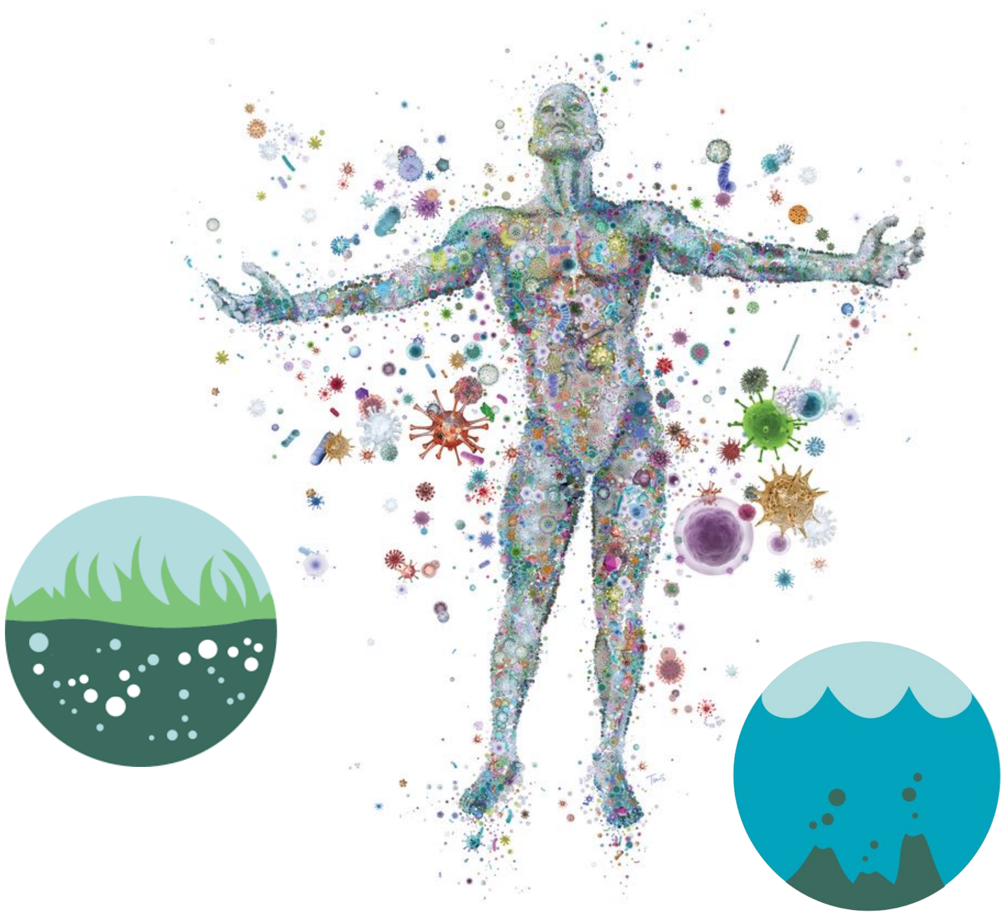
National Institute of Allergy and Infectious Diseases

METAGENOMICS OVERVIEW

Metataxonomics

MSB7105

March, 2021



Angelina Angelova, PhD

Bioinformatics and Computational Biosciences Branch (BCBB)

OCICB/OSMO/OD/NIAID/NIH

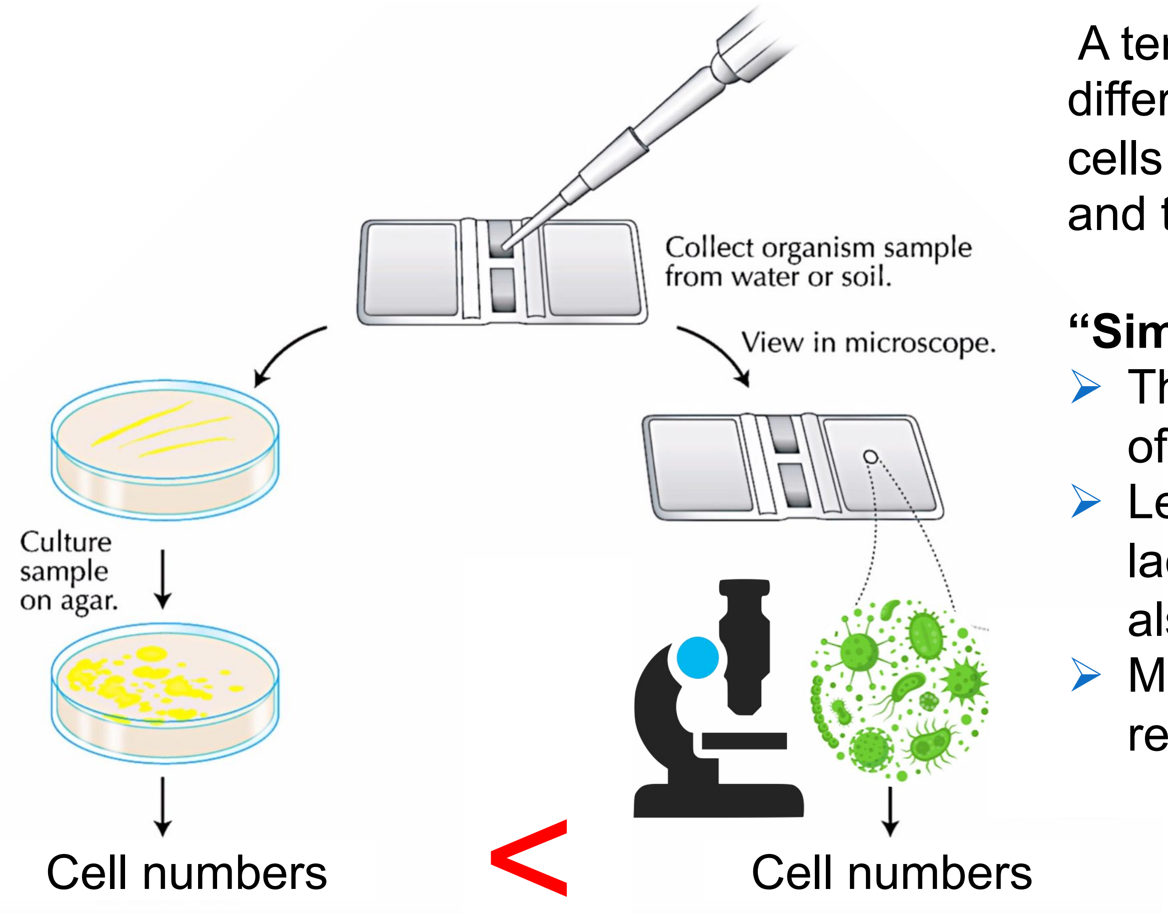
Today's instructor

Angelina Angelova, PhD
Metagenomics Analysis Specialist

Bioinformatics and Computational Biosciences Branch (BCBB)
National Institute of Allergies and Infectious Diseases (NIAID)
National Institute of Health (NIH)
Bethesda, MD, USA

Contact instructor:
angelina.angelova@nih.gov
Contact our team:
bioinformatics@niaid.nih.gov

The great plate count anomaly



A term coined by Staley & Konopka in 1985 to describe the difference (in orders of magnitude) between the number of cells from natural environments countable by microscopy and those observed after culturing on common agar media.

“Simple” genomics is not enough

- ➢ The microbial world is **extremely diverse** (construct ~1/2 of Earth’s biomass) and **largely unknown**
- ➢ Less than 1% of organisms are culturable due not only to lack of proper growth conditions for them in the lab, but also due to proper social interaction
- ➢ Microbes exist in complex communities & have complex relations between each other and larger organisms!

National Institute of Allergy and Infectious Diseases

https://www.ncbi.nlm.nih.gov/books/NBK54011/

Staley, J. T., and A. Konopka. 1985. Measurements of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu. Rev. Microbiol. 39:321-346. [PubMed] [Google Scholar]

Define ”Metagenomics”

- NGS made the field of metagenomics possible
- Metagenomics: Refers to the idea that the collection of genes (the metagenome), obtained directly from a community in its natural habitat (the microbiome), can provide an understanding of the function and characteristics of the whole community, in a similar way as the collection of genes from a single organism can provide an understanding of the function and identity of that organism.
- Metagenomics bypasses the need for isolation or cultivation of individual microbes.
- Allows for exploration of the structure (abundance & identities), interactions, strategies (communication, survival, etc.), functionality and dynamics of a community

Example microbiomes:

NIH National Institute of Allergy and Infectious Diseases

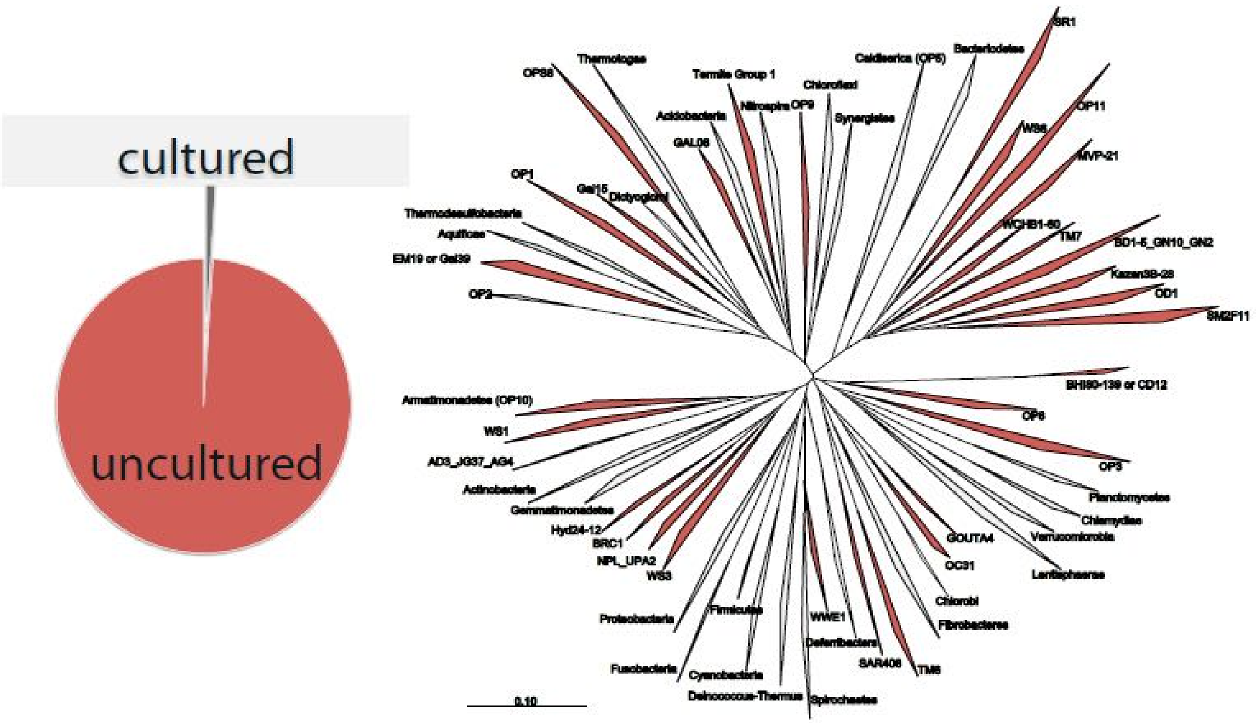
https://www.ncbi.nlm.nih.gov/books/NBK54011/

NIAID

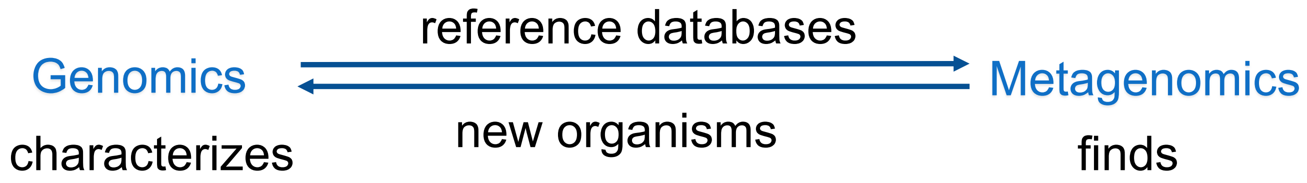
‘Shining a light’ on microbial ‘dark matter’

- Culture-based techniques are very limited to only what is cultivatable and produce a strong bias towards exploration of only cultivatable organisms, excluding > 99% of microorganisms from exploration
- Metagenomics enables scientists to explore this microbial ‘dark matter’
- Vast applications:
  - Biotechnology & Medicine
  - Environmental preservation & recovery
- The more organisms can recognize, the more we expand our capacity to ’see’ new species.

