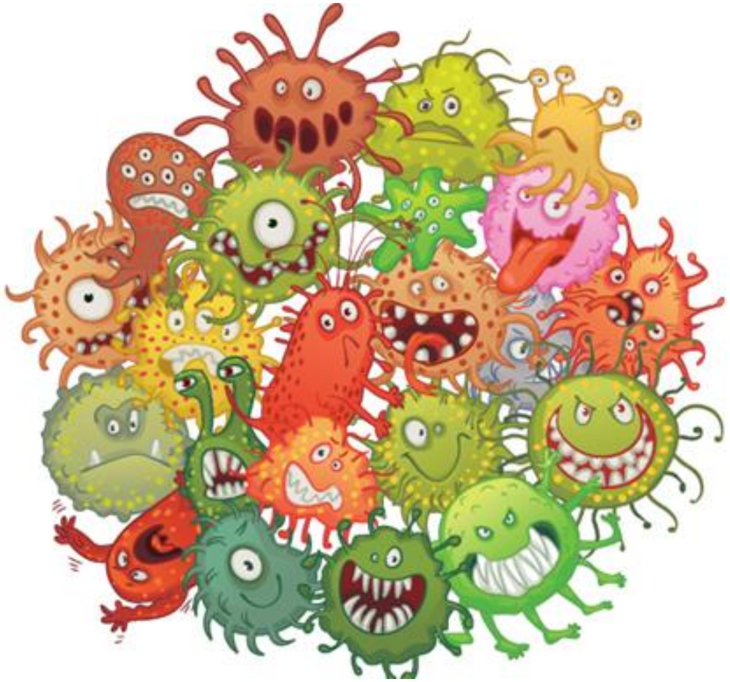


Intro to metabarcoding

Presented by Timothy Frey

Material generated by Soledad Benitez Ponce, Antonino Malacrino and Timothy Frey

Outline

- Why metabarcoding?
- What is metabarcoding?
- Metabarcoding pipeline

I. Sampling

II. DNA Extraction

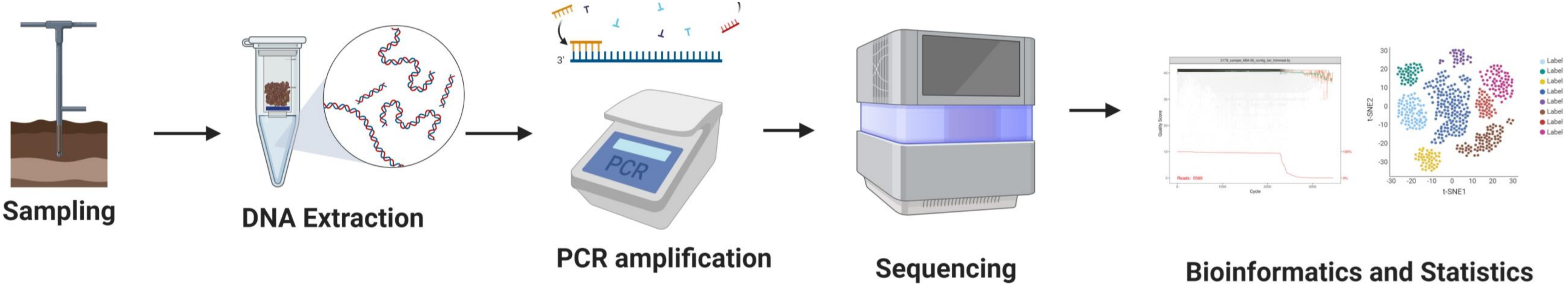
III. Target gene choices

IV. PCR amplification/Library Prep

V. Sequencing

Why Metabarcoding?

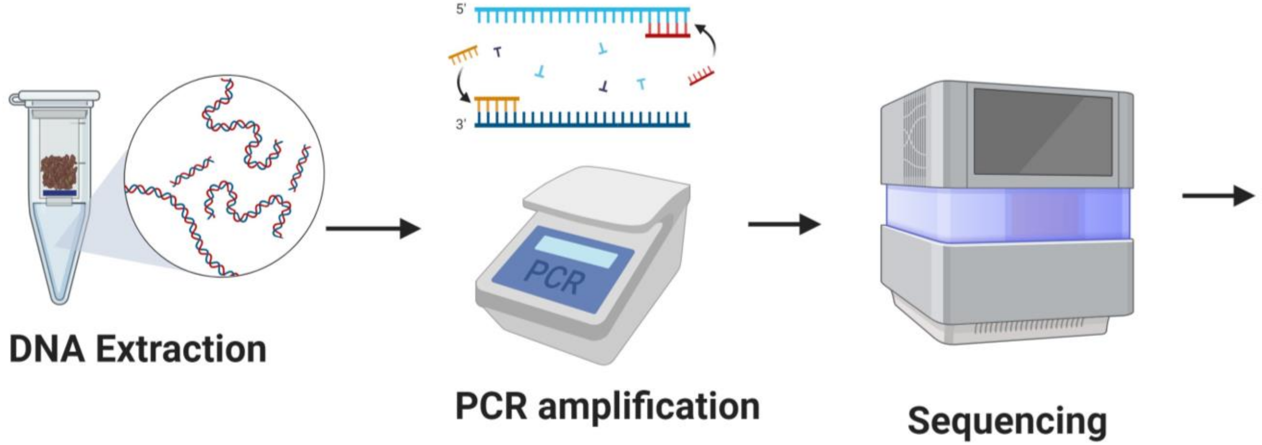
- Metabarcoding asks the question: who makes up a community?
- Metabarcoding allows us to characterize multiple species and individuals of a community simultaneously.
- A culture-independent technique



Sampling → DNA Extraction → PCR amplification → Sequencing → Bioinformatics and Statistics

Created with BioRender.com

Metabarcoding
Who are they? Which species occur in my sample?



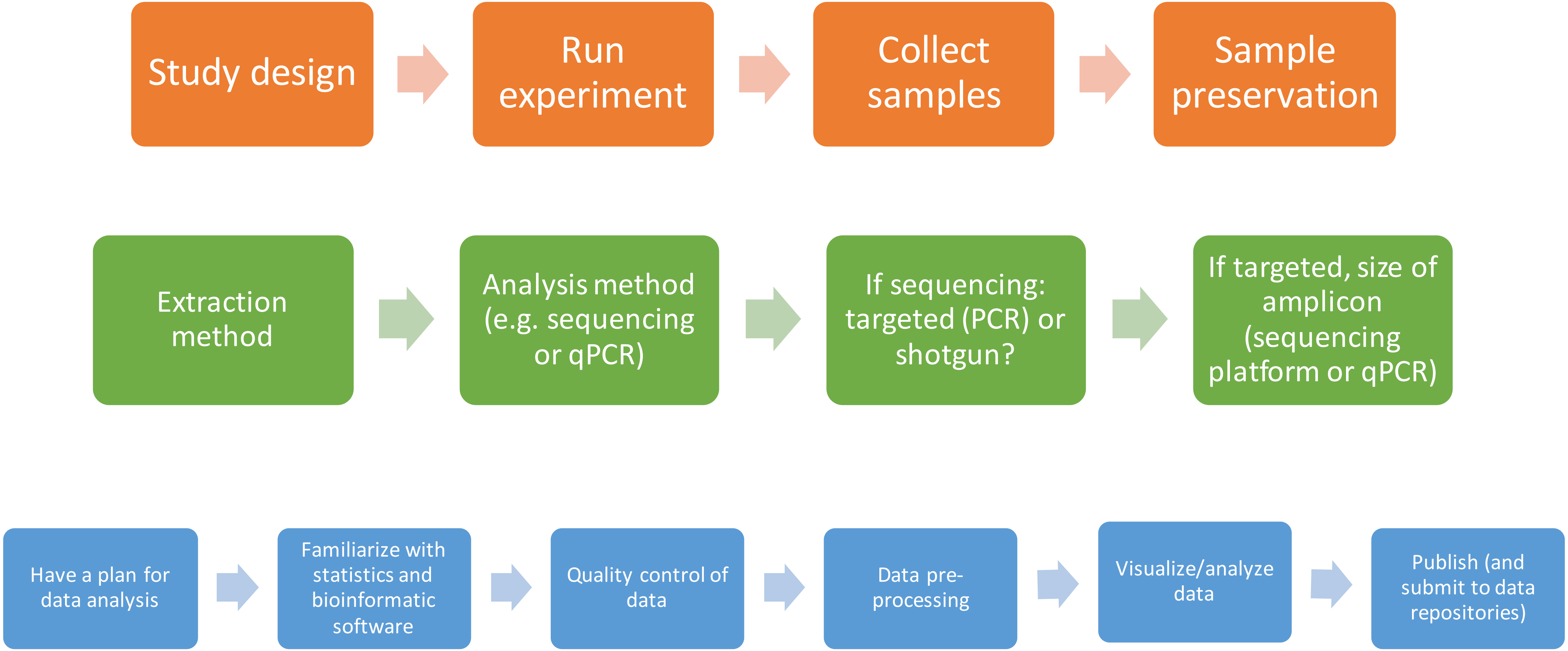
Taxa A	10	8	10
Taxa B	4	2	14
Taxa C	2	10	9
Taxa D	1	5	88

Abundance of a sequence (relative to the PCR product)

Created with BioRender.com

Modified from A. Malacrino

Decisions, decisions!



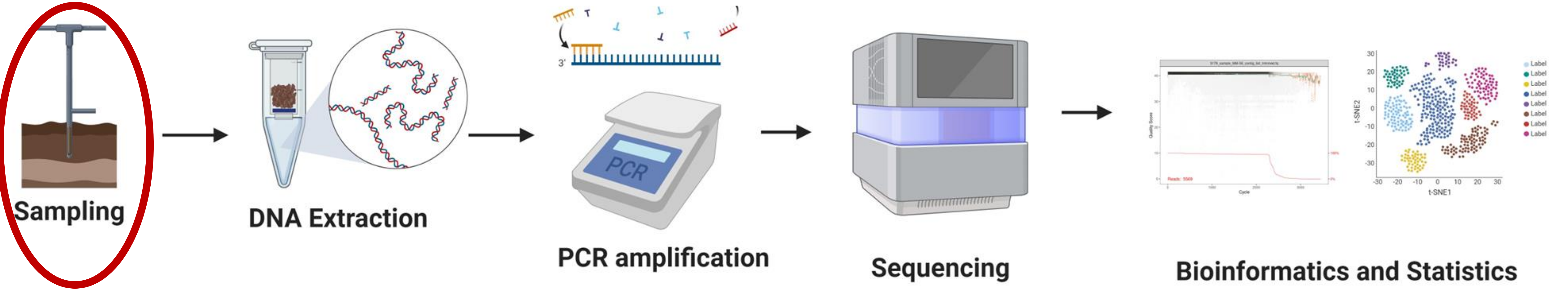
Study design → Run experiment → Collect samples → Sample preservation

Extraction method → Analysis method (e.g. sequencing or qPCR) → If sequencing: targeted (PCR) or shotgun? → If targeted, size of amplicon (sequencing platform or qPCR)

Have a plan for data analysis → Familiarize with statistics and bioinformatic software → Quality control of data → Data pre-processing → Visualize/analyze data → Publish (and submit to data repositories)

Goal: Minimize the sources of error and bias to obtain reproducible results, and maintain statistical power

