belonging to the 20 expected genera

From: IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences

		Classified to genus level[α] (%)	Groups present in the mock community							Absent from mock community[β]		
			Root	Domain	Phylum	Class	Order	Family	Genus	Order	Family	Genus
Using the RDP training set	BLAST	97.9	1	0	0	0	0	0	17	0	0	24
	IDTAXA	94.2	1	0	1	1	2	5	14	0	1	2
	MAPSeq	96.5	1	0	0	0	0	4	15	0	2	6
	QIIME	95.4	1	0	0	0	0	0	16	0	0	7
	RDP Classifier	93.3	1	1	2	3	6	8	15	0	2	6
	SINTAX	94.2	1	1	1	4	3	3	14	1	0	3
	SPINGO	96.5	1	0	0	0	0	0	17	0	0	3
With expected genera excluded from training data	BLAST	17.3	1	0	0	0	0	0	0	0	0	65
	IDTAXA	0.01	1	1	1	2	3	4	0	0	2	2
	MAPSeq	24.6	1	0	0	2	5	11	0	1	8	20
	QIIME	13.5	1	0	0	0	0	0	0	0	0	16
	RDP Classifier	3.83	1	1	2	3	6	9	0	0	3	12
	SINTAX	8.76	1	1	1	7	5	6	0	1	1	9
	SPINGO	26.7	1	0	0	0	0	0	0	0	0	15

[α] Percent of total sequences from the mock community that were classified to the genus rank

[β] Other rank levels (root, domain, phylum, and class) all had counts of zero

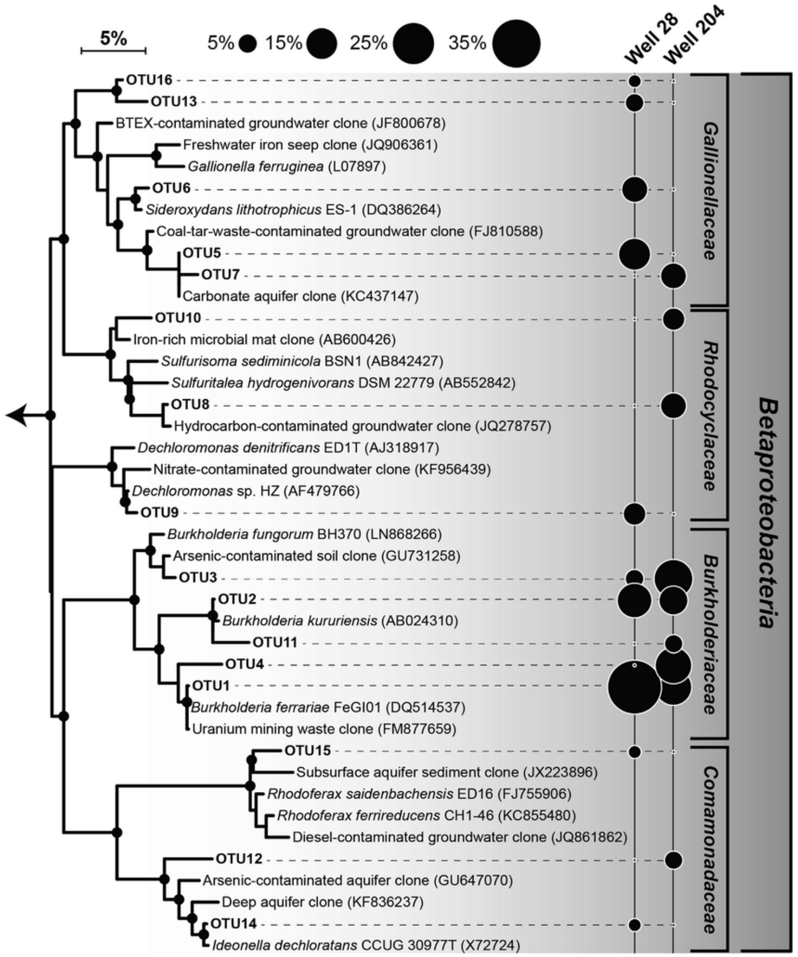
Advantages of FROGS affiliation

- Gives %cov & %id informations
- Gives all hits in a multi-affiliation file
    - Allows a smart correction sometimes (hits with same Species but different strains)

Switch to TP: FROGS affiliation OTU

Phylogeny

Phylogenetic similarity gives an other information to unknown OTUs


5%
5%
15%
25%
35%
Well 28
Well 204
OTU16
OTU13
BTEX-contaminated groundwater clone (JF800678)
Freshwater iron seep clone (JQ906361)
Gallionella ferruginea (L07897)
OTU6
Sideroxydans lithotrophicus ES-1 (DQ386264)
Coal-tar-waste-contaminated groundwater clone (FJ810588)
OTU5
OTU7
Carbonate aquifer clone (KC437147)
OTU10
Iron-rich microbial mat clone (AB600426)
Sulfurisoma sediminicola BSN1 (AB842427)
Sulfuritalea hydrogenivorans DSM 22779 (AB552842)
OTU8
Hydrocarbon-contaminated groundwater clone (JQ278757)
Dechloromonas denitrificans ED1T (AJ318917)
Nitrate-contaminated groundwater clone (KF956439)
Dechloromonas sp. HZ (AF479766)
OTU9
Burkholderia fungorum BH370 (LN868266)
Arsenic-contaminated soil clone (GU731258)
OTU3
OTU2
Burkholderia kururiensis (AB024310)
OTU11
OTU4
OTU1
Burkholderia ferrariae FeGI01 (DQ514537)
Uranium mining waste clone (FM877659)
OTU15
Subsurface aquifer sediment clone (JX223896)
Rhodoferax saidenbachensis ED16 (FJ755906)
Rhodoferax ferrireducens CH1-46 (KC855480)
Diesel-contaminated groundwater clone (JQ861862)
OTU12
Arsenic-contaminated aquifer clone (GU647070)
Deep aquifer clone (KF836237)
OTU14
Ideonella dechloratans CCUG 30977T (X72724)
Gallionellaceae
Rhodocyclaceae
Burkholderiaceae
Comamonadaceae
Betaproteobacteria

Chakraborty et al (2020)

Phylogenetic tree

- FROGS allows you to build a phylogenetic tree:
    - Mafft for doing multiple sequence alignments
    - Fasttree to build the rooted phylogentic tree
    - Essential to compute Unifrac distances
- Not always possible if sequence diversity is too high (e.g. ITS)

After this training... the real life

- Specific to FROGS
  - frogs-support@inrae.fr
- Migale support
  - help-migale@inrae.fr
- Want to collaborate with us?
  - https://migale.inrae.fr/ask-data-analysis