Our skewed view of the microbial world



16S rRNA tree of known bacterial phyla



https://www.scientificamerican.com/article/most-microbial-species-are-dark-matter/
http://www.calit2.net/newsroom/print_page.php?id=2160

Reference genomic databases

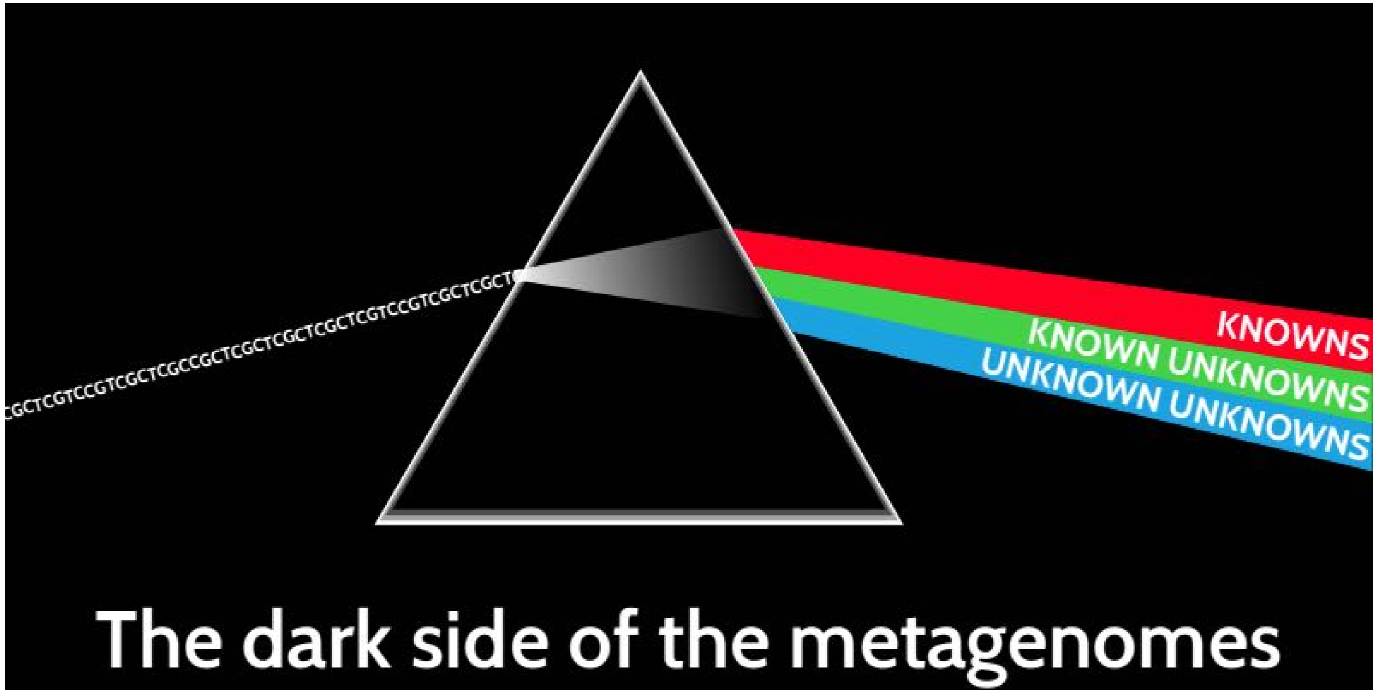
A reference genomic databases are a collection of DNA sequences that are idealistic genomic representations of recognized organisms. These sequences are sourced either from individual cultivated organisms (a type strain representing that lineage) or in case of more complex organisms – from multiple organisms from the same species (e.g. human).

- RefDBs allow for the characterization of
  - a specific species through identification of conserved genes within that organism's genome (genetic markers).
  - a specific function through identification of known genes/proteins observed within the organism

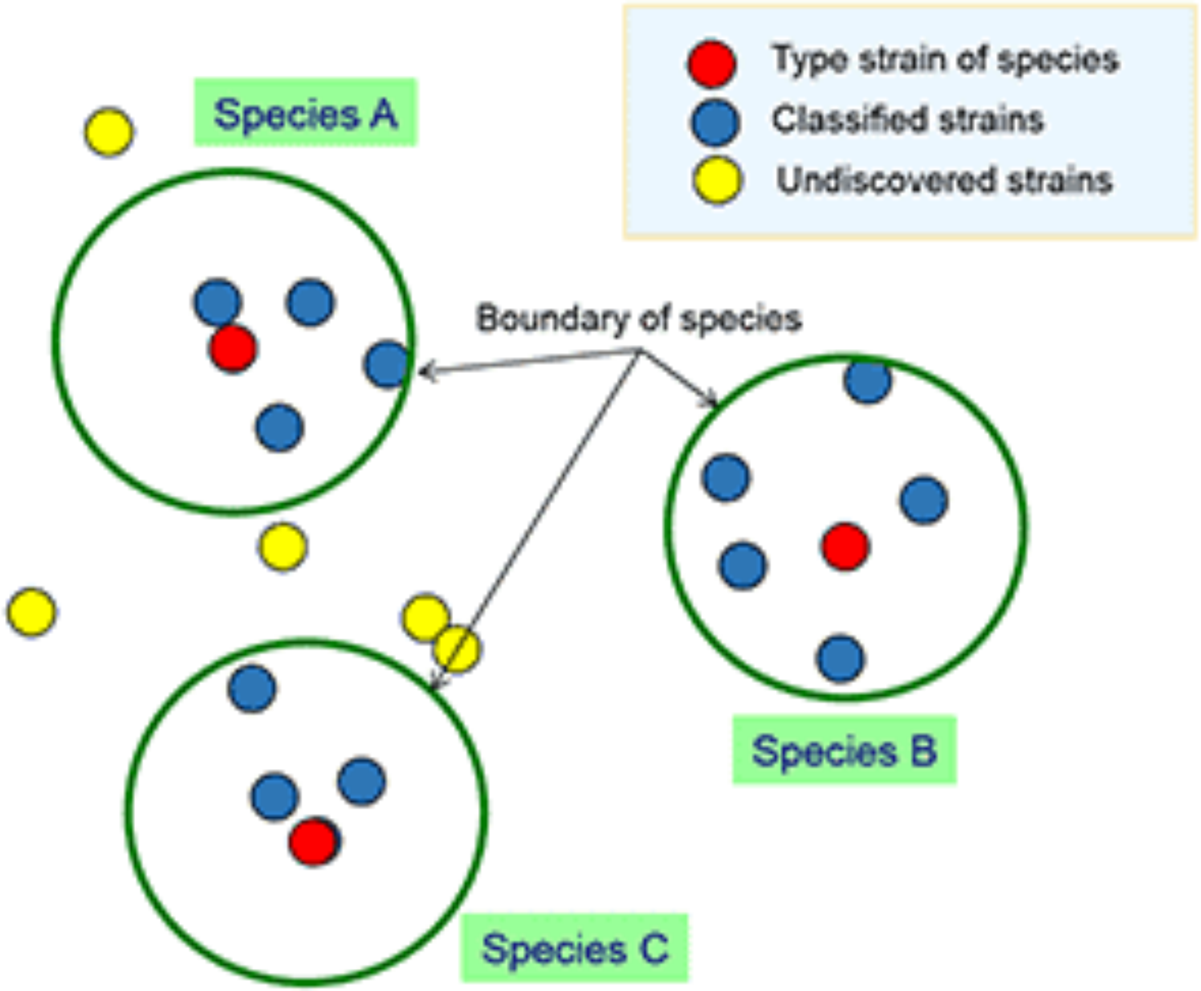
- RefDB are important for
  - Propper phylogenic & functional assignments of unknown sequences
  - Understanding genomic structures, functional capacity & survival strategies of organisms
  - Guiding assembly software & genome mining tools

Can you think of any genetic markers commonly used in microbial identification?

Genetic markers are used for identification of organisms or function. These are ideally a single-copy genes with universal presence but internal variability in all organisms


CGCTCGTCCGTCGCTCGCCGCTCGCTCGCTCGCTCGTCCGTCGCTCGCT
KNOWNS
KNOWN UNKNOWNS
UNKNOWN UNKNOWNS
The dark side of the metagenomes

Can you recall a name of a reference sequence database?



National Institute of Allergy and Infectious Diseases

Shotgun VS Amplicon

Shotgun Strategy

ALL the DNA from ALL the genomes within the ENTIRE community, is fragmented to the “bite-size” capacity of a sequencing platform. ALL DNA is sequenced. The sequences are used to explore taxonomic composition *and* functional capacity of the entire community

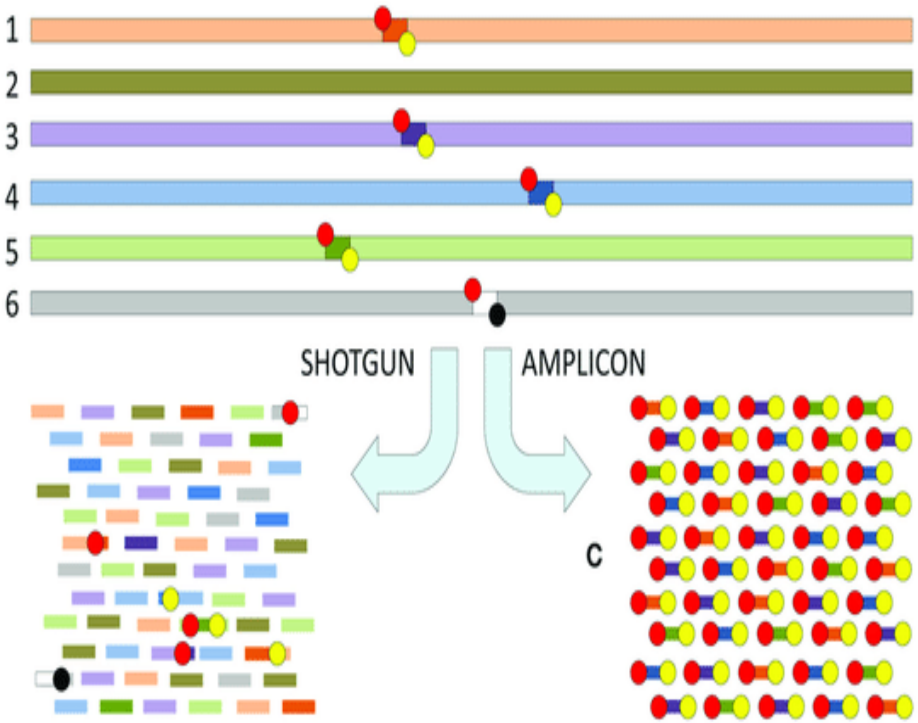
Long-read Strategy

ALL the DNA from ALL the genomes within the ENTIRE community, is sequenced in “large bites”. The sequences are used to explore taxonomic composition *and* functional capacity of the entire community

Common platforms

For long reads:

- PacBio, Nanopore (MinIon)



Common platforms

For Short reads:

- Illumina HiSeq, NextSeq, NovoSeq

Amplicon Strategy

One gene (a marker gene or a fraction of it) from ALL the genes from within ALL the genomes of ALL the organisms in a community, is targeted for amplification. Its sequence is used to explore the taxonomic composition of the entire community.