Mod. from A. Testen



Sampling

DNA Extraction

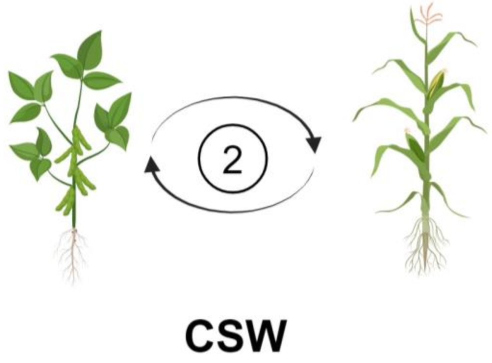
PCR amplification

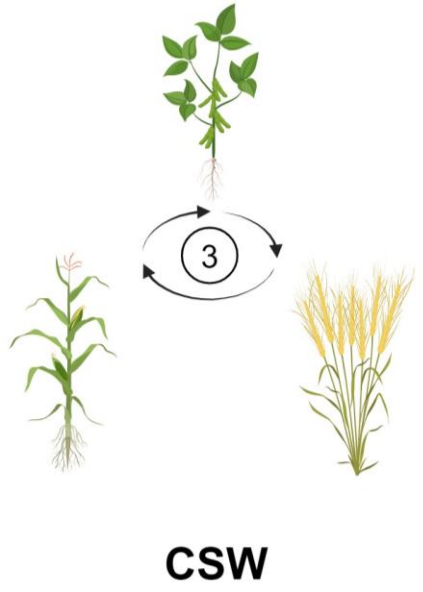
Sequencing

Bioinformatics and Statistics

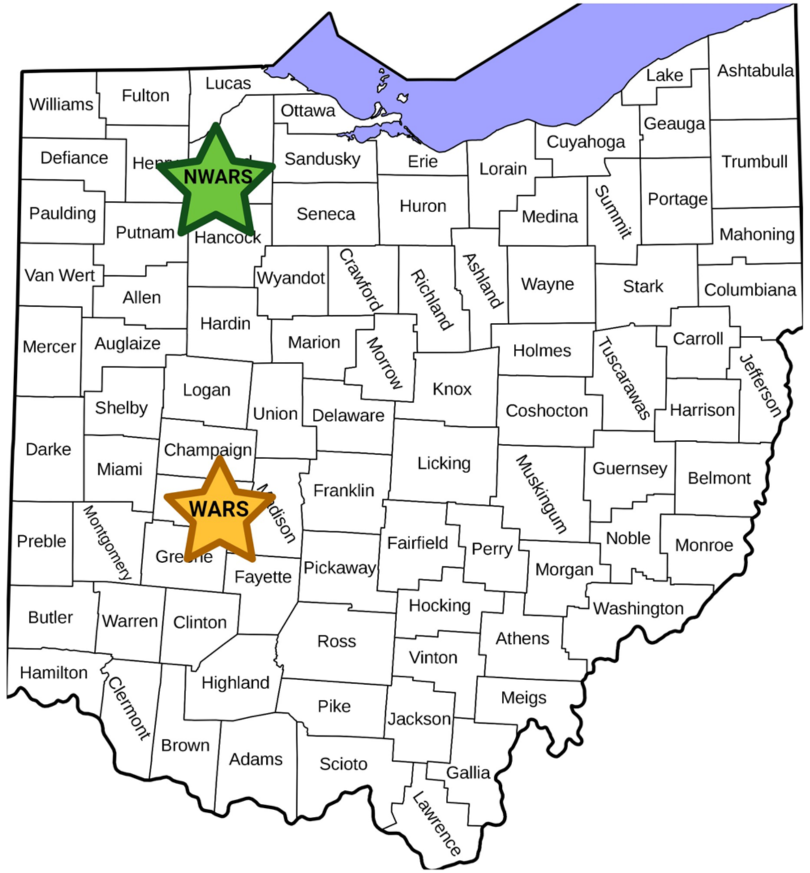
Sampling – A brief description of our dataset

- Two Rotations – Corn/Soy vs Corn/Soy/Wheat



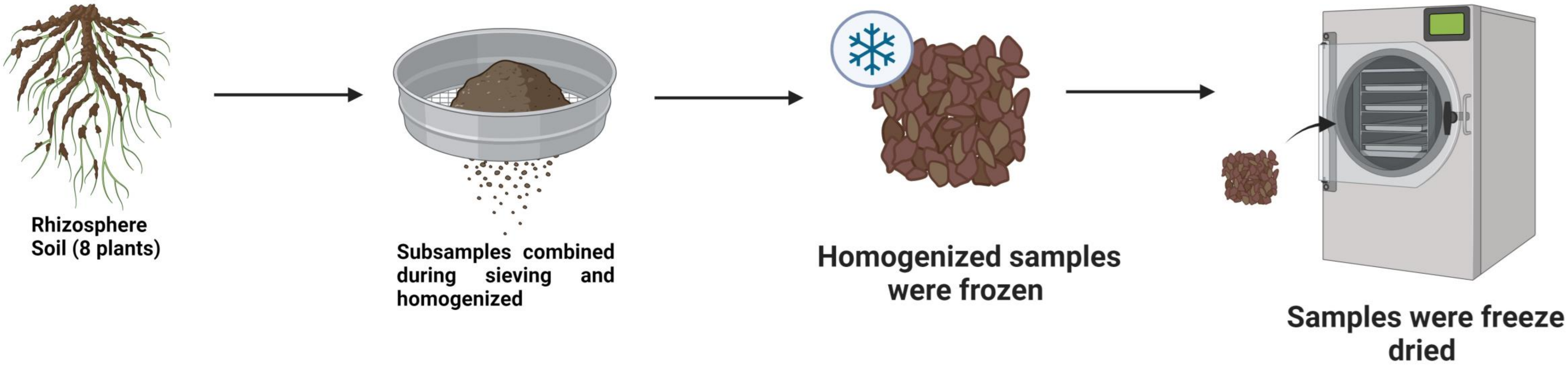


- Two locations - NWARS



Created with BioRender.com

Sampling and processing (prior to DNA extraction)



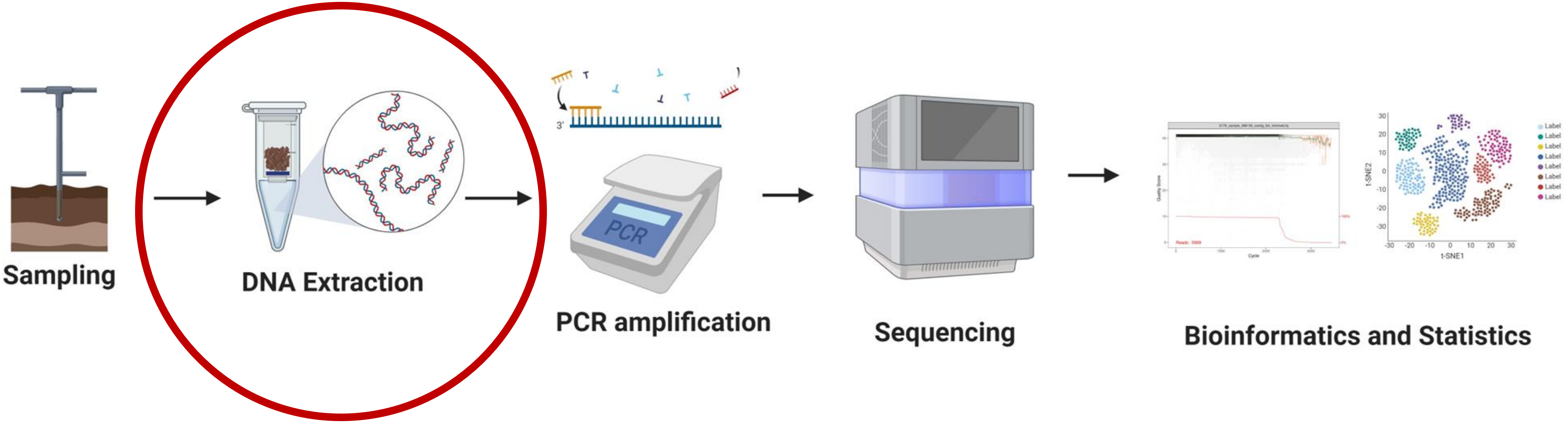
Rhizosphere Soil (8 plants)

Subsamples combined during sieving and homogenized

Homogenized samples were frozen

Samples were freeze dried

Created with BioRender.com



Sampling

DNA Extraction

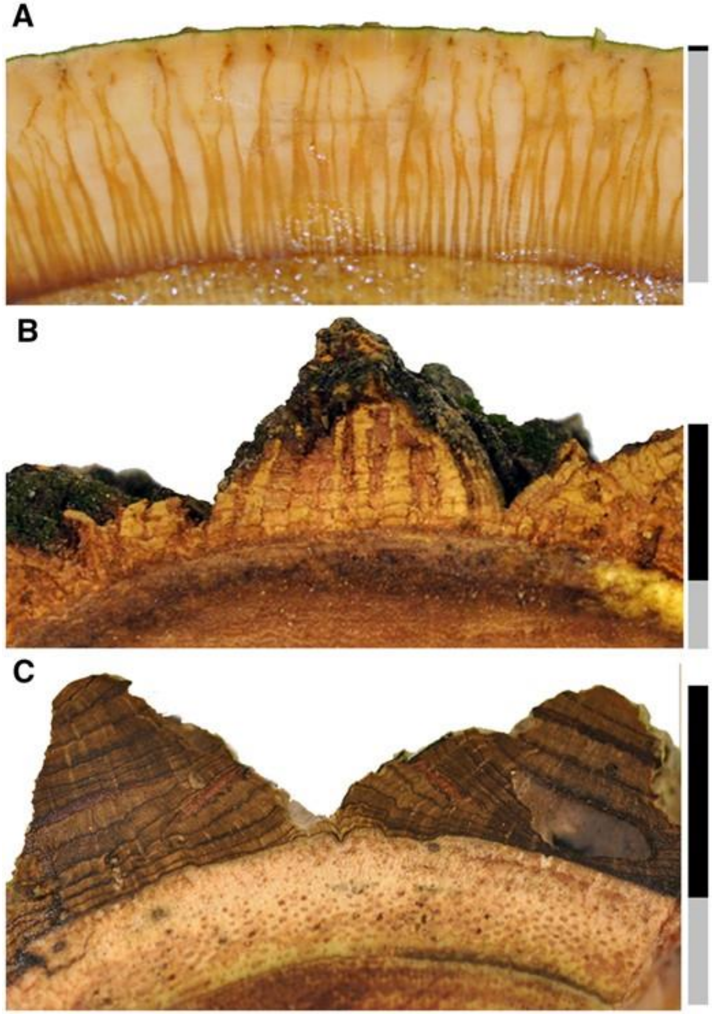
PCR amplification

Sequencing

Bioinformatics and Statistics

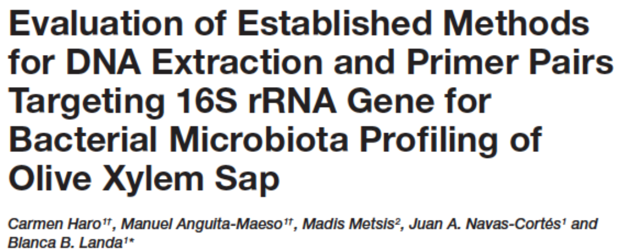
DNA extraction - efficiency depends on your sample origin and extraction method

smithsonian.org



Rosell et al 2019

kew.org



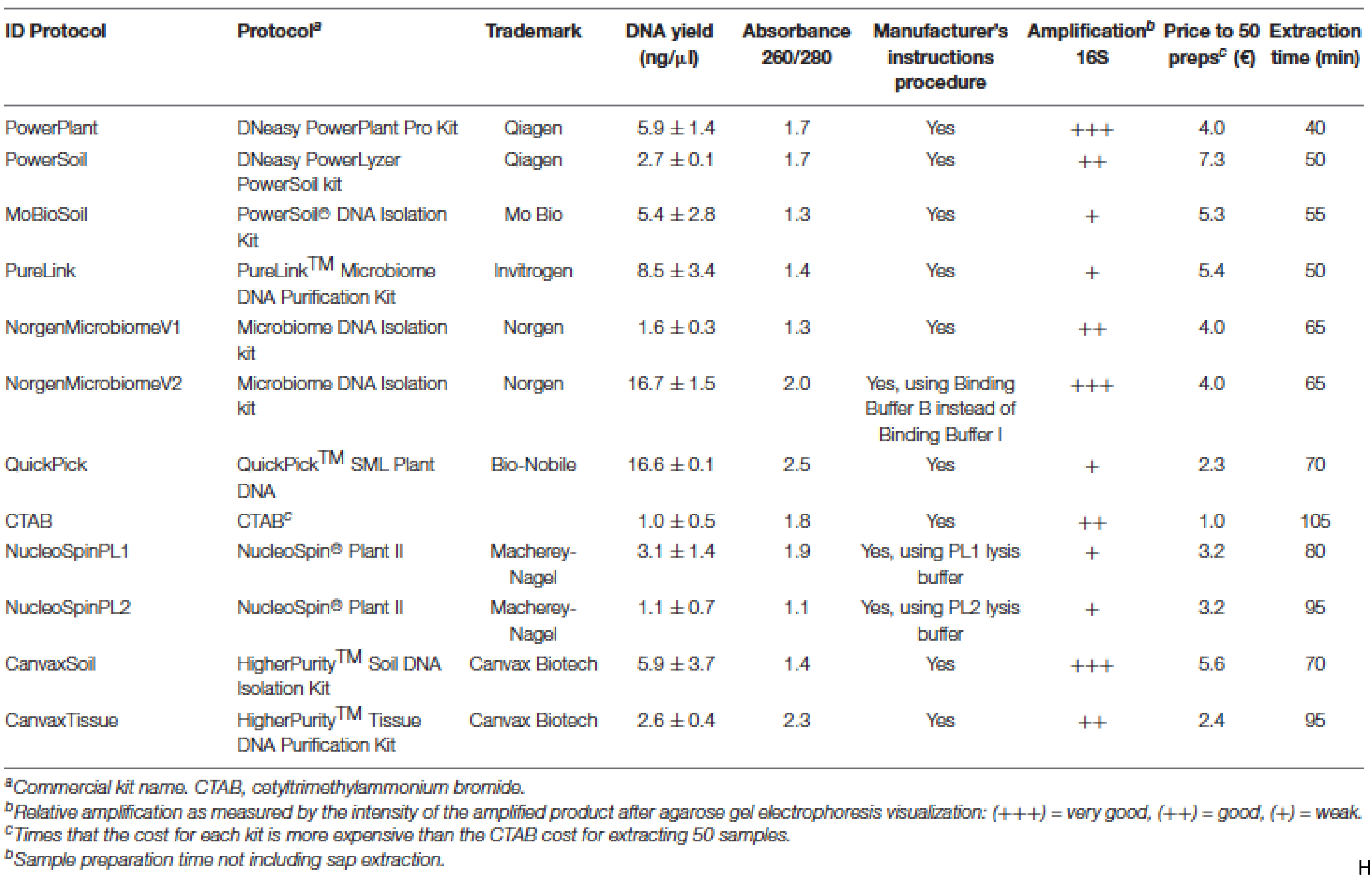
Evaluation of Established Methods for DNA Extraction and Primer Pairs Targeting 16S rRNA Gene for Bacterial Microbiota Profiling of Olive Xylem Sap

Carmen Haro[1†], Manuel Anguita-Maeso[1†], Madis Metsis[2], Juan A. Navas-Cortés[1] and Blanca B. Landa[1]*

DNA extraction kits

DNA extraction efficiency depends on your sample origin and extraction method

.E 1 | Characteristics of the DNA extraction protocols used in the study.