Common marker genes:

- For Bacterial & Archaeal organisms:
  - 16S rRNA gene
- For Eukaryotic organisms:
  - 18S rRNA *gene* (less conserved)
  - ITS: internal transcribed spacer region

Common platforms

For Short reads:

- Illumina MiSeq, NextSeq
- TermoFisher IonTorrent

NIH National Institute of Allergy and Infectious Diseases

JOURNAL OF CLINICAL MICROBIOLOGY, Sept. 2007, 45 (9), https://jcm.asm.org/content/45/9/2761.short
PNAS April 17, 2012 109 (16) 6241-6246; https://doi.org/10.1073/pnas.1117018109
Nature Biotechnology. Sept. 2017. https://doi.org/10.1038/nbt.3935

NIAID

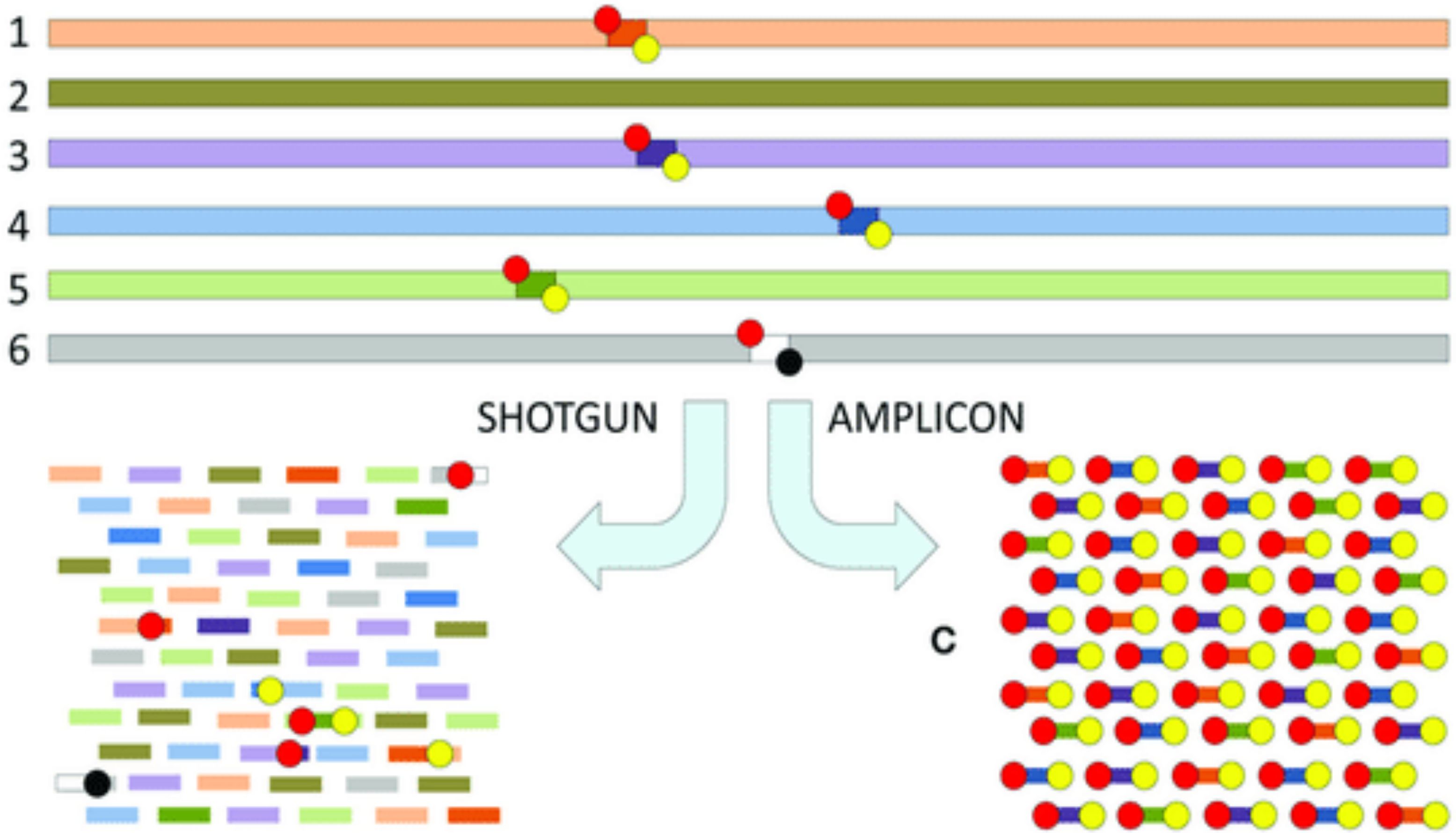
Why is long-read sequencing not commonly used?

- Because DNA gets commonly shredded during extraction, anyway.
- Because **long-read platforms have a lower resolution for the rarer organisms or organisms with smaller genomes**, so such will often be unrepresented
- Because bacterial genomes are often not “that big” and long-read strategy is often overkill.
- Because long-read strategy still creates sequence errors which inflates the diversity of a community

Long-read sequencing can still be used & be quite useful depending on what one is looking to explore (e.g. micro-eukaryotic communities!)

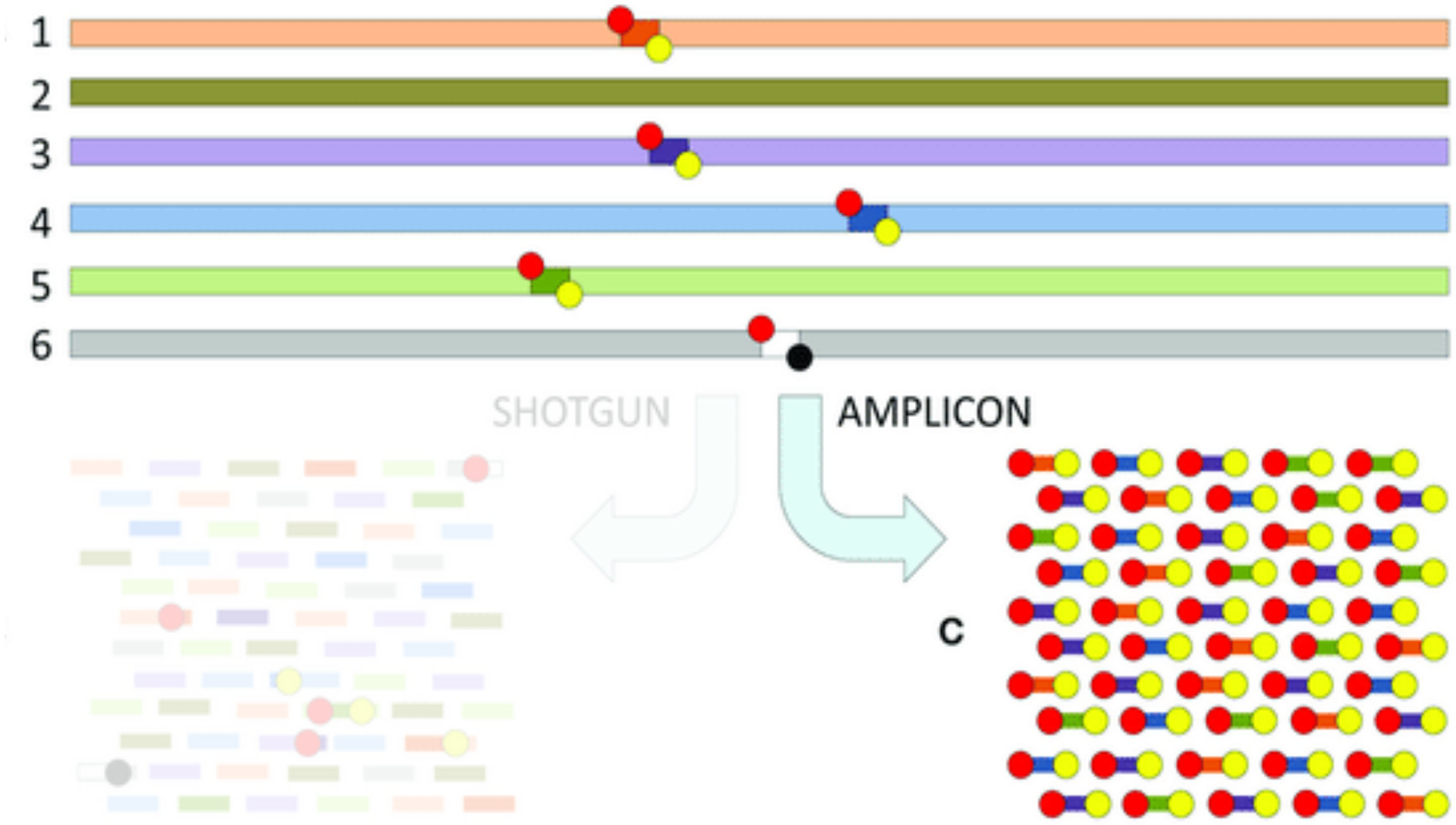
Microbiome exploration strategies

Metagenomics



Metataxonomics

Amplicon-based community exploration



Metataxonomics

DNA indexing
DNA barcoding
Marker gene sequencing
Amplicon sequencing

National Institute of Allergy and Infectious Diseases

Down rabbit hole #1

NIAID

NIAID

16S rRNA gene characteristics

Ubiquitous gene, found in all prokaryotes

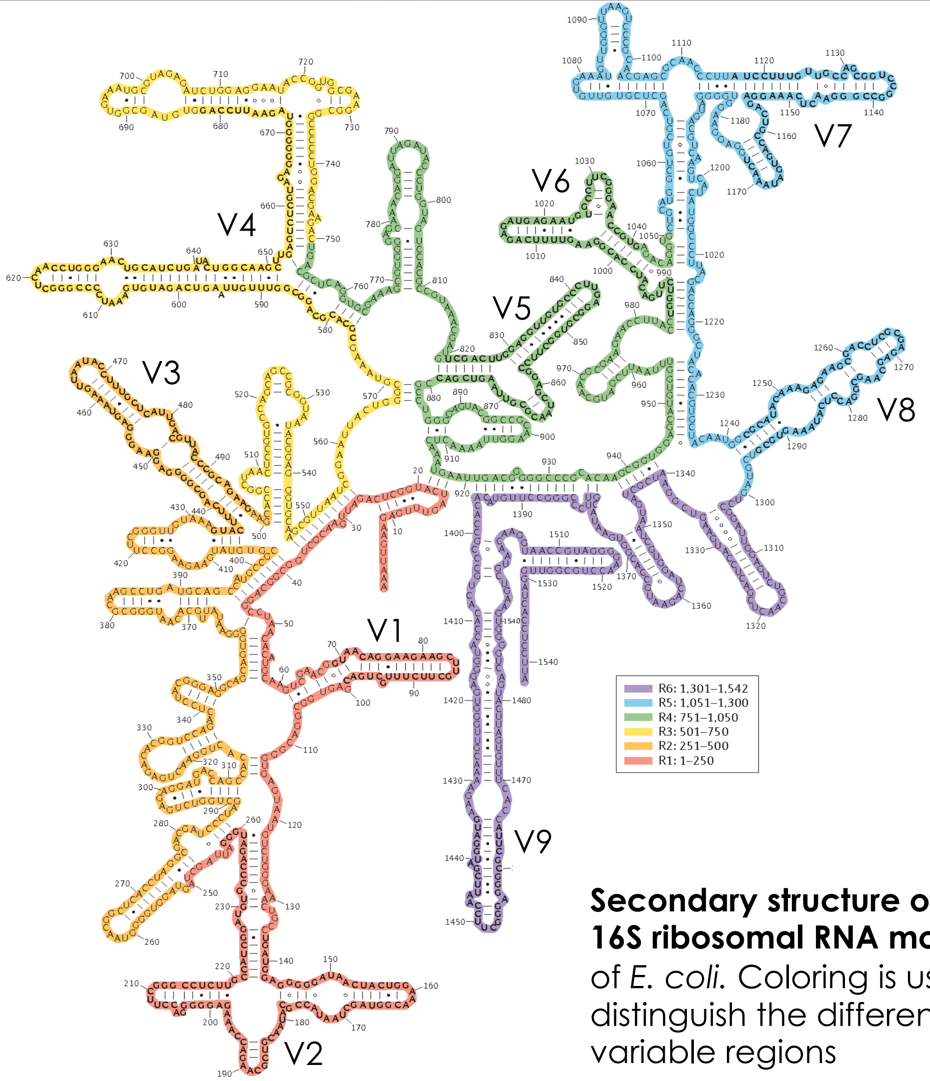
Most sequenced gene!

16S rRNA gene-specific primers are not only numerous, but also well characterized.

Current gold standard as evolutionary marker

Cost effective approach for exploring a community

Availability of large 16S rRNA gene databases (e.g. RDP, SILVA).