ID Protocol	Protocol[a]	Trademark	DNA yield (ng/µl)	Absorbance 260/280	Manufacturer's instructions procedure	Amplification[b] 16S	Price to 50 preps[c] (€)	Extraction time (min)
PowerPlant	DNeasy PowerPlant Pro Kit	Qiagen	5.9 ± 1.4	1.7	Yes	+++	4.0	40
PowerSoil	DNeasy PowerLyzer PowerSoil kit	Qiagen	2.7 ± 0.1	1.7	Yes	++	7.3	50
MoBioSoil	PowerSoil® DNA Isolation Kit	Mo Bio	5.4 ± 2.8	1.3	Yes	+	5.3	55
PureLink	PureLink™ Microbiome DNA Purification Kit	Invitrogen	8.5 ± 3.4	1.4	Yes	+	5.4	50
NorgenMicrobiomeV1	Microbiome DNA Isolation kit	Norgen	1.6 ± 0.3	1.3	Yes	++	4.0	65
NorgenMicrobiomeV2	Microbiome DNA Isolation kit	Norgen	16.7 ± 1.5	2.0	Yes, using Binding Buffer B instead of Binding Buffer I	+++	4.0	65
QuickPick	QuickPick™ SML Plant DNA	Bio-Nobile	16.6 ± 0.1	2.5	Yes	+	2.3	70
CTAB	CTAB[c]		1.0 ± 0.5	1.8	Yes	++	1.0	105
NucleoSpinPL1	NucleoSpin® Plant II	Macherey-Nagel	3.1 ± 1.4	1.9	Yes, using PL1 lysis buffer	+	3.2	80
NucleoSpinPL2	NucleoSpin® Plant II	Macherey-Nagel	1.1 ± 0.7	1.1	Yes, using PL2 lysis buffer	+	3.2	95
CanvaxSoil	HigherPurity™ Soil DNA Isolation Kit	Canvax Biotech	5.9 ± 3.7	1.4	Yes	+++	5.6	70
CanvaxTissue	HigherPurity™ Tissue DNA Purification Kit	Canvax Biotech	2.6 ± 0.4	2.3	Yes	++	2.4	95

[a] *Commercial kit name. CTAB, cetyltrimethylammonium bromide.*

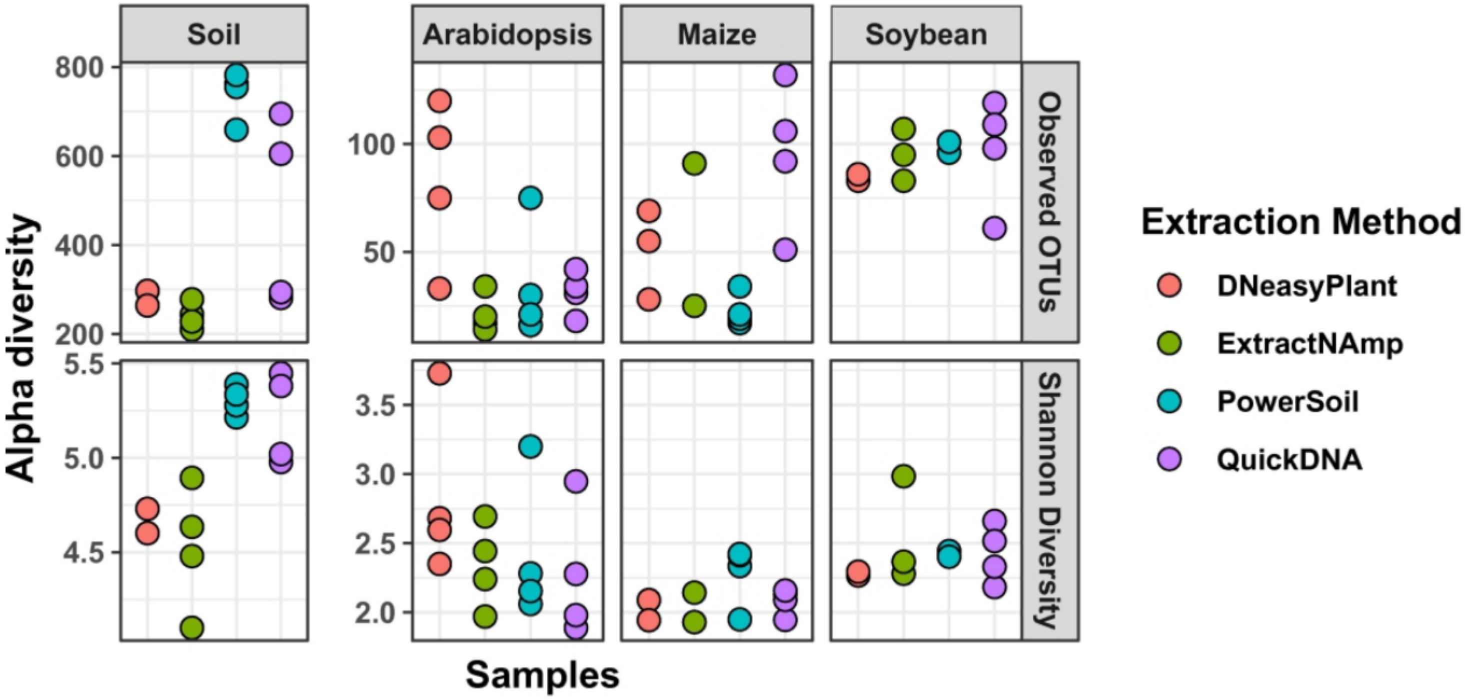
[b] *Relative amplification as measured by the intensity of the amplified product after agarose gel electrophoresis visualization: (+++) = very good, (++) = good, (+) = weak.*

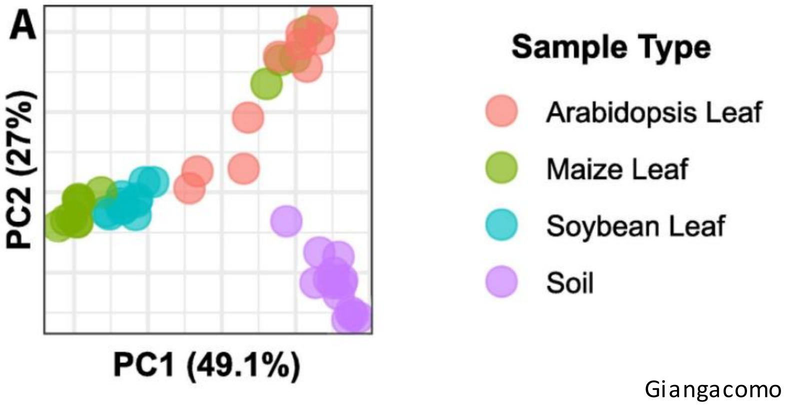
[c] *Times that the cost for each kit is more expensive than the CTAB cost for extracting 50 samples.*

[D] *Sample preparation time not including sap extraction.*

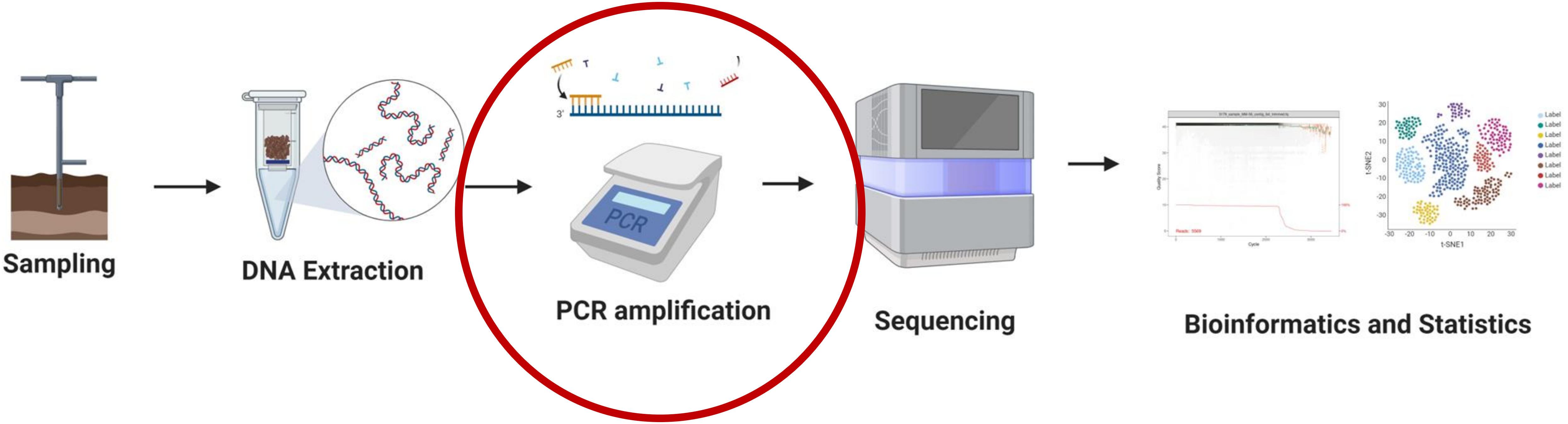
Haro et al 2021

DNA Extraction can bias metabarcoding experiments





Giangacomo et al, 2021



Sampling → DNA Extraction → PCR amplification → Sequencing → Bioinformatics and Statistics

Which analysis to use?

(nucleic-acid based)

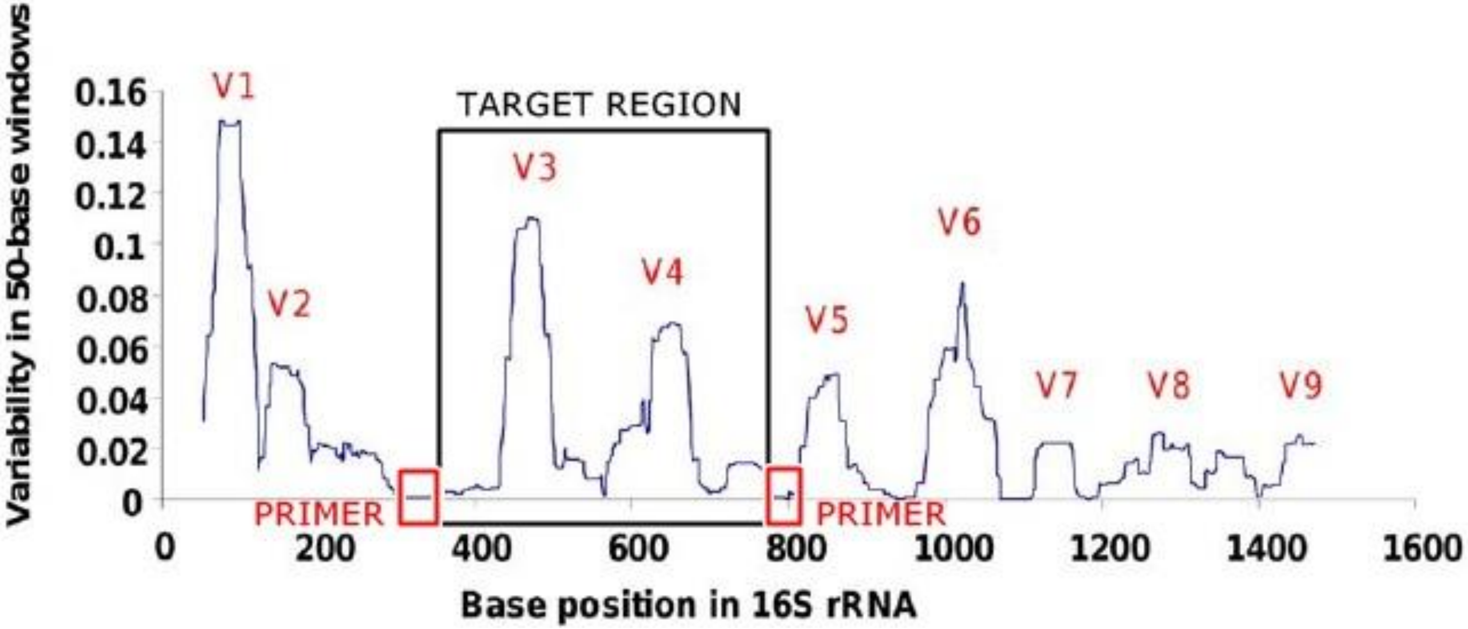
- Shotgun vs. **Targeted**
- If targeted: which is your marker gene?
- What level of phylogenetic resolution do you need?
- Which sequencing methodology will be a good fit for your research question? (Read length/depth of sampling)
- Do you need quantitative data?
- What are potential sources of bias and controls to be used?

What is the target gene?

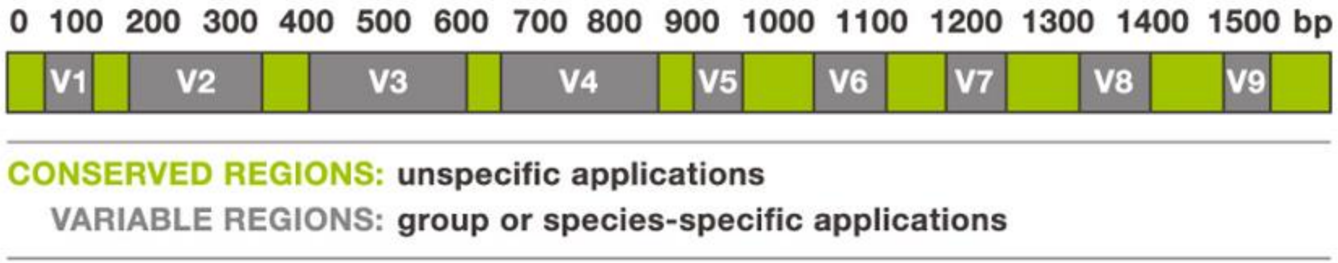
- Gene of choice
  - Taxonomic survey (e.g. gene diversity: rDNA, Btub, rpoB)
  - Metabolic diversity (functional genes, e.g. *nifh*, laccasse)
- Universal or taxa specific? Who is our target?
- Resolution of a short-read?
- What databases are available, or how would you construct your own?
- Is copy number an issue?
- What are potential sources of bias?

Target gene considerations

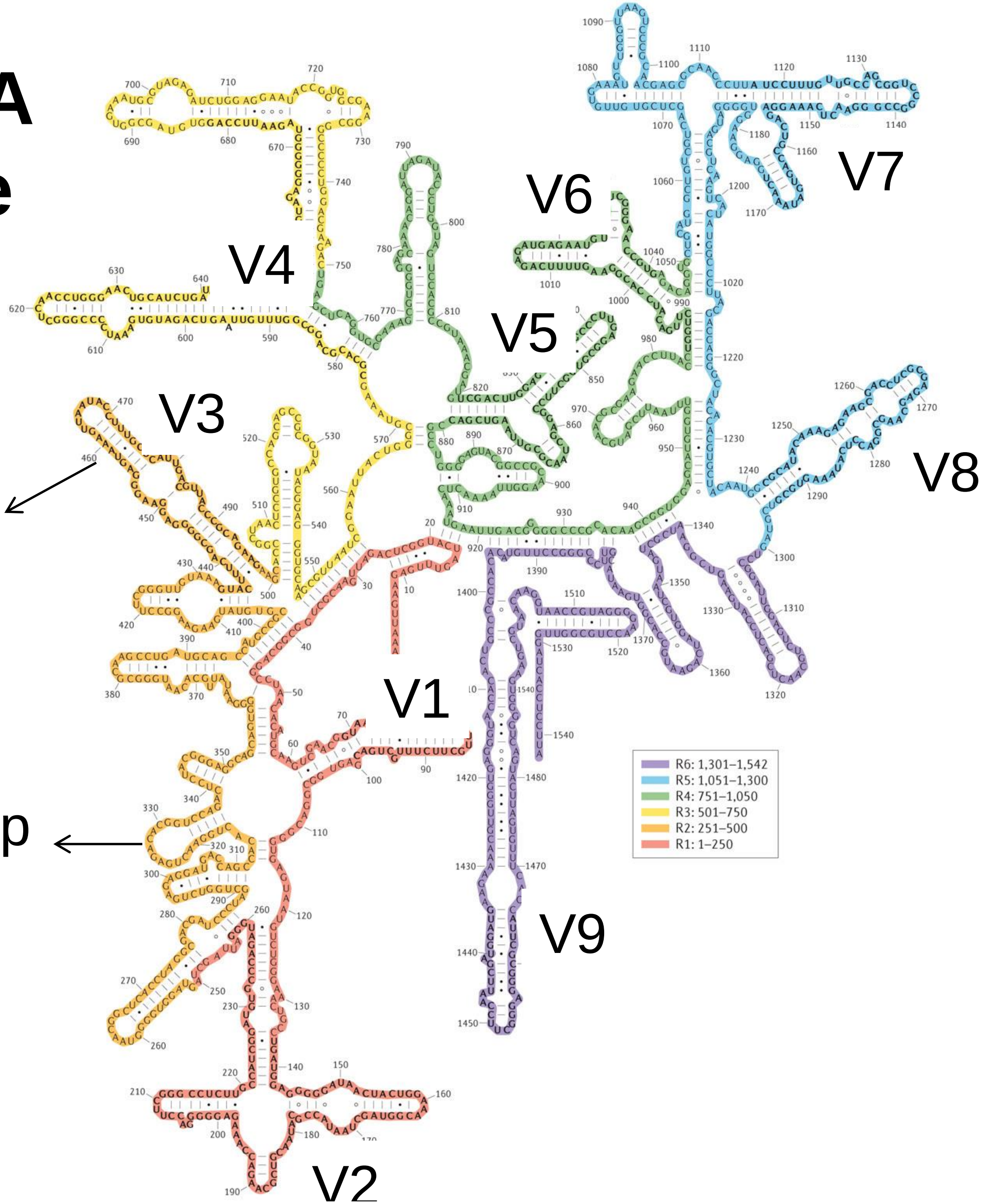
- Universal primers – If you want to target as many species as possible in a metabarcoding experiment
- Use of rRNA region is advantageous because it has alternating conserved and variable regions



Bodilis, Josselin, et al., 2012


0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500 bp
V1
V2
V3
V4
V5
V6
V7
V8
V9
CONSERVED REGIONS: unspecific applications
VARIABLE REGIONS: group or species-specific applications

16S rRNA structure



V7

V6

V4

V5

V3

stem

V8

V1

loop

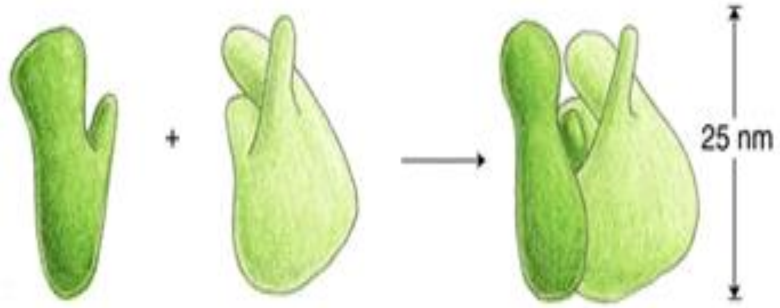
V9

V2

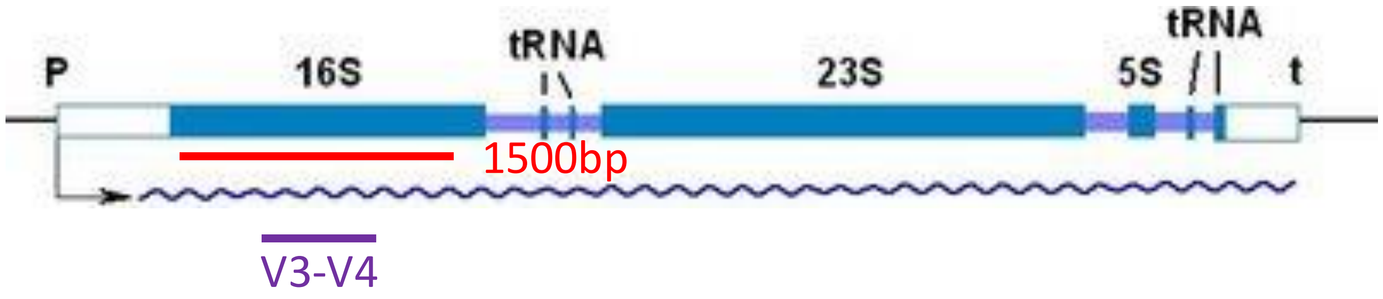
Nature Reviews Microbiology **12**, 635–645 (2014) | doi:10.1038/nrmicro3330

Ribosomal markers as taxonomic barcodes

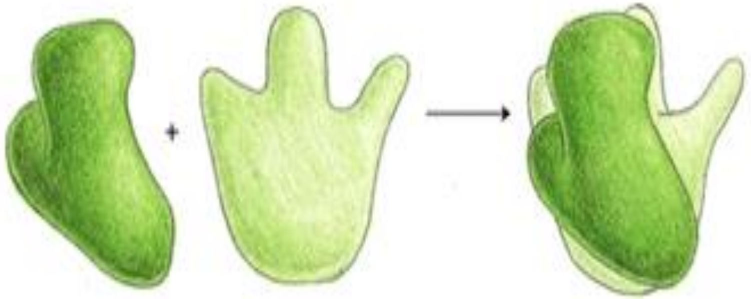
30S 50S



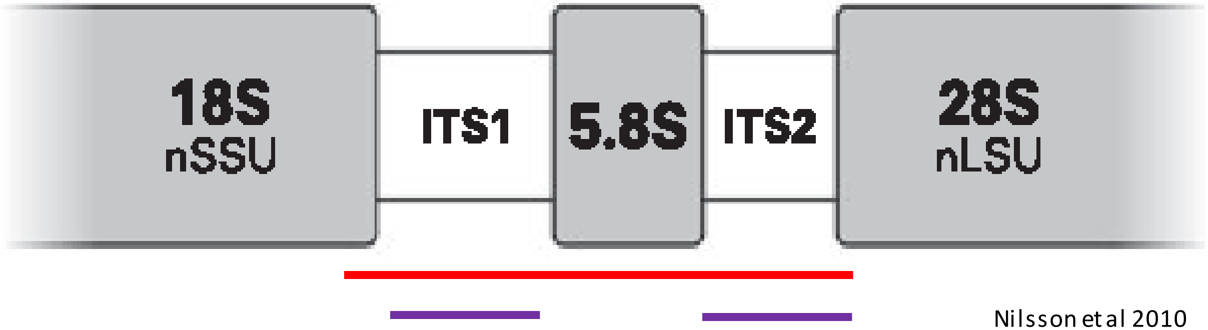
Bacteria (chloroplast, mitochondria)



40S 60S



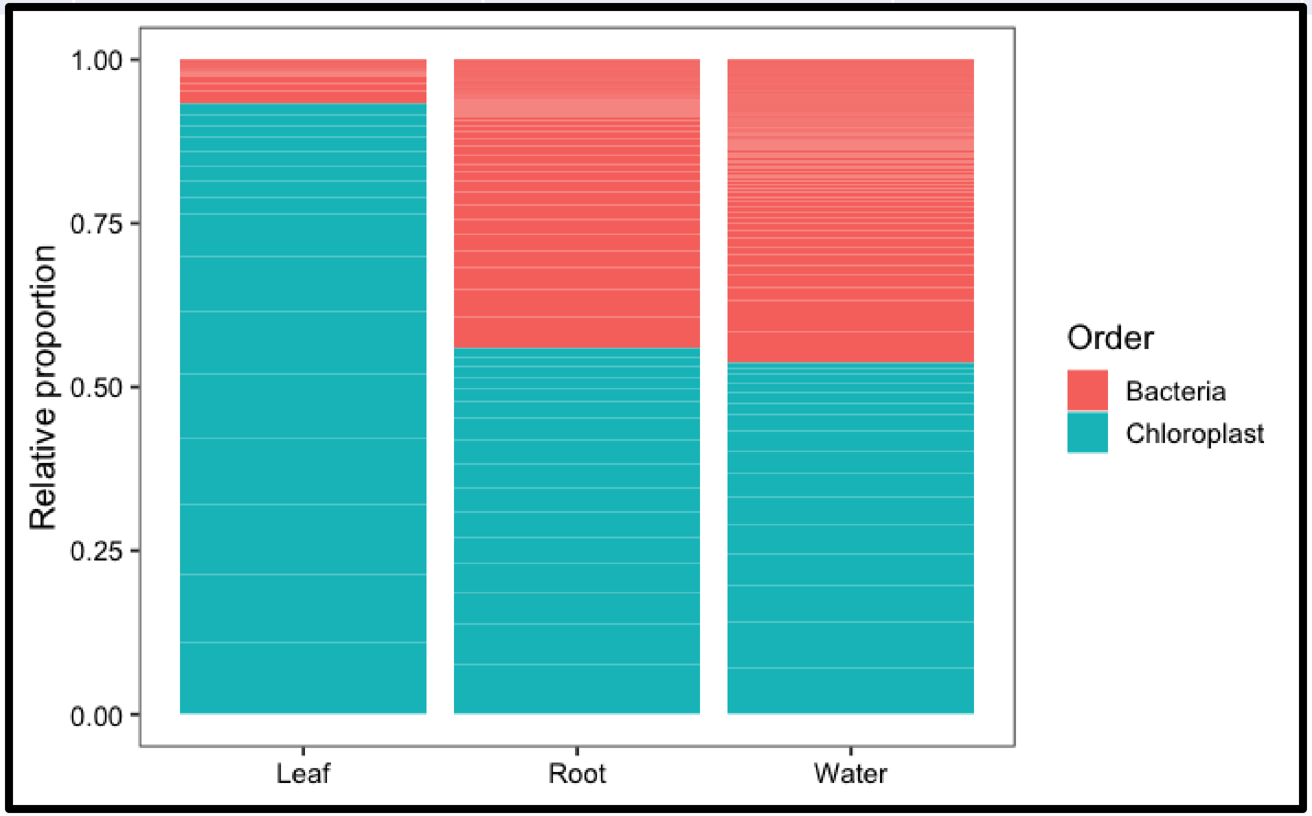
Eukarya (Fungi, Nematodes, Protists, Plants)