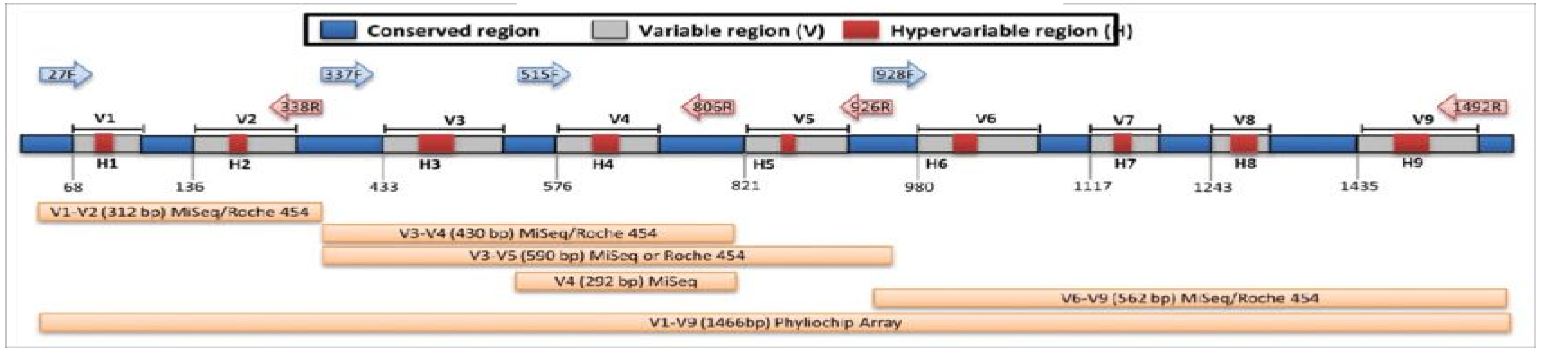


Secondary structure of the 16S ribosomal RNA molecule of *E. coli*. Coloring is used to distinguish the different variable regions

Length: ~1,600bp

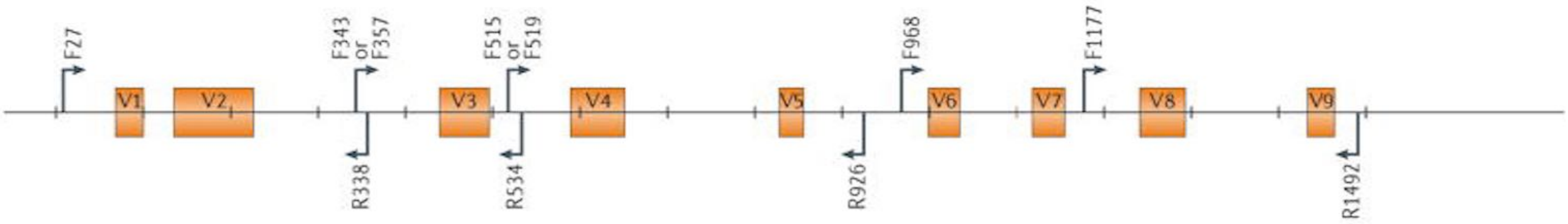
Structure: contains

- highly conserved regions: enabling gene targeting across species; and
- highly variable regions: enabling taxonomic characterization



Int J Syst Bacteriol. 1992. doi: 10.1099/00207713-42-1-166
Nat Rev Microboil, 2014. https://doi.org/10.1038/nrmicro3330

Common universal 16S rRNA Primers



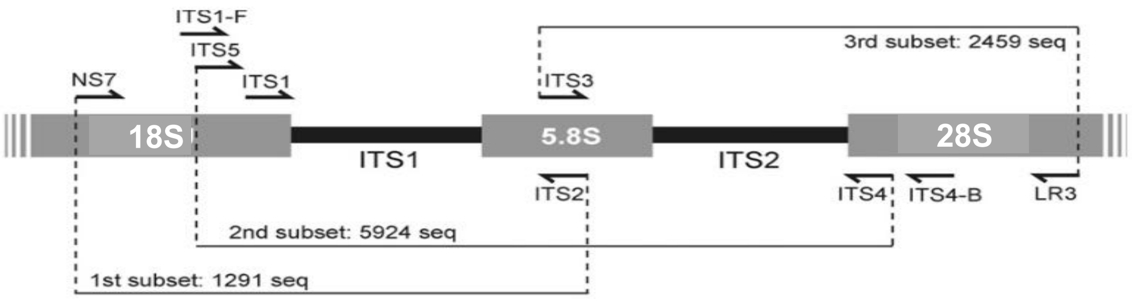
- F27: AGAGTTTGATCCTGGCTCAG
- F343: CCTACGGGNGGCWGCAG
- F515: GTGYCAGCMGCCGCGGTAA
- F968: ACGCGHNRAACCTTACC
- F1177: ACGTCATCCCCACCTTCC

- R1492: CGGNTACCTTGTTACGACTT
- R926: CCGYCAATTCMTTTRAGTTT
- R785: GACTACHVGGGTATCTAATCC
- R534: GTWTTACCGCGGCTGCTGG
- R338: GCTGCCTCCCGTAGGAGT

National Institute of
Allergy and
Infectious Diseases

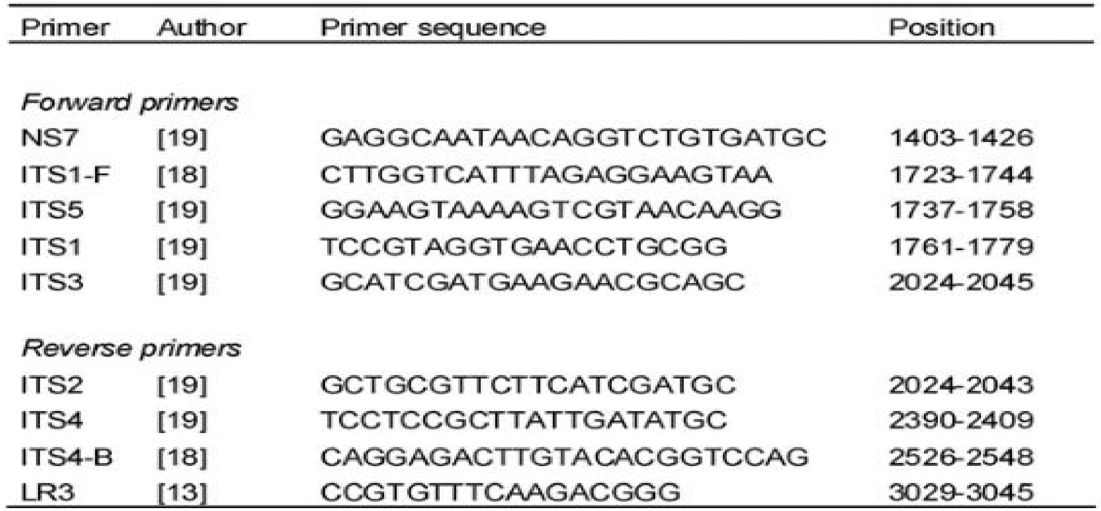
Kuczynski et al. 2012. Nat Rev Genet 13. doi: https://doi.org/10.1038/nrg3129

Common universal ITS & 18S primers



Fungal equivalent of 16S rRNA marker gene is the **ITS region** (Internal transcribed spacer region) Unlike 16S or 18S rRNA amplicons, the ITS regions can drastically vary in length and sequence between species, complicating sequence processing

Common ITS primers



Primer	Author	Primer sequence	Position
Forward primers			
NS7	[19]	GAGGCAATAACAGGTCTGTGATGC	1403-1426
ITS1-F	[18]	CTTGGTCATTTAGAGGAAGTAA	1723-1744
ITS5	[19]	GGAAGTAAAAGTCGTAACAAGG	1737-1758
ITS1	[19]	TCCGTAGGTGAACCTGCGG	1761-1779
ITS3	[19]	GCATCGATGAAGAACGCAGC	2024-2045
Reverse primers			
ITS2	[19]	GCTGCGTTCTTCATCGATGC	2024-2043
ITS4	[19]	TCCTCCGCTTATTGATATGC	2390-2409
ITS4-B	[18]	CAGGAGACTTGTACACGGTCCAG	2526-2548
LR3	[13]	CCGTGTTTCAAGACGGG	3029-3045

Figure 1 Commonly used primers for amplifying parts or the entirety of the ITS region. a) Relative position of the primers, design of the subsets and number of sequences in each subset. b) Primer sequences, references and position of the primer sequence according to a reference sequence of *Serpula himantioides* (AM946630) stretching the entire nrDNA repeat.

Common 18S rRNA gene primers

- F563: TGC CAG CAG CCG CGG TAA TTC C
- R1150: CCG TCA ATT CCT TTA AGT TT
- F1267: GGT GGT GCA TGG CCG TTC TTA G
- R1644: GAC GGG CGG TGT GTA CAA AGG

Hadziavdic et al. 2014. https://doi.org/10.1371/journal.pone.0087624

National Institute of Allergy and Infectious Diseases

Belleman et al. 2010. BMC Bicrobi. doi: https://doi.org/10.1186/1471-2180-10-189
Usyk et al. 2017. doi: 10.1128/mSphere.00488-17

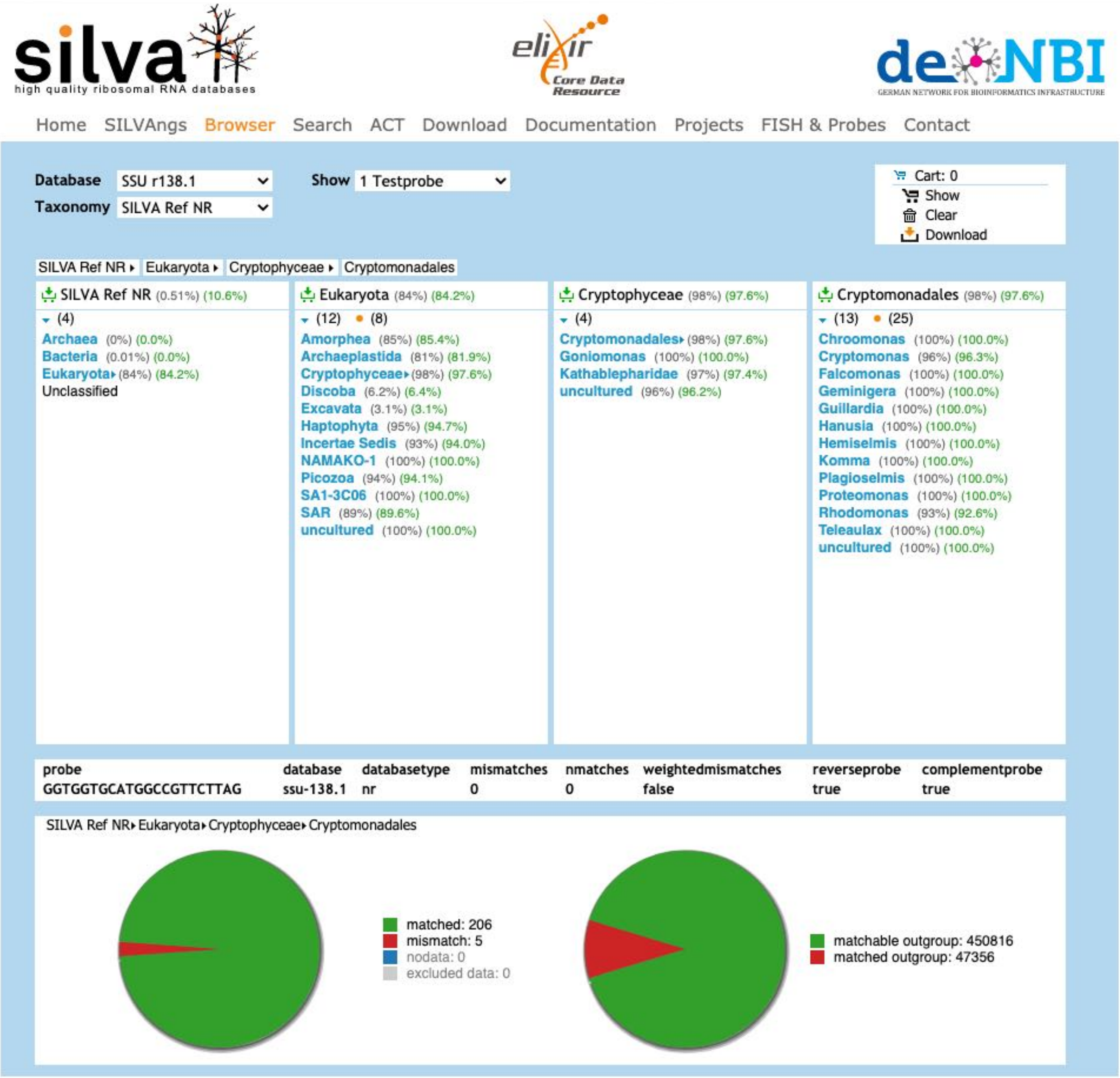
NIAID

Selecting your primers (TestProbe 3.0)

Mystery primers

Df514	TCC AGC TCC AAT AGC GTA
Dr1069	TCT TTA AGT TTC AGC CTT GC

Primers you select will drastically affect the resolution and visibility of different phylogenies. You can test ANY primer at https://www.arb-silva.de/search/testprobe/ to determine *in silico* a primer's ability to "recognize" specific lineages of organisms, prior to sequencing



Database SSU r138.1
Taxonomy SILVA Ref NR
Show 1 Testprobe

SILVA Ref NR ▸ Eukaryota ▸ Cryptophyceae ▸ Cryptomonadales

SILVA Ref NR (0.51%) (10.6%)
(4)
Archaea (0%) (0.0%)
Bacteria (0.01%) (0.0%)
Eukaryota ▸ (84%) (84.2%)
Unclassified

Eukaryota (84%) (84.2%)
(12) ● (8)
Amorphea (85%) (85.4%)
Archaeplastida (81%) (81.9%)
Cryptophyceae ▸ (98%) (97.6%)
Discoba (6.2%) (6.4%)
Excavata (3.1%) (3.1%)
Haptophyta (95%) (94.7%)
Incertae Sedis (93%) (94.0%)
NAMAKO-1 (100%) (100.0%)
Picozoa (94%) (94.1%)
SA1-3C06 (100%) (100.0%)
SAR (89%) (89.6%)
uncultured (100%) (100.0%)

Cryptophyceae (98%) (97.6%)
(4)
Cryptomonadales ▸ (98%) (97.6%)
Goniomonas (100%) (100.0%)
Kathablepharidae (97%) (97.4%)
uncultured (96%) (96.2%)

Cryptomonadales (98%) (97.6%)
(13) ● (25)
Chroomonas (100%) (100.0%)
Cryptomonas (96%) (96.3%)
Falcomonas (100%) (100.0%)
Geminigera (100%) (100.0%)
Guillardia (100%) (100.0%)
Hanusia (100%) (100.0%)
Hemiselmis (100%) (100.0%)
Komma (100%) (100.0%)
Plagioselmis (100%) (100.0%)
Proteomonas (100%) (100.0%)
Rhodomonas (93%) (92.6%)
Teleaulax (100%) (100.0%)
uncultured (100%) (100.0%)

probe	database	databasetype	mismatches	nmatches	weightedmismatches	reverseprobe	complementprobe
GGTGGTGCATGGCCGTTCTTAG	ssu-138.1	nr	0	0	false	true	true

SILVA Ref NR ▸ Eukaryota ▸ Cryptophyceae ▸ Cryptomonadales

PCR primer scaffolds for high-throughput sequencing of 16S rRNA gene

Primers scaffold:

Adaptors help amplicon for sequencing attach to flow cell

Primer pads help avoid hairpin and primer dimers within the primer scaffolds

Locus-specific sequence is targets & anneals to the region of interest

FWD scaffold

5'-	AATGATACGGCGACCACCGAGATCTACACGCT	XXXXXXXXXXXX	TATGGTAATT	GT	GTGYCAGCMGCCGCGGTAA	-3'
	5' Illumina adapter P5	Barcode	Primer pad	Primer linker	Locus-specific sequence	515F primer sequence

REV scaffold

5'-	CAAGCAGAAGACGGCATACGAGAT	XXXXXXXXXXXX	AGTCAGCCAG	CC	GGACTACNVGGGTWTCAAT	-3'
	3' Illumina adapter P7	Barcode	Primer pad	Primer linker	Locus-specific sequence	806R primer sequence

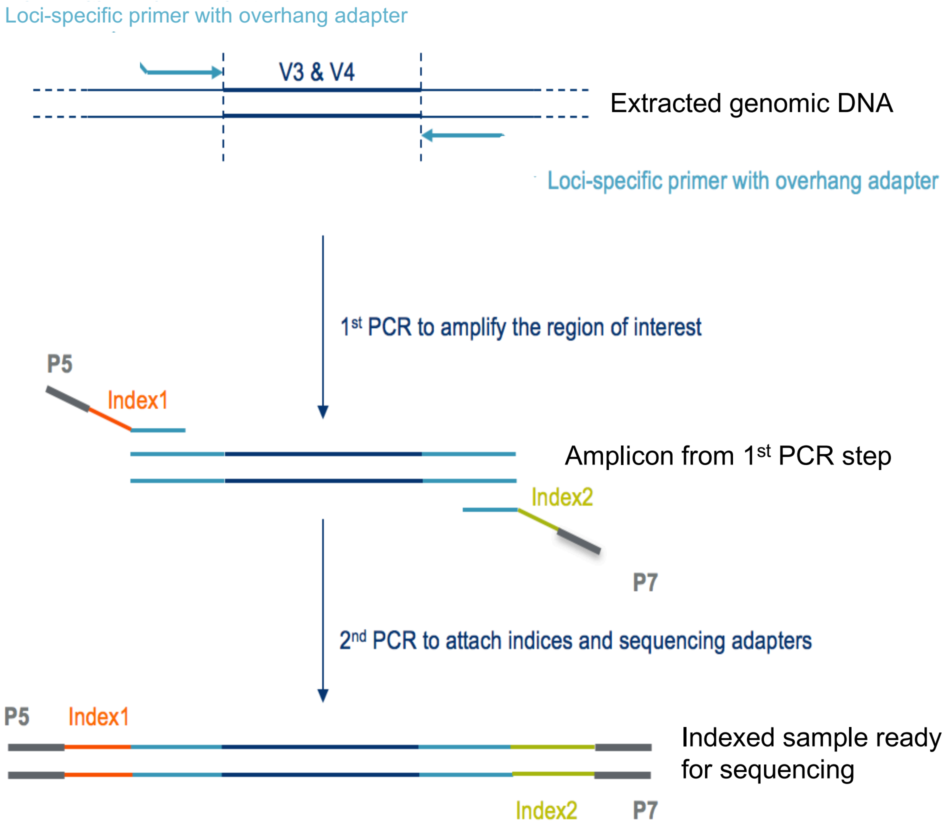
Barcodes are attached to the primers (indexing) to differentiate amplicons from each sample. Afterwards libraries can be been pooled together for sequencing. Distinguishing the produced reads based on the inserted barcodes after sequencing, is called demultiplexing

Primer linkers help introduce a kink / bend in the molecular shape of the primer & provide spacing between the locus-specific primer & the rest of the scaffold

National Institute of
Allergy and
Infectious Diseases

https://www.idtdna.com/pages/education/decoded/article/16s-rrna-indexed-primers-amplify-phylogenic-markers-for-microbiome-sequencing-analysis

Library preparation for high-throughput sequencing of 16S rRNA gene

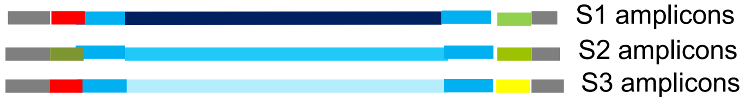

Loci-specific primer with overhang adapter
V3 & V4
Extracted genomic DNA
Loci-specific primer with overhang adapter
1st PCR to amplify the region of interest
P5
Index1
Amplicon from 1st PCR step
Index2
P7
2nd PCR to attach indices and sequencing adapters
P5
Index1
Indexed sample ready for sequencing
Index2
P7

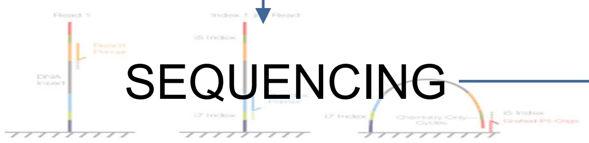
Commonly a 2 step PCR is used in library preparation, to reduce PCR bias

Sample ID	Index 1	Index 2
Sample 1	ATGAATAG	AGGCTATA
Sample 2	TCCGAGA	AGGCTATA
Sample 3	ATGAATAG	ACCTAGCA

Amplicons from each sample are uniquely indexed

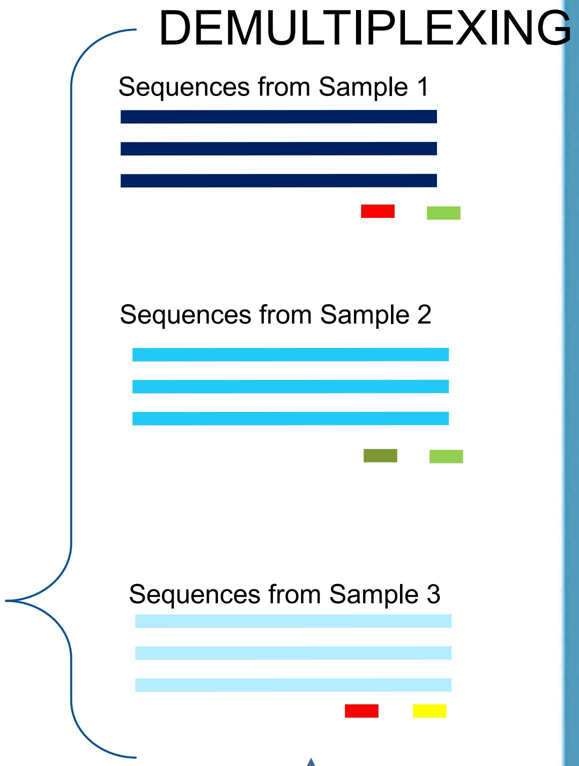
Indexed amplicons from each sample are pooled together at equimolar concentration





SEQUENCING

SEQUENCE DATA



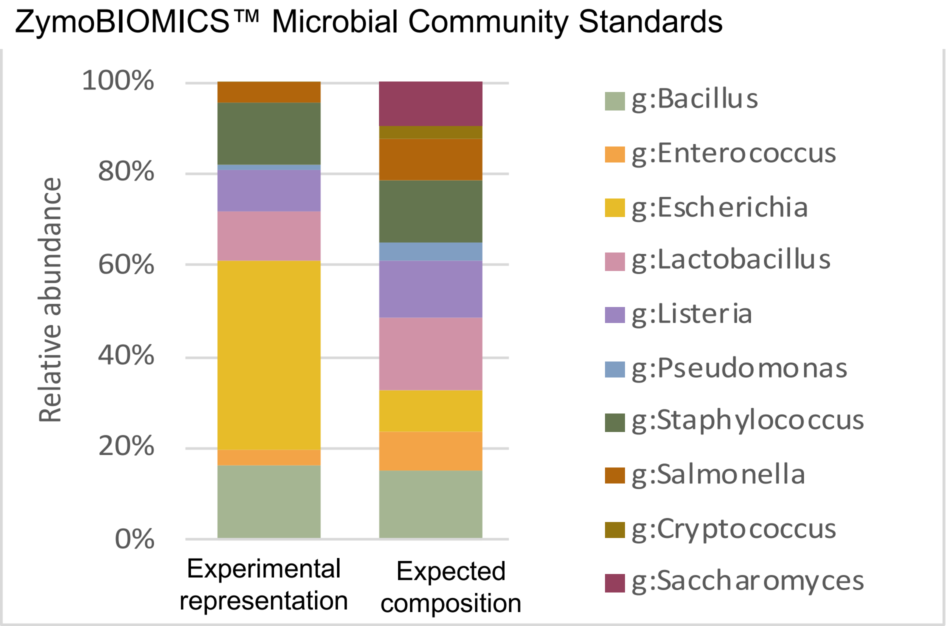
DEMULTIPLEXING

National Institute of Allergy and Infectious Diseases

https://www.idtdna.com/pages/education/decoded/article/16s-rrna-indexed-primers-amplify-phylogenic-markers-for-microbiome-sequencing-analysis

Problems with PCR: inherent biases

- Inherent biases:
  - PCR primer bias: bias introduced from the different primer annealing efficiency to different templates /phylogenic groups
  - PCR efficiency bias: bias introduced from the different PCR rates among different templates/ phylogenic groups
- Resulting in inaccurate representation of template distributions
- PCR bias cannot be avoided in amplification-based studies, but it can (and should!) be:
  - constrained with molecular biology techniques (e.g. 2 step PCR, degenerate bases); or
  - assessed in the downstream analysis, acknowledged and presented in each study!