Nilsson et al 2010

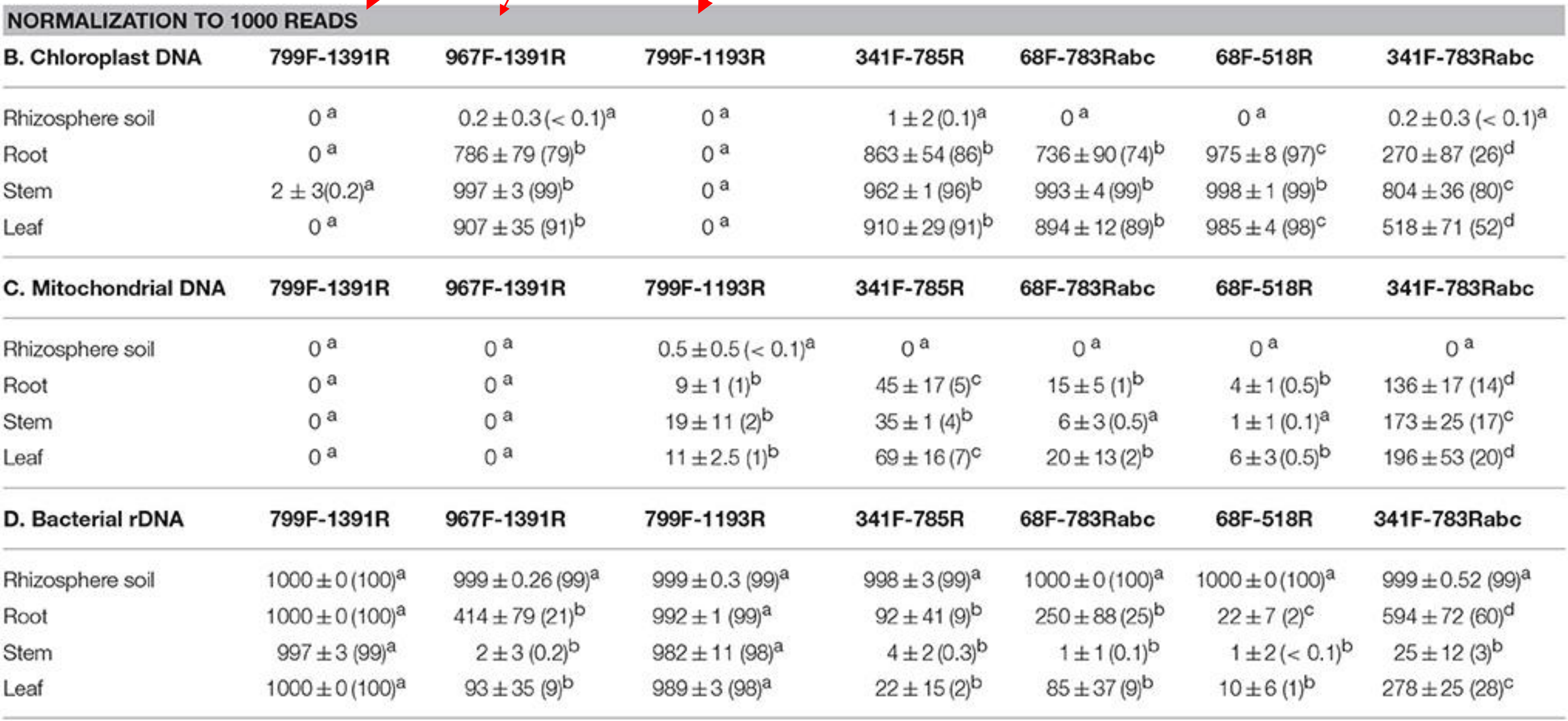
Most commonly used bacterial primer set in soil (and plant studies): 515R-806R

Primer name	% coverage			
	Bacteria	Archaeae	Plant mitochondria	Plant plastids
515F	79.1	50.6	55.4	98.4



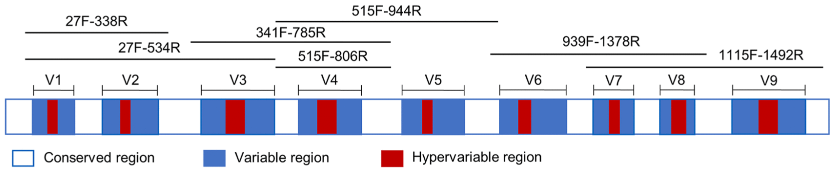
Reinhold-Hurek et al 2015

Some alternatives



NORMALIZATION TO 1000 READS

B. Chloroplast DNA	799F-1391R	967F-1391R	799F-1193R	341F-785R	68F-783Rabc	68F-518R	341F-783Rabc
Rhizosphere soil	0^{a}	0.2 ± 0.3 (< 0.1)^{a}	0^{a}	1 ± 2 (0.1)^{a}	0^{a}	0^{a}	0.2 ± 0.3 (< 0.1)^{a}
Root	0^{a}	786 ± 79 (79)^{b}	0^{a}	863 ± 54 (86)^{b}	736 ± 90 (74)^{b}	975 ± 8 (97)^{c}	270 ± 87 (26)^{d}
Stem	2 ± 3(0.2)^{a}	997 ± 3 (99)^{b}	0^{a}	962 ± 1 (96)^{b}	993 ± 4 (99)^{b}	998 ± 1 (99)^{b}	804 ± 36 (80)^{c}
Leaf	0^{a}	907 ± 35 (91)^{b}	0^{a}	910 ± 29 (91)^{b}	894 ± 12 (89)^{b}	985 ± 4 (98)^{c}	518 ± 71 (52)^{d}
C. Mitochondrial DNA	**799F-1391R**	**967F-1391R**	**799F-1193R**	**341F-785R**	**68F-783Rabc**	**68F-518R**	**341F-783Rabc**
Rhizosphere soil	0^{a}	0^{a}	0.5 ± 0.5 (< 0.1)^{a}	0^{a}	0^{a}	0^{a}	0^{a}
Root	0^{a}	0^{a}	9 ± 1 (1)^{b}	45 ± 17 (5)^{c}	15 ± 5 (1)^{b}	4 ± 1 (0.5)^{b}	136 ± 17 (14)^{d}
Stem	0^{a}	0^{a}	19 ± 11 (2)^{b}	35 ± 1 (4)^{b}	6 ± 3 (0.5)^{a}	1 ± 1 (0.1)^{a}	173 ± 25 (17)^{c}
Leaf	0^{a}	0^{a}	11 ± 2.5 (1)^{b}	69 ± 16 (7)^{c}	20 ± 13 (2)^{b}	6 ± 3 (0.5)^{b}	196 ± 53 (20)^{d}
D. Bacterial rDNA	**799F-1391R**	**967F-1391R**	**799F-1193R**	**341F-785R**	**68F-783Rabc**	**68F-518R**	**341F-783Rabc**
Rhizosphere soil	1000 ± 0 (100)^{a}	999 ± 0.26 (99)^{a}	999 ± 0.3 (99)^{a}	998 ± 3 (99)^{a}	1000 ± 0 (100)^{a}	1000 ± 0 (100)^{a}	999 ± 0.52 (99)^{a}
Root	1000 ± 0 (100)^{a}	414 ± 79 (21)^{b}	992 ± 1 (99)^{a}	92 ± 41 (9)^{b}	250 ± 88 (25)^{b}	22 ± 7 (2)^{c}	594 ± 72 (60)^{d}
Stem	997 ± 3 (99)^{a}	2 ± 3 (0.2)^{b}	982 ± 11 (98)^{a}	4 ± 2 (0.3)^{b}	1 ± 1 (0.1)^{b}	1 ± 2 (< 0.1)^{b}	25 ± 12 (3)^{b}
Leaf	1000 ± 0 (100)^{a}	93 ± 35 (9)^{b}	989 ± 3 (98)^{a}	22 ± 15 (2)^{b}	85 ± 37 (9)^{b}	10 ± 6 (1)^{b}	278 ± 25 (28)^{c}



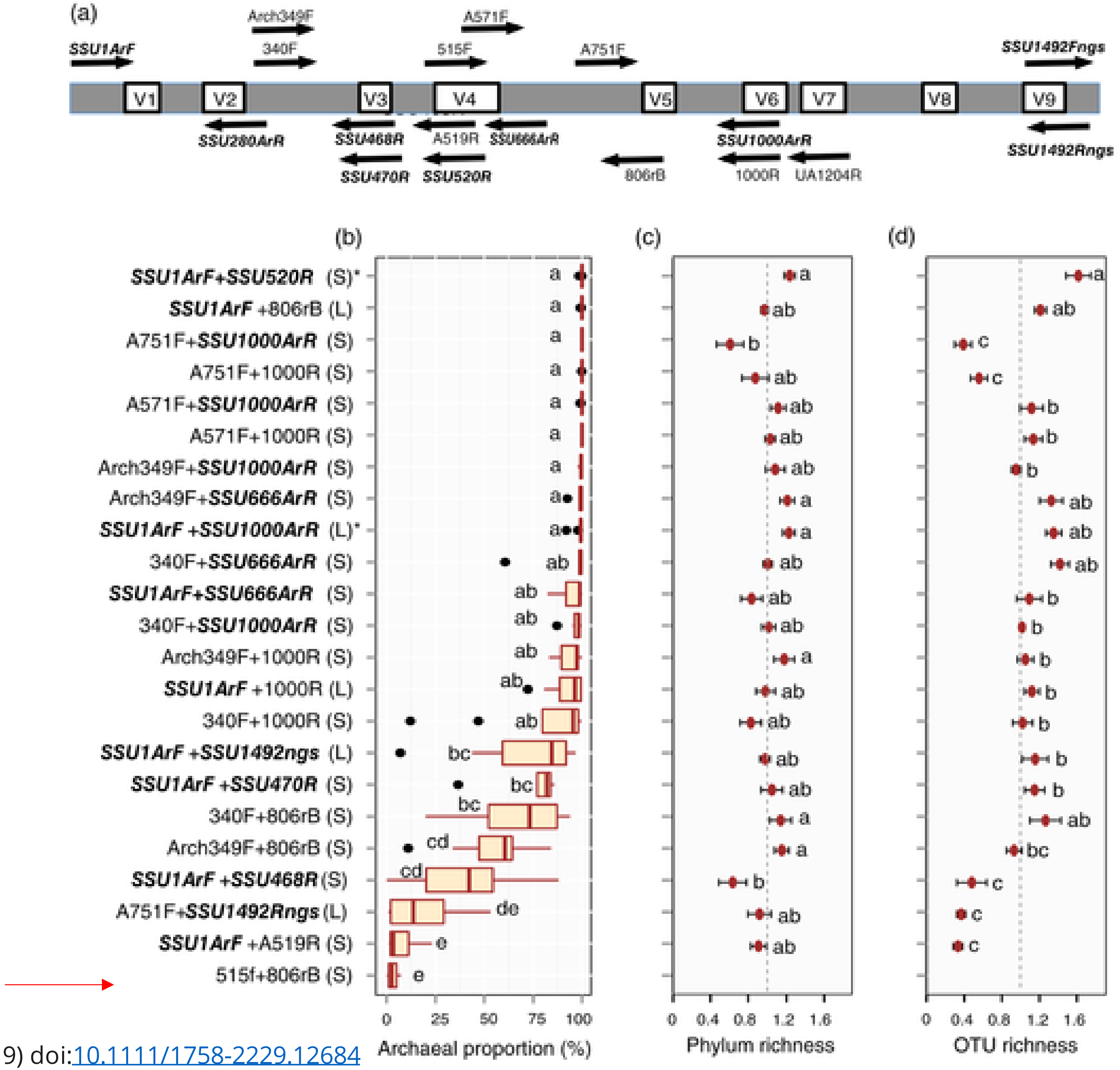
Beckers et al 2016

Other marker choice considerations

Primer bias

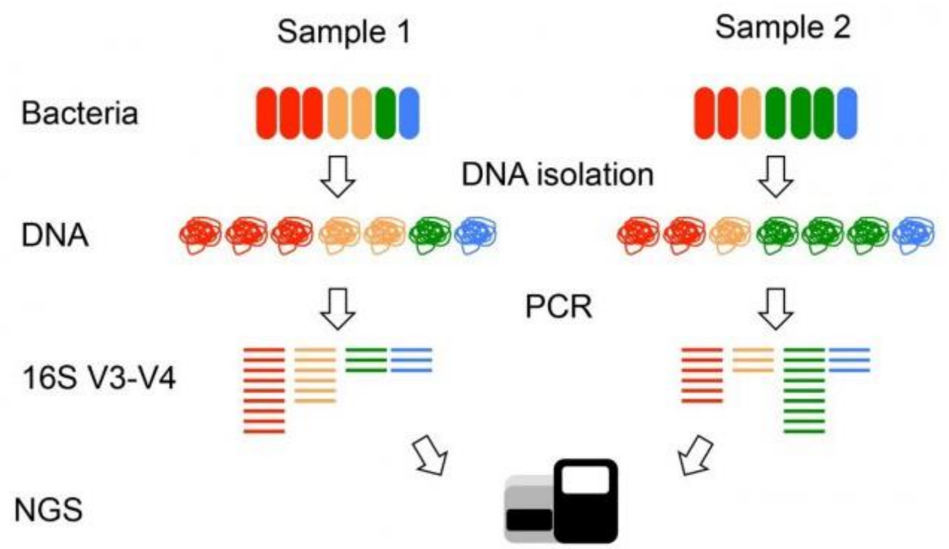
“The choice of primers dictates what [taxa] fungi will be recovered from the sample, and we recommend spending substantial time evaluating and choosing primers”

Nilsson et al 2019 doi: 10.1038/s41579-018-0116-y



Bahram et al (2019) doi:10.1111/1758-2229.12684

Controls

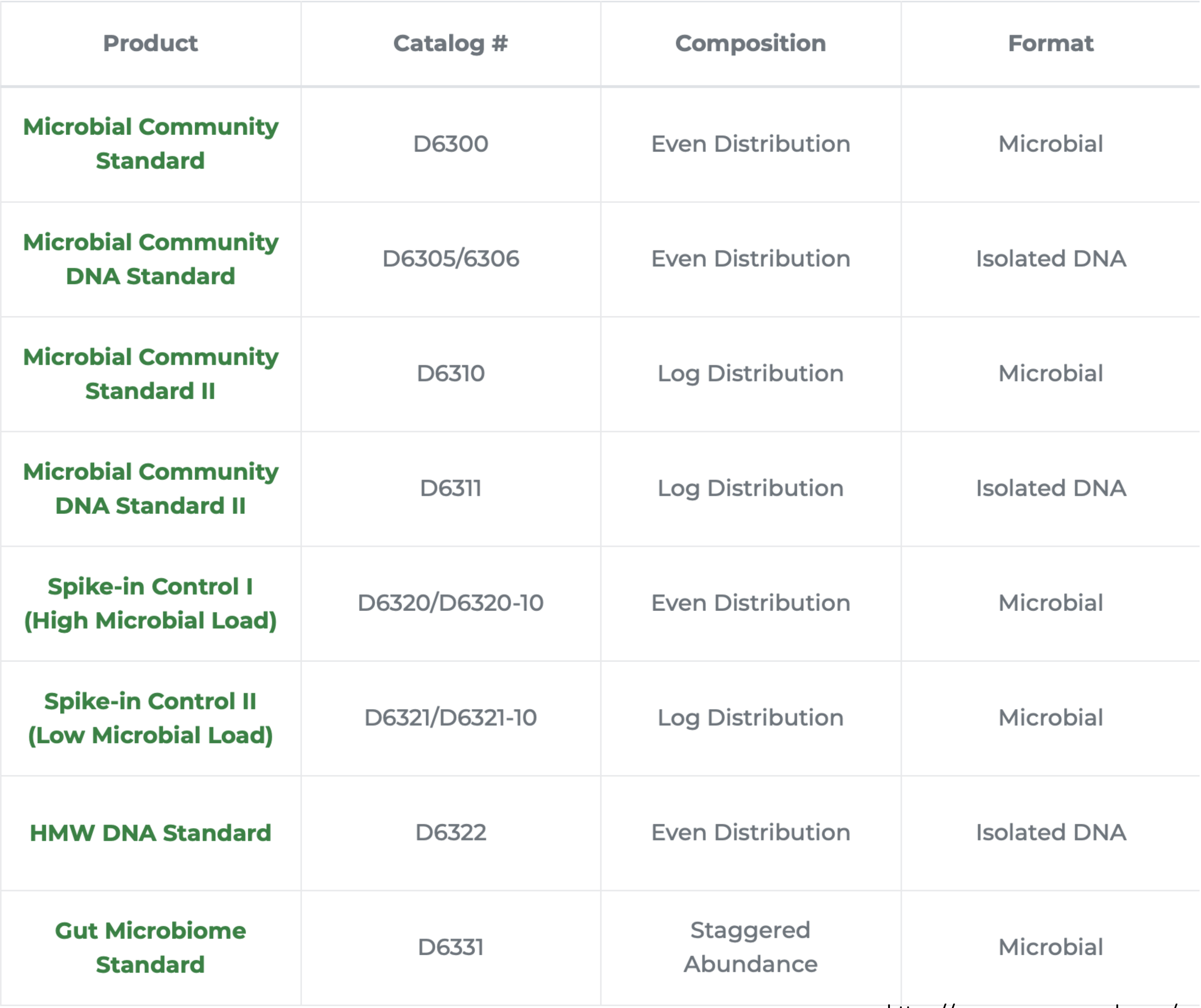


Positive **Negative**

Controls to consider

- Negative control 1 – Molecular biology grade water with no processing.
- Negative control 2 – Run molecular biology grade water through the entire pipeline, pool with other samples even if you do not see amplification.
- Positive Control – Mock communities – Homemade vs industry standard (Zymo, etc), DNA and known cultures

Mock communities and positive controls

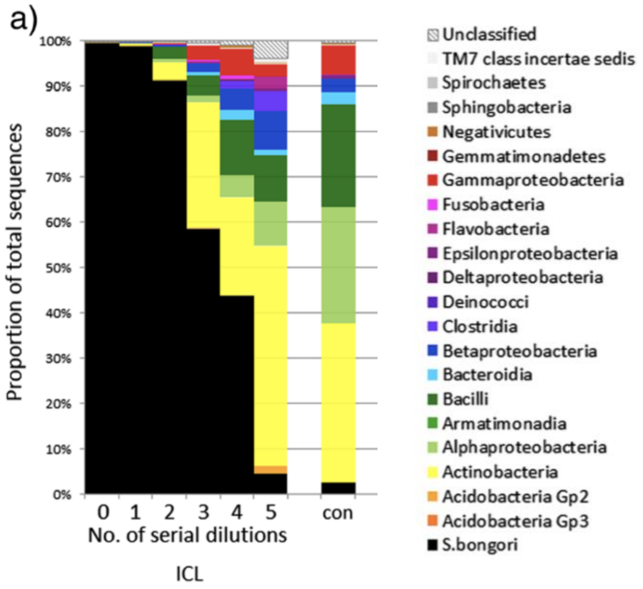


Product	Catalog #	Composition	Format
Microbial Community Standard	D6300	Even Distribution	Microbial
Microbial Community DNA Standard	D6305/6306	Even Distribution	Isolated DNA
Microbial Community Standard II	D6310	Log Distribution	Microbial
Microbial Community DNA Standard II	D6311	Log Distribution	Isolated DNA
Spike-in Control I (High Microbial Load)	D6320/D6320-10	Even Distribution	Microbial
Spike-in Control II (Low Microbial Load)	D6321/D6321-10	Log Distribution	Microbial
HMW DNA Standard	D6322	Even Distribution	Isolated DNA
Gut Microbiome Standard	D6331	Staggered Abundance	Microbial

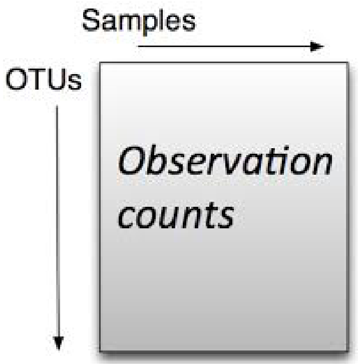
https://www.zymoresearch.com/pages/microbiome-standards

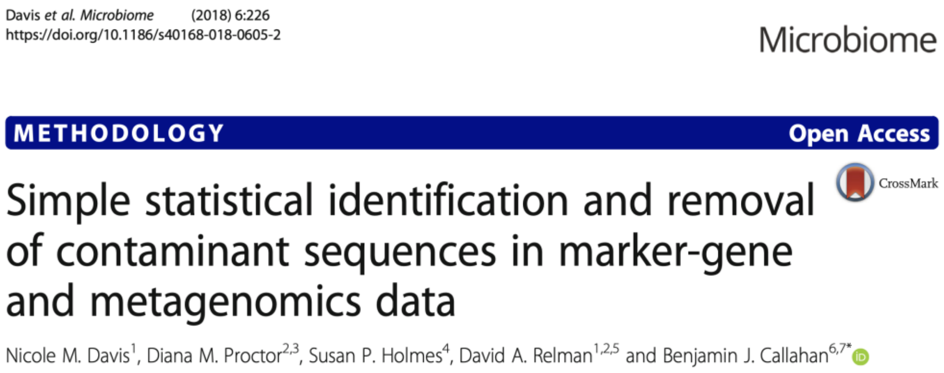
Negative controls and identifying contaminants

Contaminants in low biomass samples



con=template free PCR
http://www.biomedcentral.com/1741-7007/12/87





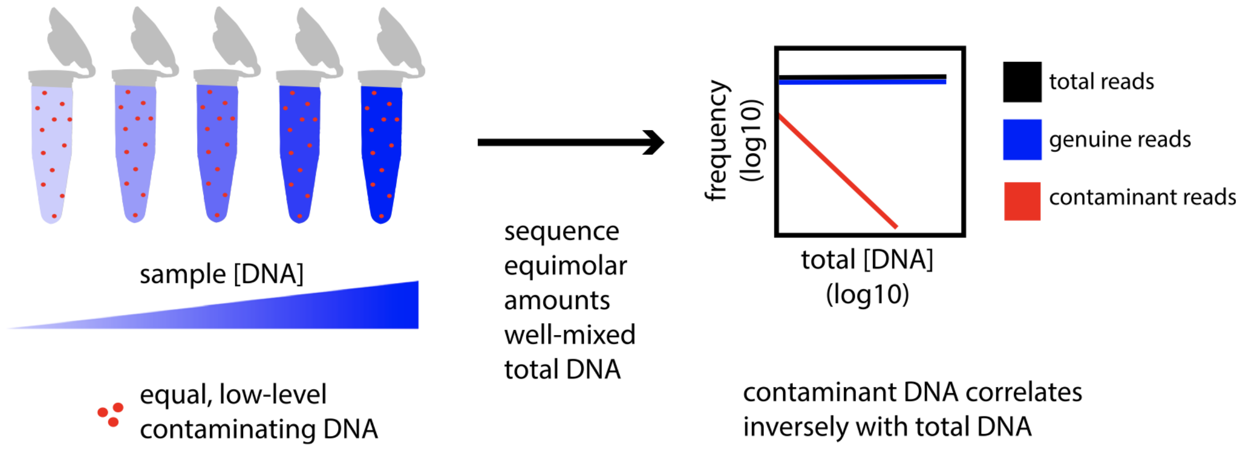
Davis *et al. Microbiome* (2018) 6:226
https://doi.org/10.1186/s40168-018-0605-2