It is pivotal to include microbial community standards to your experimental samples to assess PCR bias

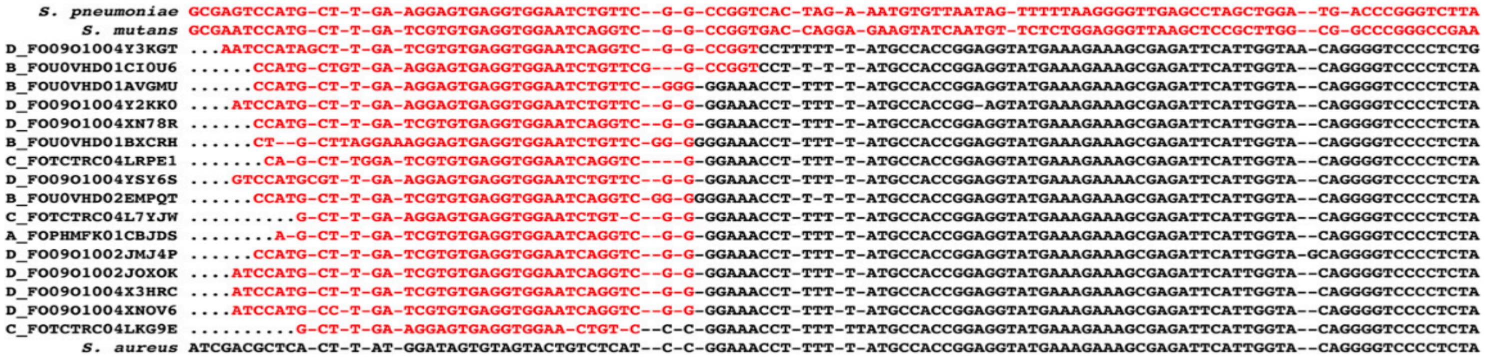
e.g. ATCC microbiome standards, ZymoBiomics Microbial Community standards

National Institute of Allergy and Infectious Diseases

Acinas *et al*. 2005. Appl Environ Microbiol 7(12). doi: 10.1128/AEM.71.12.8966-8969.2005

Problems with PCR: Chimera formation

- Chimeras:
  - Artifacts of PCR amplification, sequencing or read merging
  - Hybrid products between multiple parent sequences
- Inflate apparent community organismal diversity, by suggesting presence of non-existent organisms
- Chimeric formation is more common between more closely related organisms



```
   S. pneumoniae GCGAGTCCATG-CT-T-GA-AGGAGTGAGGTGGAATCTGTTC--G-G-CCGGTCAC-TAG-A-AATGTGTTAATAG-TTTTTAAGGGGTTGAGCCTAGCTGGA--TG-ACCCGGGTCTTA
       S. mutans GCGAATCCATG-CT-T-GA-TCGTGTGAGGTGGAATCAGGTC--G-G-CCGGTGAC-CAGGA-GAAGTATCAATGT-TCTCTGGAGGGTTAAGCTCCGCTTGG--CG-GCCCGGGCCGAA
D_FO09O1004Y3KGT ...AATCCATAGCT-T-GA-TCGTGTGAGGTGGAATCAGGTC--G-G-CCGGTCCTTTTT-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTAA-CAGGGGTCCCCTCTG
B_FOU0VHD01CI0U6 ......CCATG-CTGT-GA-AGGAGTGAGGTGGAATCTGTTCG---G-CCGGTCCT-T-T-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
B_FOU0VHD01AVGMU ......CCATG-CT-T-GA-AGGAGTGAGGTGGAATCTGTTC--GGG-GGAAACCT-TTT-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
D_FO09O1004Y2KK0 ....ATCCATG-CT-T-GA-AGGAGTGAGGTGGAATCTGTTC--G-G-GGAAACCT-TTT-T-ATGCCACCGG-AGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
D_FO09O1004XN78R ......CCATG-CT-T-GA-TCGTGTGAGGTGGAATCAGGTC--G-G-GGAAACCT-TTT-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
B_FOU0VHD01BXCRH ......CT--G-CTTAGGAAAGGAGTGAGGTGGAATCTGTTC-GG-GGGGAAACCT-TTT-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
C_FOTCTRC04LRPE1 .......CA-G-CT-TGGA-TCGTGTGAGGTGGAATCAGGTC----G-GGAAACCT-TTT-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
D_FO09O1004YSY6S ....GTCCATGCGT-T-GA-AGGAGTGAGGTGGAATCTGTTC--G-G-GGAAACCT-TTT-T-ATGCCACCGGAGGTATGAAAGAAAACGAGATTCATTGGTA--CAGGGGTCCCCTCTA
B_FOU0VHD02EMPQT ......CCATG-CT-T-GA-TCGTGTGAGGTGGAATCAGGTC-GG-GGGGAAACCT-T-T-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
C_FOTCTRC04L7YJW ..........G-CT-T-GA-AGGAGTGAGGTGGAATCTGT-C--G-G-GGAAACCT-TTT-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
A_FOPHMFK01CBJDS ........A-G-CT-T-GA-TCGTGTGAGGTGGAATCAGGTC--G-G-GGAAACCT-TTT-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
D_FO09O1002JMJ4P ......CCATG-CT-T-GA-TCGTGTGAGGTGGAATCAGGTC--G-G-GGAAACCT-TTT-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA-GCAGGGGTCCCCTCTA
D_FO09O1002JOXOK ....ATCCATG-CT-T-GA-TCGTGTGAGGTGGAATCAGGTC--G-G-GGAAACCT-TTT-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
D_FO09O1004X3HRC ....ATCCATG-CT-T-GA-TCGTGTGAGGTGGAATCAGGTC--G-G-GGAAACCT-TTT-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
D_FO09O1004XNOV6 ....ATCCATG-CC-T-GA-TCGTGTGAGGTGGAATCAGGTC--G-G-GGAAACCT-TTT-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
C_FOTCTRC04LKG9E ..........G-CT-T-GA-AGGAGTGAGGTGGAA-CTGT-C--C-C-GGAAACCT-TTT-TTATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
       S. aureus ATCGACGCTCA-CT-T-AT-GGATAGTGTAGTACTGTCTCAT--C-C-GGAAACCT-TTT-T-ATGCCACCGGAGGTATGAAAGAAAGCGAGATTCATTGGTA--CAGGGGTCCCCTCTA
```

Figure 4. Alignment of sequences corresponding to chimeras between *Streptococcus* and *Staphylococcus* 16S rRNA genes. Only columns from the NAST multiple alignment containing nonidentical nucleotides between the reference sequences (*top* and *bottom*) are shown. Nucleotides matching *Streptococcus* sequences are colored red. Sequence prefixes correspond to the four experimental replicates A–D.

PCR

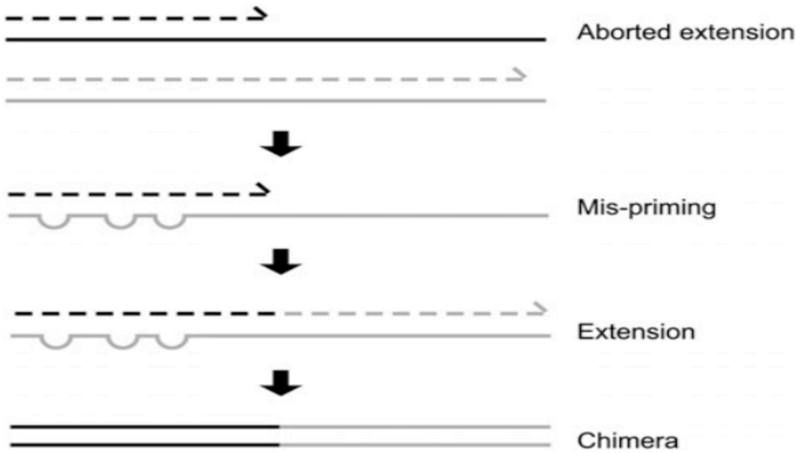


Figure 1. Formation of chimeric sequences during PCR. An aborted extension product from an earlier cycle of PCR can function as a primer in a subsequent PCR cycle. If this aborted extension product anneals to and primes DNA synthesis from an improper template, a chimeric molecule is formed.

Computational tools are used to identify and remove chimeric sequences

Haas (2011), Genome Res 21:494

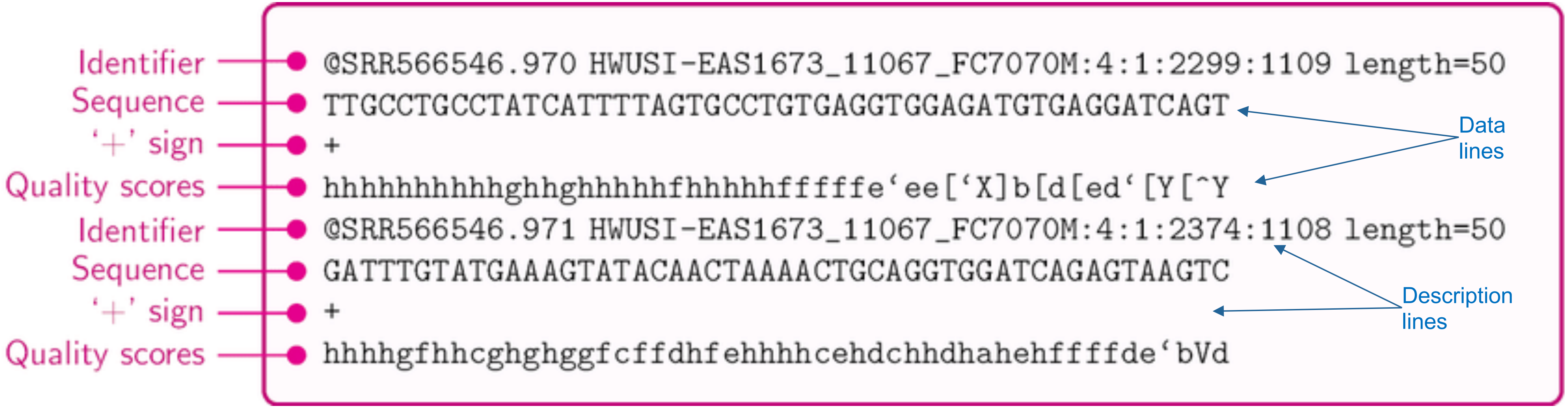
Raw sequences: FastQ file structure

Output files provided from sequencer (normally): Fastq files in archived format

Sample1…_**R1**_...fastq.gz (Forward Reads)
Sample1…_**R2**_...fastq.gz (Reverse Reads)
Sample1…_**R0**_...fastq.gz (Undetermined)