Microbiome

METHODOLOGY **Open Access**

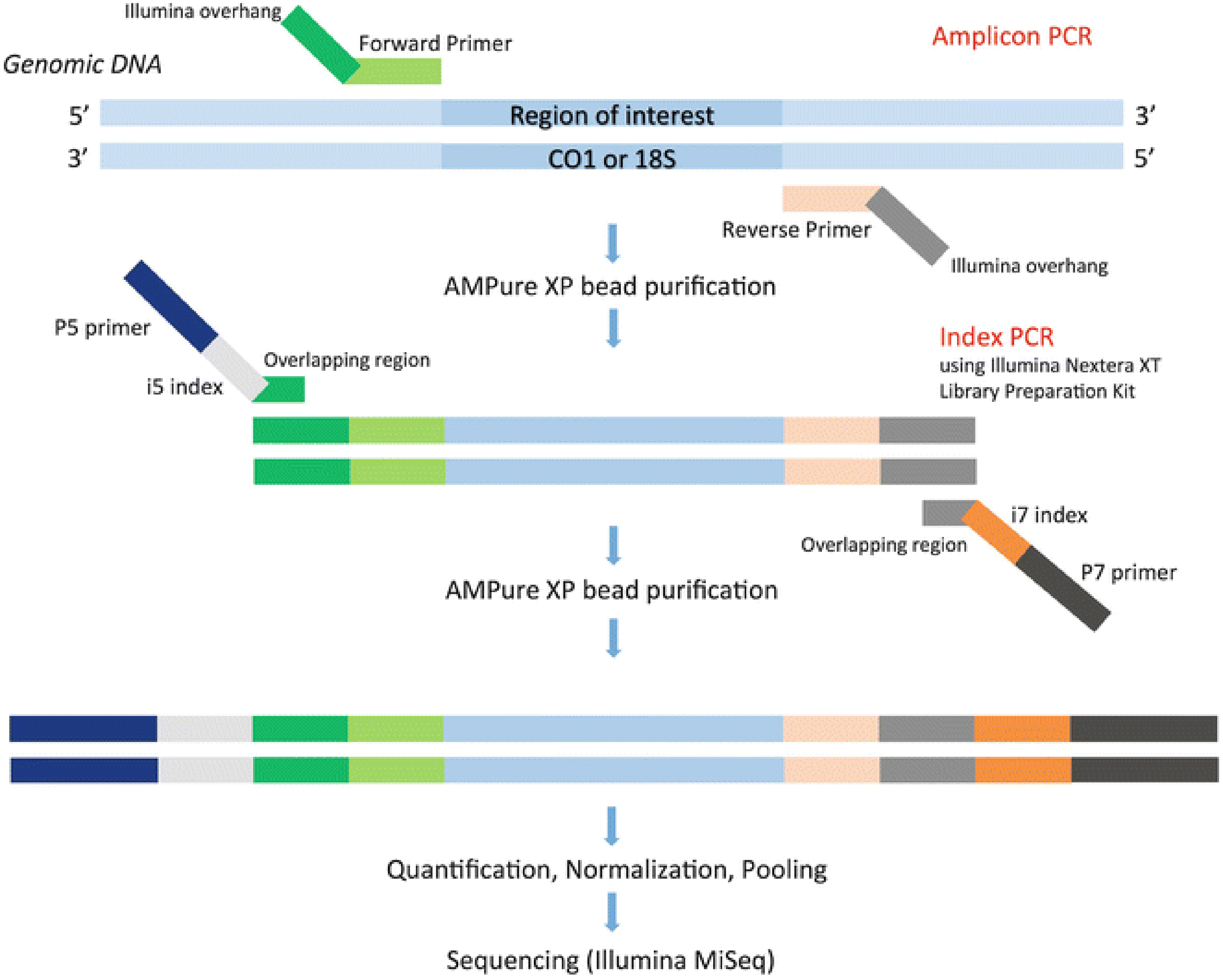
Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data

Nicole M. Davis[1], Diana M. Proctor[2,3], Susan P. Holmes[4], David A. Relman[1,2,5] and Benjamin J. Callahan[6,7*]


sample [DNA]
equal, low-level contaminating DNA
sequence equimolar amounts well-mixed total DNA
frequency (log10)
total [DNA] (log10)
total reads
genuine reads
contaminant reads
contaminant DNA correlates inversely with total DNA

Library Prep

1. PCR amplification - Amplify region of interest of our target gene, add sequencing specific adaptors
2. Adapter ligation - Add indexing and sequencing primers
3. Quantify, normalize, pool

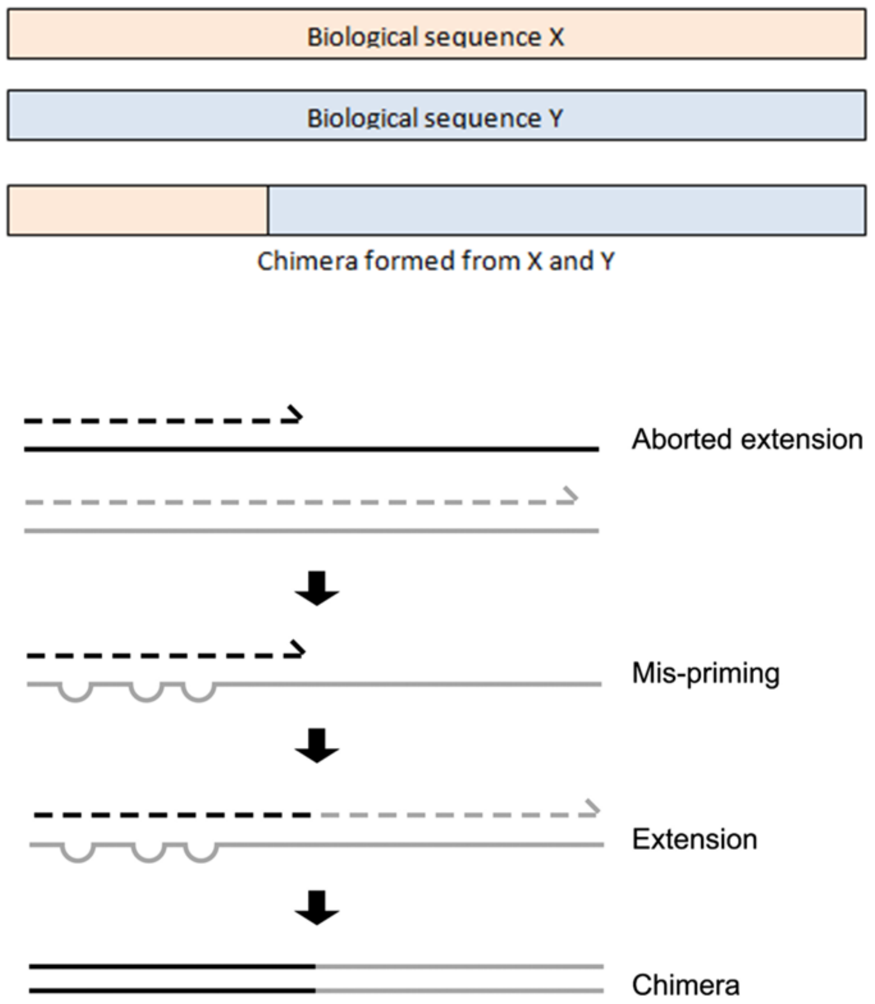


Bourlat, Sarah J., et al. 2016

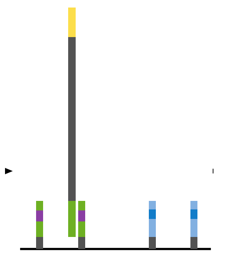
PCR also introduces bias

- Taq polymerases are error prone – making one error/1000bases – Solution: Use a Hi-Fi Polymerase
- Chimeras – Sequences formed by two or more biological sequences joined together
- To reduce chimeras - Use optimal annealing temperatures
- Minimize # of PCR cycles

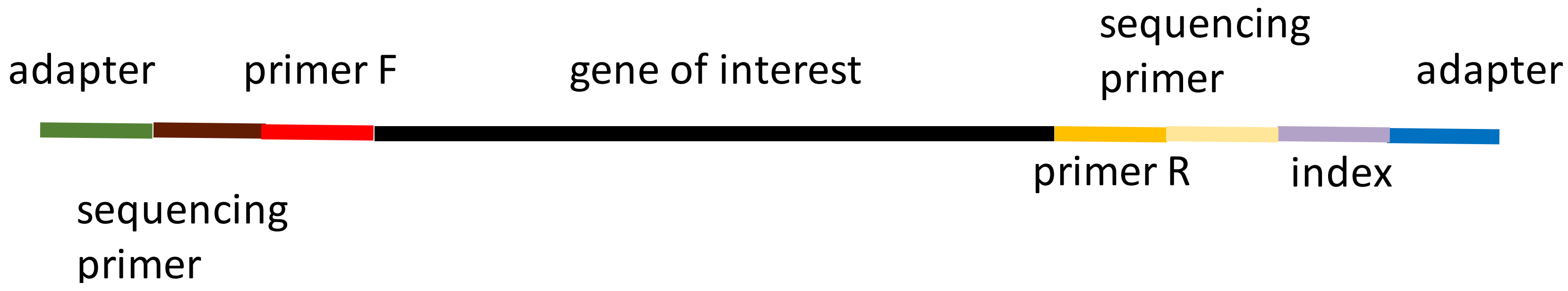
PCR and Chimeras

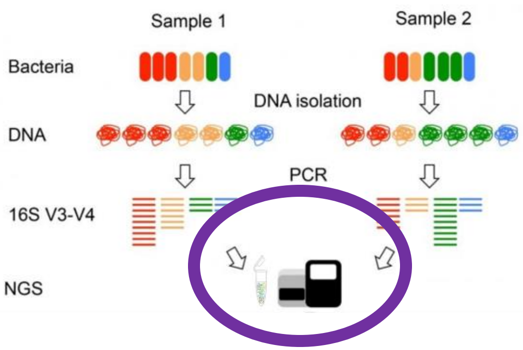


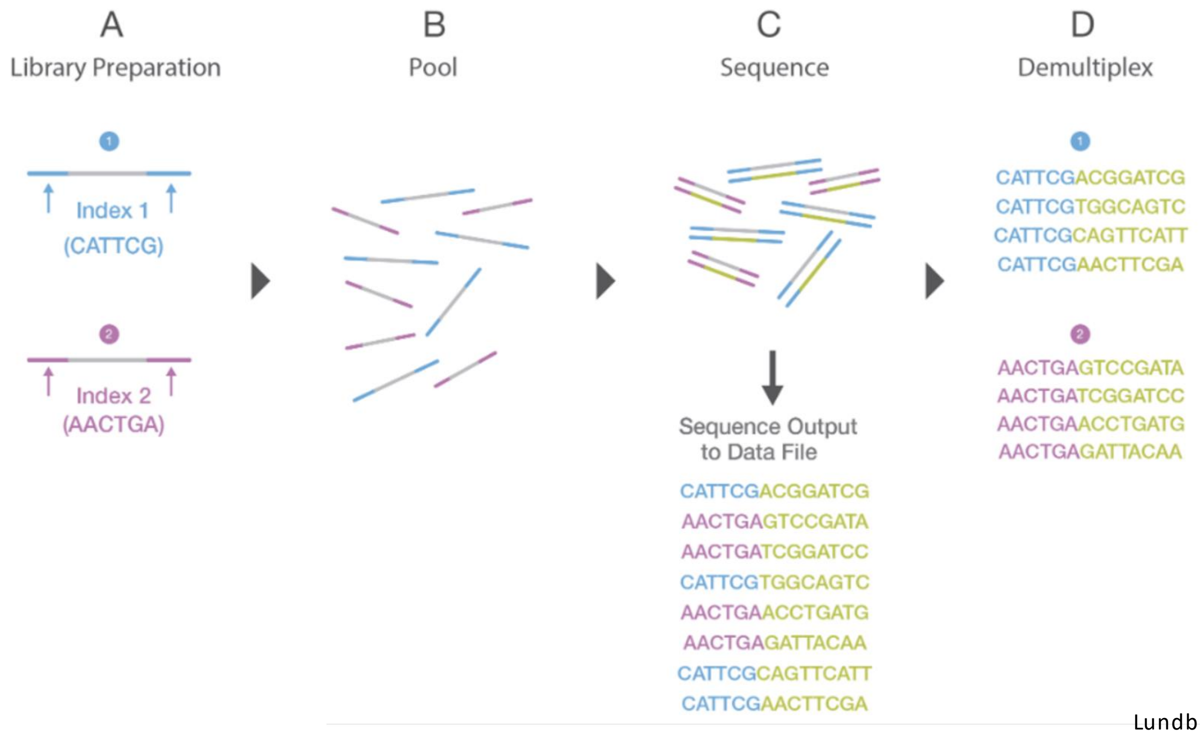
Structure of the amplicon (read) after library preparation