What you will mostly need

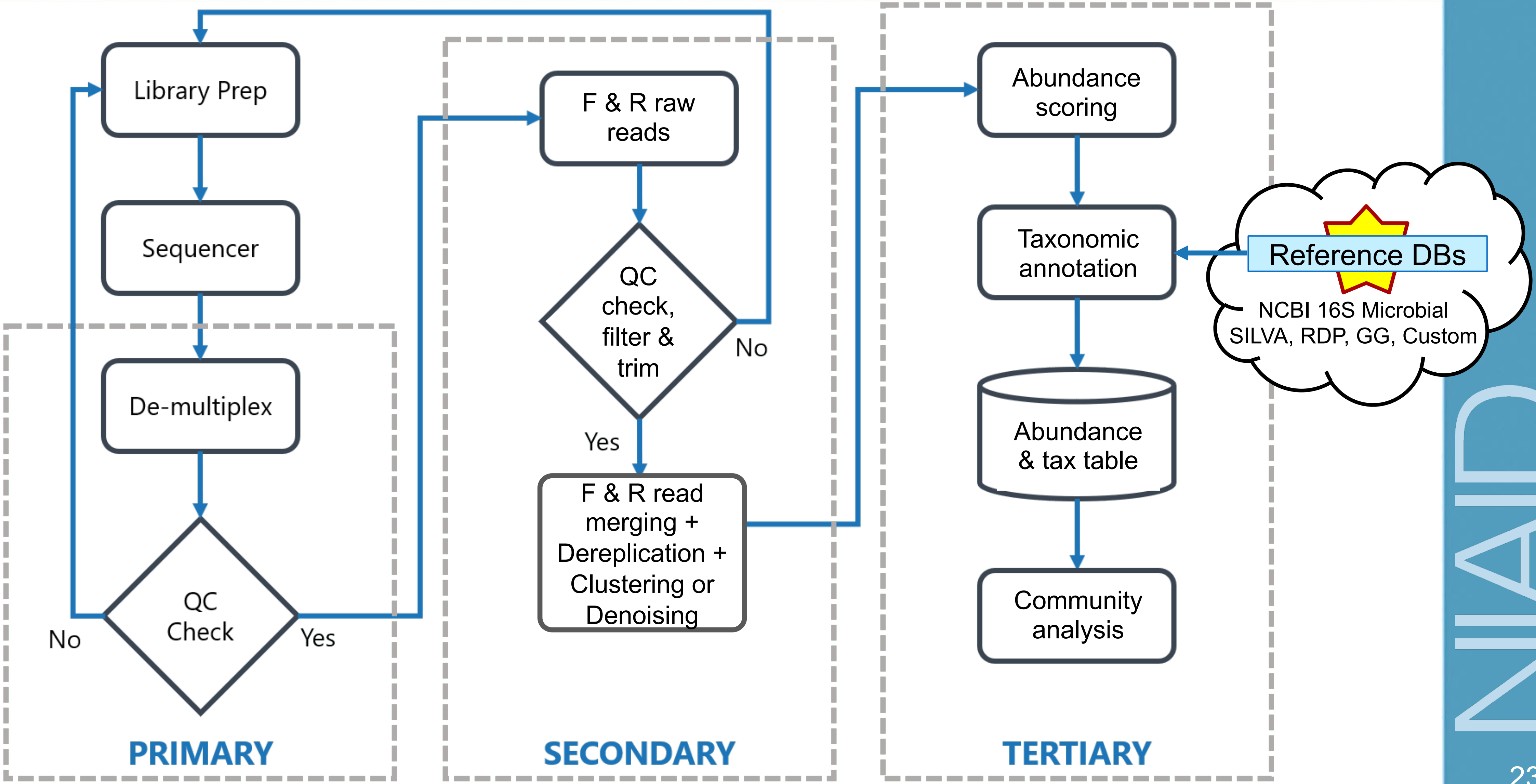
A text-based format for both sequence and associated *phred quality scores*, developed by Sanger Institute



```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhhghhghhhhhfhhhhhfffffe`ee[`X]b[d[ed`[Y[^Y
@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde`bVd
```

National Institute of Allergy and Infectious Diseases

General Bioinformatics workflow
(metataxonomics)



Library Prep

Sequencer

De-multiplex

QC Check

No

Yes

PRIMARY

F & R raw reads

QC check, filter & trim

No

Yes

F & R read merging + Dereplication + Clustering or Denoising

SECONDARY

Abundance scoring

Taxonomic annotation

Abundance & tax table

Community analysis

TERTIARY

Reference DBs

NCBI 16S Microbial
SILVA, RDP, GG, Custom

Processing strategies

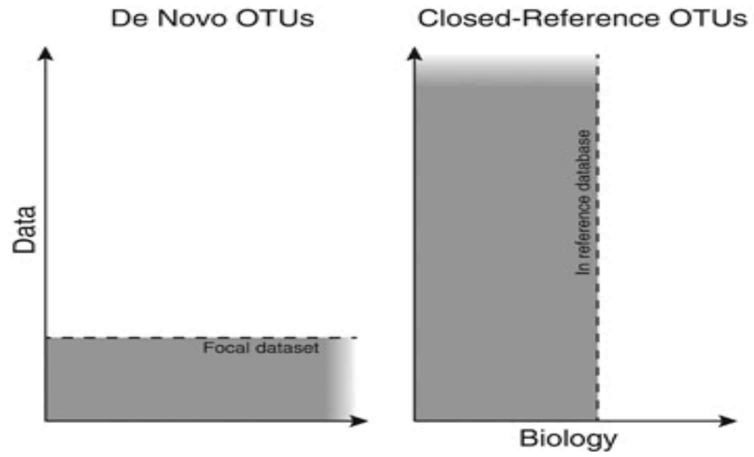
- Cluster-based strategy: OTU formation
  - OTU: Operational taxonomic unit
  - A sequence (usually the longest) representing a cluster of similar sequences
  - recently depreciated for carrying over too many errors
- Denoising strategy: ASV formation
  - ASV: Amplicon Sequence Variants
  - A sequence (error-corrected) representing an amplicon variant
  - common practice / more correct representation

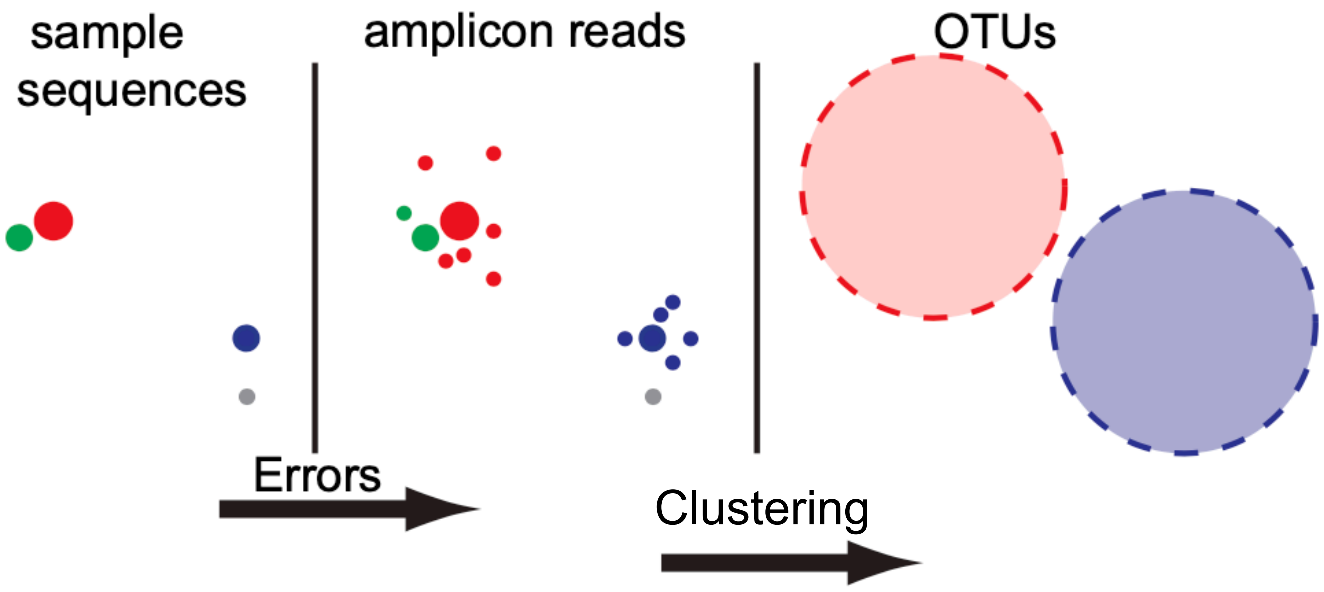
National Institute of Allergy and Infectious Diseases

Clustering (OTU formation)

OTUs = Operational Taxonomic Units
Cluster representative sequence

- Created by clustering of all sequences, based on a fixed similarity threshold (97% seq similarity). A **representative read** is chosen (usually the longest one within the cluster) to represent the whole cluster and its abundance is determined by the number of reads in its cluster.
- This is a depreciated method because: Lingering sequencing errors artificially inflate community diversity
- More stringent downstream filtering is required, loosing information about rarer species

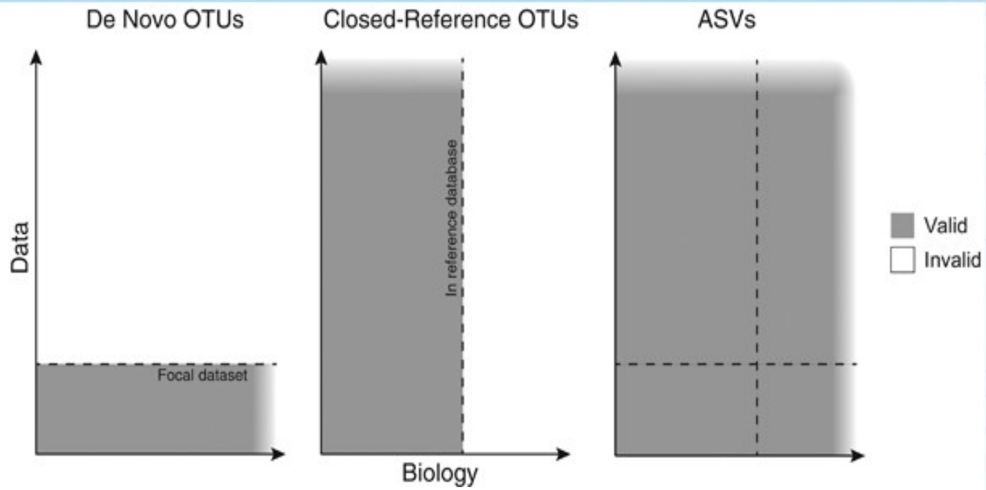


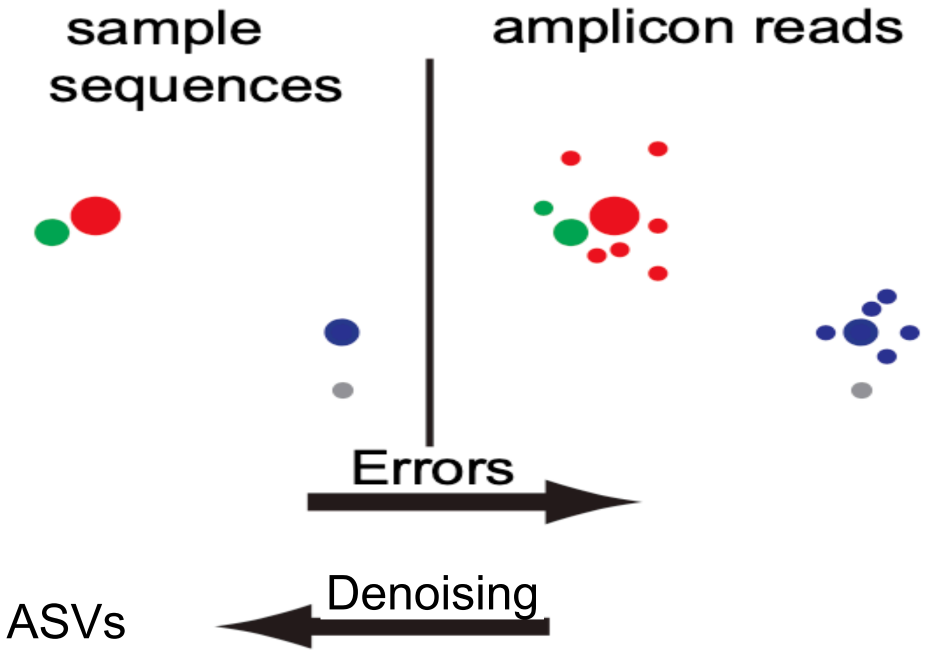


Denoising (ASVs)

ASVs = Amplicon Sequence Variant

- Using nucleotide identity and quality, an error model is created individually for F & R reads, to identify artificial variations. These models are then used to correct the sequence errors, prior to read merging and abundance scoring.
- The amplicon sequence variants (ASVs), provide a more accurate and fine-scale resolution into the *real* diversity of the amplicons, than any clustering algorithm.
- The error model may vary from one sequencing run to another (batch effect)
- Algorithm is reference-free and works on any genetic locus, highly similar sequences (amplicons of the same locus).





sample sequences

amplicon reads

Errors

Denoising

ASVs

National Institute of
Allergy and
Infectious Diseases

Callahan *et al*. 2017 *ISME J* **11**, 2639–2643. https://doi.org/10.1038/ismej.2017.119