Illumina flow cell







CATTCGACGGATCG
AACTGAGTCCGATA
AACTGATCGGATCC
CATTCGTGGCAGTC
AACTGAACCTGATG
AACTGAGATTACAA
CATTCGCAGTTCATT
CATTCGAACTTCGA

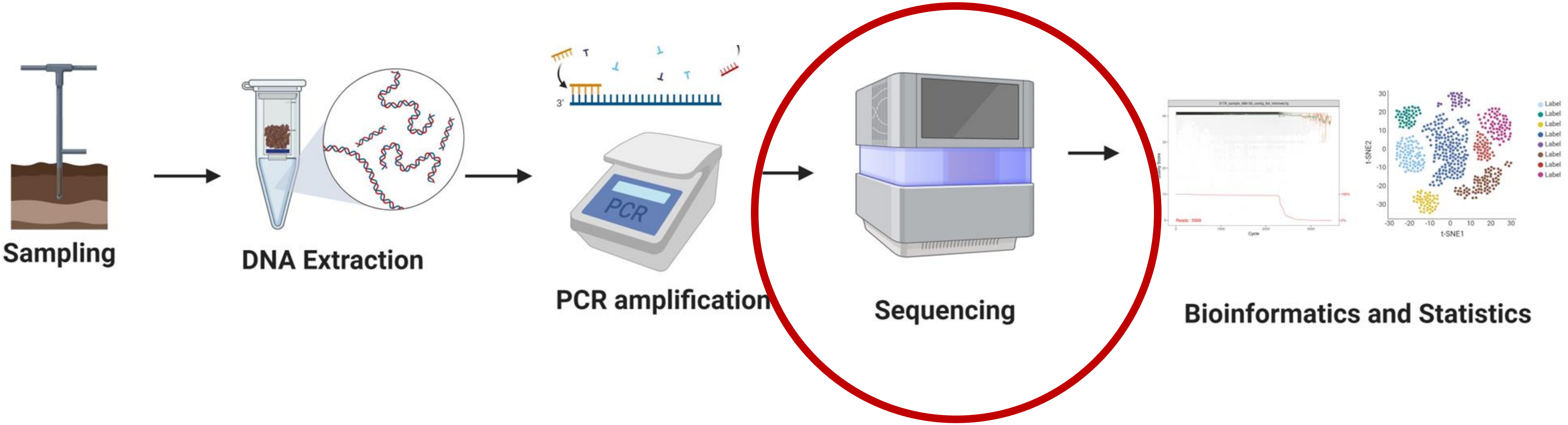
CATTCGACGGATCG
CATTCGTGGCAGTC
CATTCGCAGTTCATT
CATTCGAACTTCGA

AACTGAGTCCGATA
AACTGATCGGATCC
AACTGAACCTGATG
AACTGAGATTACAA

Lundberg et al 2013, Caporaso et al 2011,www.illumina.com

Metabarcoding – different indexes allows the sequencing of multiple samples at once

https://www.youtube.com/watch?v=mI0Fo9kaWqo



Sampling

DNA Extraction

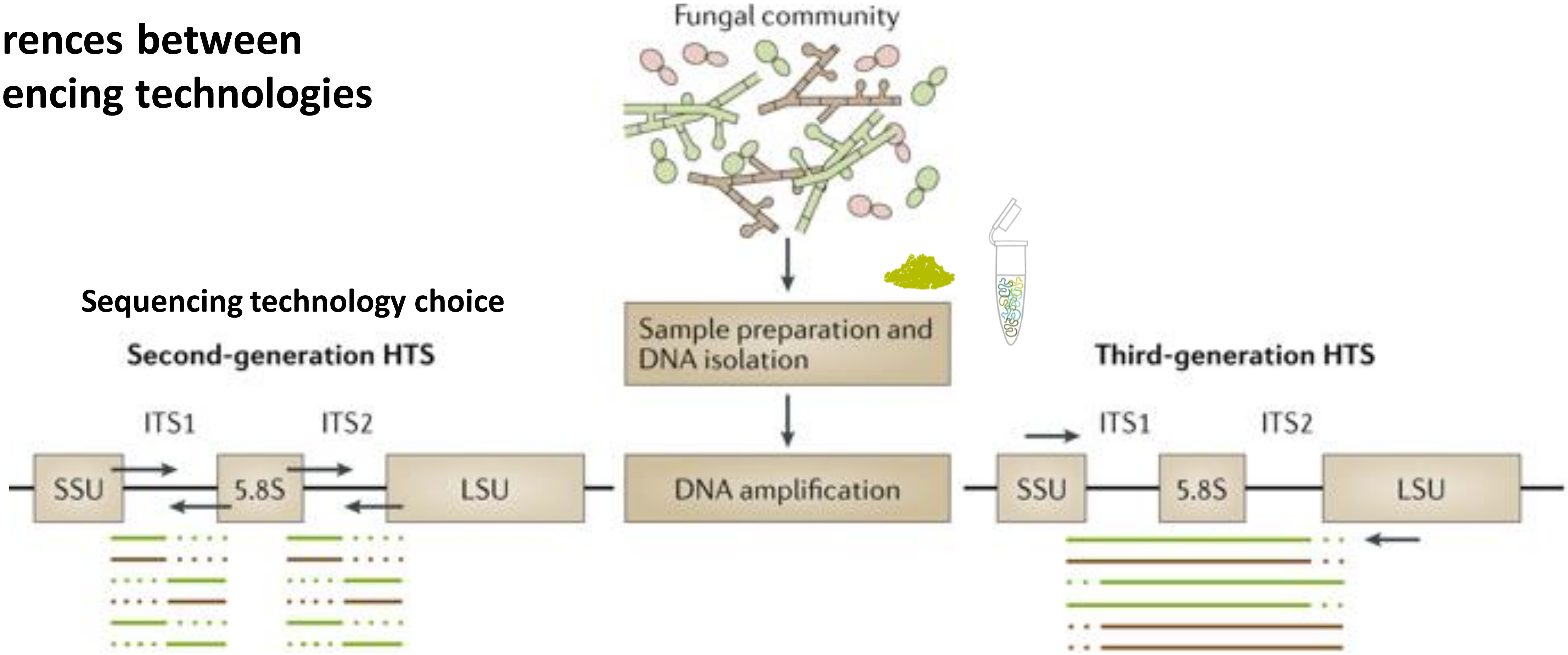
PCR amplification

Sequencing

Bioinformatics and Statistics

Differences between sequencing technologies

Sequencing technology choice



Illumina MiSeq $2300/run
2x300 (12 GB)
384 indexes available
~ 400 bp after quality trimming

Illumina NextSeq $2300/run
2x300 (60 GB)
384 indexes available
~ 400 bp after quality trimming

NovaSeq $5000/run
2x250 (0.5 TB of data)
~ 400 bp after quality trimming

PacBio $3000/run

- Full length ribosomal region
- 196 indexes available

Oxford Nanopore $1200/run

- Full-length ribosomal region kits (bact)
- 24 indexes available

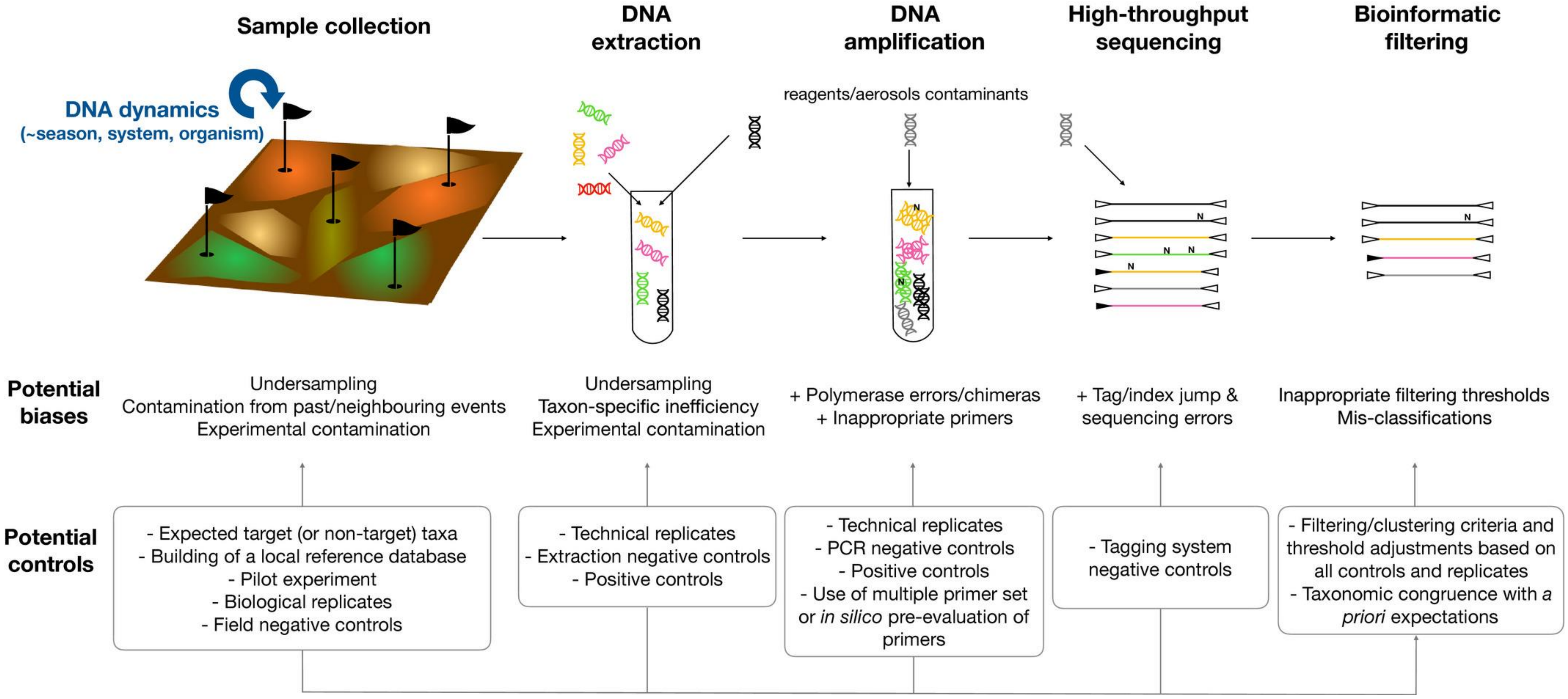
https://www.youtube.com/watch?v=mI0Fo9kaWqo

Nilsson et al 2019

Bias: Systematic error (vs. random error)

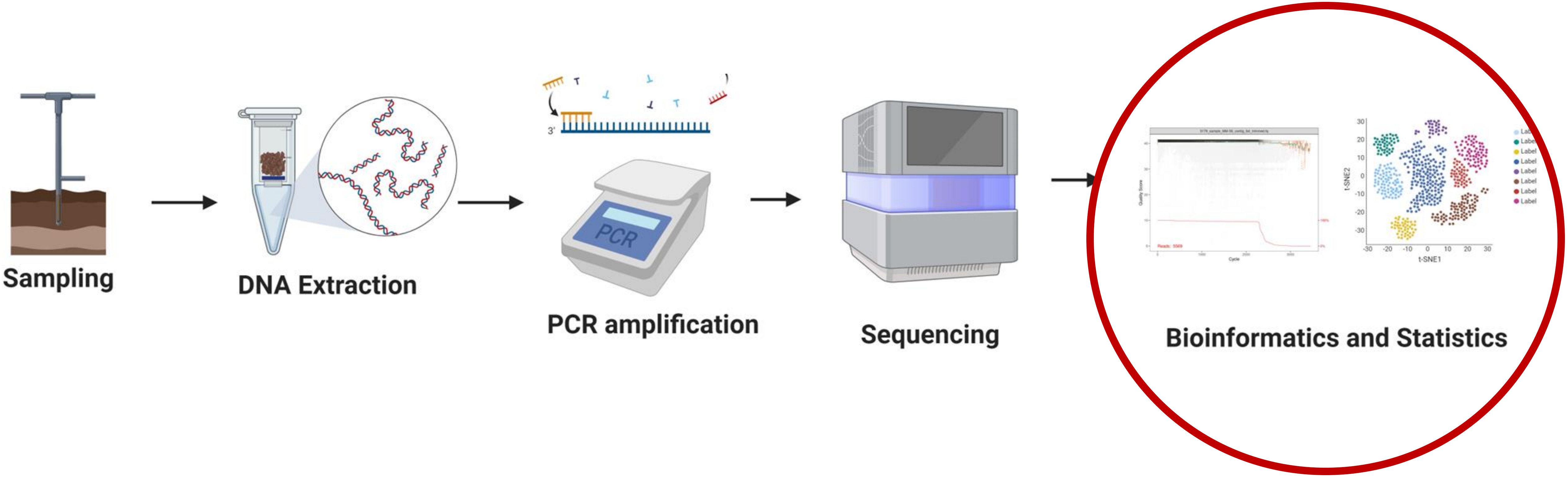
"MGS measurements are biased: The measured relative abundances of the taxa and genes in the sample are systematically distorted from their true values (Brooks, 2016; Sinha et al., 2017). Bias arises because each step in an experimental MGS workflow preferentially measures (i.e. preserves, extracts, amplifies, sequences, or bioinformatically identifies) some taxa over others."

McLaren et al 2019



	Sample collection	DNA extraction	DNA amplification	High-throughput sequencing	Bioinformatic filtering
Potential biases	Undersampling Contamination from past/neighbouring events Experimental contamination	Undersampling Taxon-specific inefficiency Experimental contamination	+ Polymerase errors/chimeras + Inappropriate primers	+ Tag/index jump & sequencing errors	Inappropriate filtering thresholds Mis-classifications
Potential controls	- Expected target (or non-target) taxa - Building of a local reference database - Pilot experiment - Biological replicates - Field negative controls	- Technical replicates - Extraction negative controls - Positive controls	- Technical replicates - PCR negative controls - Positive controls - Use of multiple primer set or *in silico* pre-evaluation of primers	- Tagging system negative controls	- Filtering/clustering criteria and threshold adjustments based on all controls and replicates - Taxonomic congruence with *a priori* expectations

Zinger etal 2019



Sampling → DNA Extraction → PCR amplification → Sequencing → Bioinformatics and Statistics