Denoising (ASVs) VS Clustering (OTUs)

```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
     |                 |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

$$p(r|s) = \prod_{i=1}^{L} p(r(i)|s(i), q_r(i), Z)$$

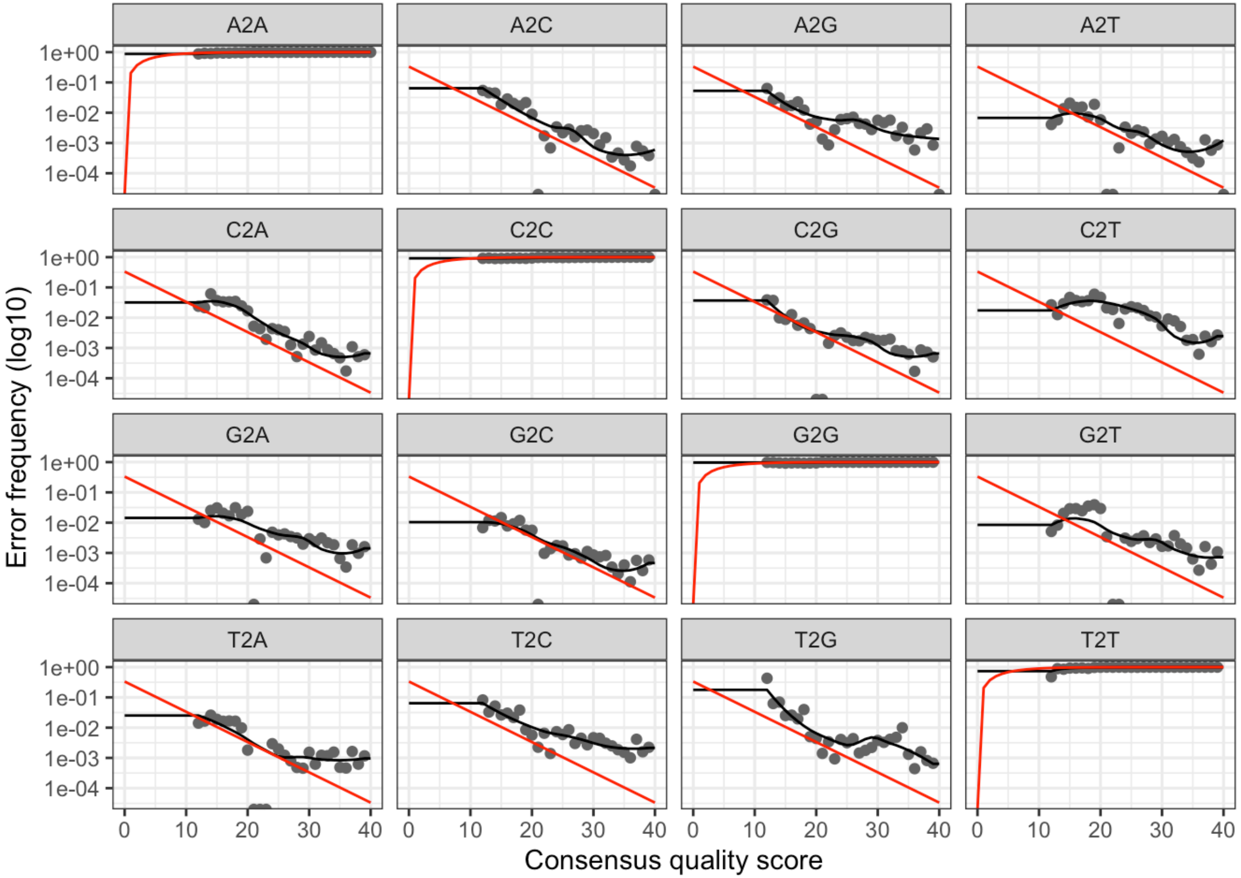
Error rates depend on....

- Substitution (eg. A->C)
- Quality score (eg. Q=30)
- Batch effect (eg. run)

Using more data!

In a good model, the observed error rates (black line) will decrease with increase of quality score (the x axis), keeping the trend of the expected error rates (red line)

Denoising error model



National Institute of Allergy and Infectious Diseases

NIAID

Callahan *et al*. 2017 *ISME J* **11,** 2639–2643. https://doi.org/10.1038/ismej.2017.119

Bioinformatics DADA2 workflow

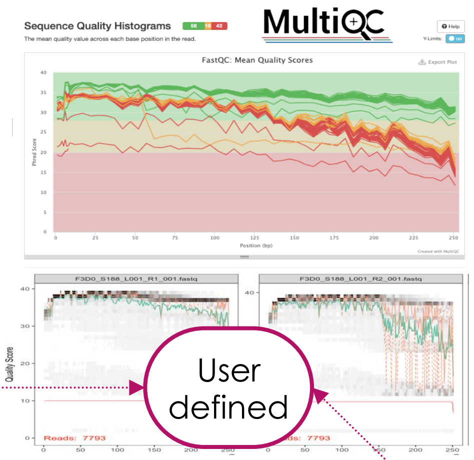
Amplicon Sequencing. Exactly. Version 1.14

Raw Reads

BBDuk

Primer trim

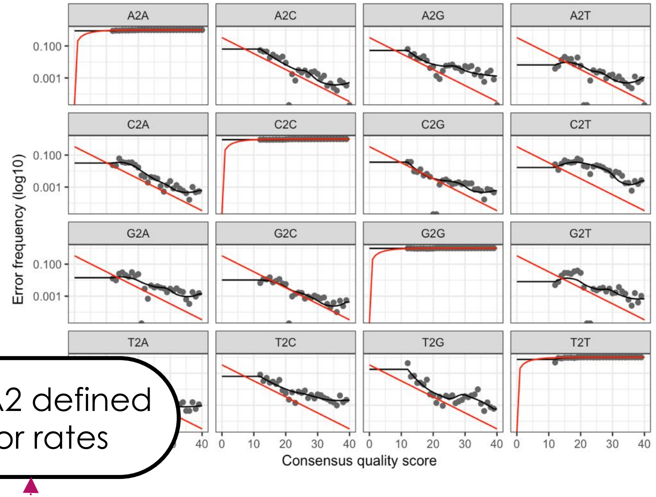
Quality Control



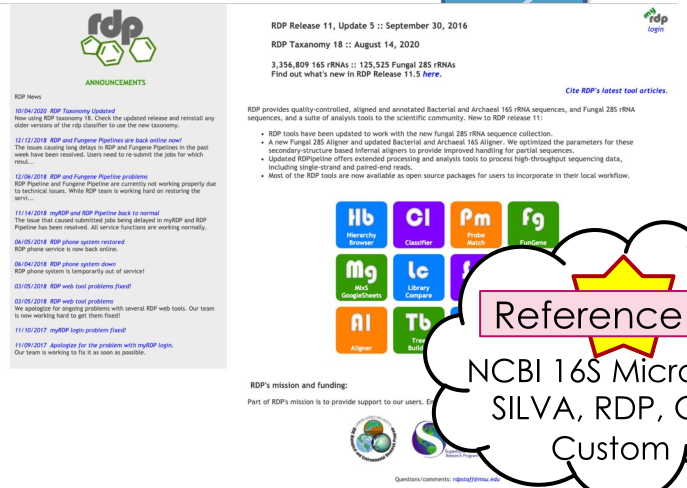
Sequence Quality Histograms 56 16 42

The mean quality value across each base position in the read.

User defined



DADA2 defined error rates



ANNOUNCEMENTS

RDP News

10/04/2020 RDP Taxonomy Updated
Now using RDP taxonomy 18. Check the updated release and reinstall any older versions of the rdp classifier to use the new taxonomy.

12/12/2018 RDP and Fungene Pipelines are back online now!
The issues causing long delays in RDP and Fungene Pipelines in the past week have been resolved. Users need to re-submit the jobs for which resul...

12/06/2018 RDP and Fungene Pipeline problems
RDP Pipeline and Fungene Pipeline are currently not working properly due to technical issues. While RDP team is working hard on restoring the servi...

11/14/2018 myRDP and RDP Pipeline back to normal
The issue that caused submitted jobs being delayed in myRDP and RDP Pipeline has been resolved. All service functions are working normally.

06/05/2018 RDP phone system restored
RDP phone service is now back online.

06/04/2018 RDP phone system down
RDP phone system is temporarily out of service!

03/05/2018 RDP web tool problems fixed!

03/05/2018 RDP web tool problems
We apologize for ongoing problems with several RDP web tools. Our team is now working hard to get them fixed!

11/10/2017 myRDP login problem fixed!

11/09/2017 Apologize for the problem with myRDP login.
Our team is working to fix it as soon as possible.

RDP Release 11, Update 5 :: September 30, 2016

RDP Taxonomy 18 :: August 14, 2020

3,356,809 16S rRNAs :: 125,525 Fungal 28S rRNAs
Find out what's new in RDP Release 11.5 here.

Cite RDP's latest tool articles.

RDP provides quality-controlled, aligned and annotated Bacterial and Archaeal 16S rRNA sequences, and Fungal 28S rRNA sequences, and a suite of analysis tools to the scientific community. New to RDP release 11:

- RDP tools have been updated to work with the new fungal 28S rRNA sequence collection.
- A new Fungal 28S Aligner and updated Bacterial and Archaeal 16S Aligner. We optimized the parameters for these secondary-structure based Infernal aligners to provide improved handling for partial sequences.
- Updated RDPipeline offers extended processing and analysis tools to process high-throughput sequencing data, including single-strand and paired-end reads.
- Most of the RDP tools are now available as open source packages for users to incorporate in their local workflow.

RDP's mission and funding:

Part of RDP's mission is to provide support to our users. En

Questions/comments: rdpstaff@msu.edu

Reference DBs

NCBI 16S Microbial
SILVA, RDP, GG
Custom

Read in fastQ files → Define F & R reads → Quality filtration → Denoise (from error rates) → Dereplicate → Merge F&R reads → Chimera Removal → Taxonomic Annotation

Amplicon Sequencing. Exactly. Version 1.14

Divisive Amplicon Denoising Algorithm

Abundance scoring

ASV formation

ASV taxonomy

Abundance and tax table

Community analysis

Sample ID	FWD.reads	REV.reads	Gene.Region	Sample.Type	Time
Sample 1	Smp1_R1.fastq.gz	Smp1_R2.fastq.gz	V3-V4	Control	0
Sample 2	Smp2_R1.fastq.gz	Smp2_R2.fastq.gz	V3-V4	Experimental	1
Sample 3	Smp3_R1.fastq.gz	Smp3_R2.fastq.gz	V3-V4	Experimental	2

Meta-data table

Dada2 tutorial

Denoising (ASVs) VS Clustering (OTUs)

- Denoising algorithms
  - DADA2 (DADA2,R, QIIME2)
  - UNOISE3 (USEARCH, VSEARCH)
  - Deblur (Deblur, QIIME)
- Clustering algorithms
  - UCLUST (QIIME)
  - UPARSE (USEARCH, VSEARCH)
  - Mothur (mothur)

DADA2

Amplicon Sequencing. Exactly. Version 1.14

Divisive
Amplicon
Denoising
Algorithm

Databases for taxonomic assignment

16S / Bacterial & Archaeal Databases

SILVA

- 2,225,272 full 16S & 18S rRNA gene sequences + guide tree
- Latest release: v138.1 from August 2020

RDP

- 3,356,809 full 16S rRNA gene sequences + 125,525 fungal 28S rRNA gene sequences
- latest release: v18 from August 2020

NCBI's 16S Microbial

- 20,845 full 16S rRNA gene sequences from type strains
- Latest release: Oct 202 (regularly updated)

GreenGenes

- Outdated, latest release 2013

silva
high quality ribosomal RNA databases

eHOMD

rdp

OPEN Tree of Life

NCBI
National Center for Biotechnology Information

GREEN GENES

MicrobiomeDB
A Microbiome Resource
Release 20 17 Dec 2020

Databases for taxonomic assignment

ITS / Fungal Databases

unite community

UNITE Community

- 35,077 ITS gene region sequences
- Latest release: v8.2 from Feb 2020

GlobalFungi

- 145,873,740 ITS sequence variants
- Latest release: v0.9.8 from Jan 2020

FungiDB

- 6,632 ITS gene regions
- Latest release: 50 beta from Dec 2020

FungiDB
Fungal and Oomycetes Genomics Resources
http://FungiDB.org

R-Syst DBs

- A collection of custom databases specific for different phylogenies across kingdoms

PR2

R SYST

silva
high quality ribosomal RNA databases

NIH National Institute of Allergy and Infectious Diseases

18S protozoan databases

SILVA

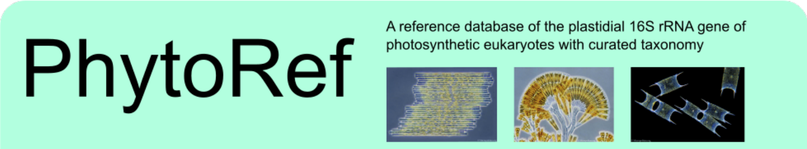
- 2,225,272 full 16S & 18S rRNA gene sequences + guide tree
- Latest release: v138.1 from August 2020

PR2 database

- 184,000 18S rRNA gene region sequences
- Latest release: v14.12.0, from August 2019
- manually curated & metadata available

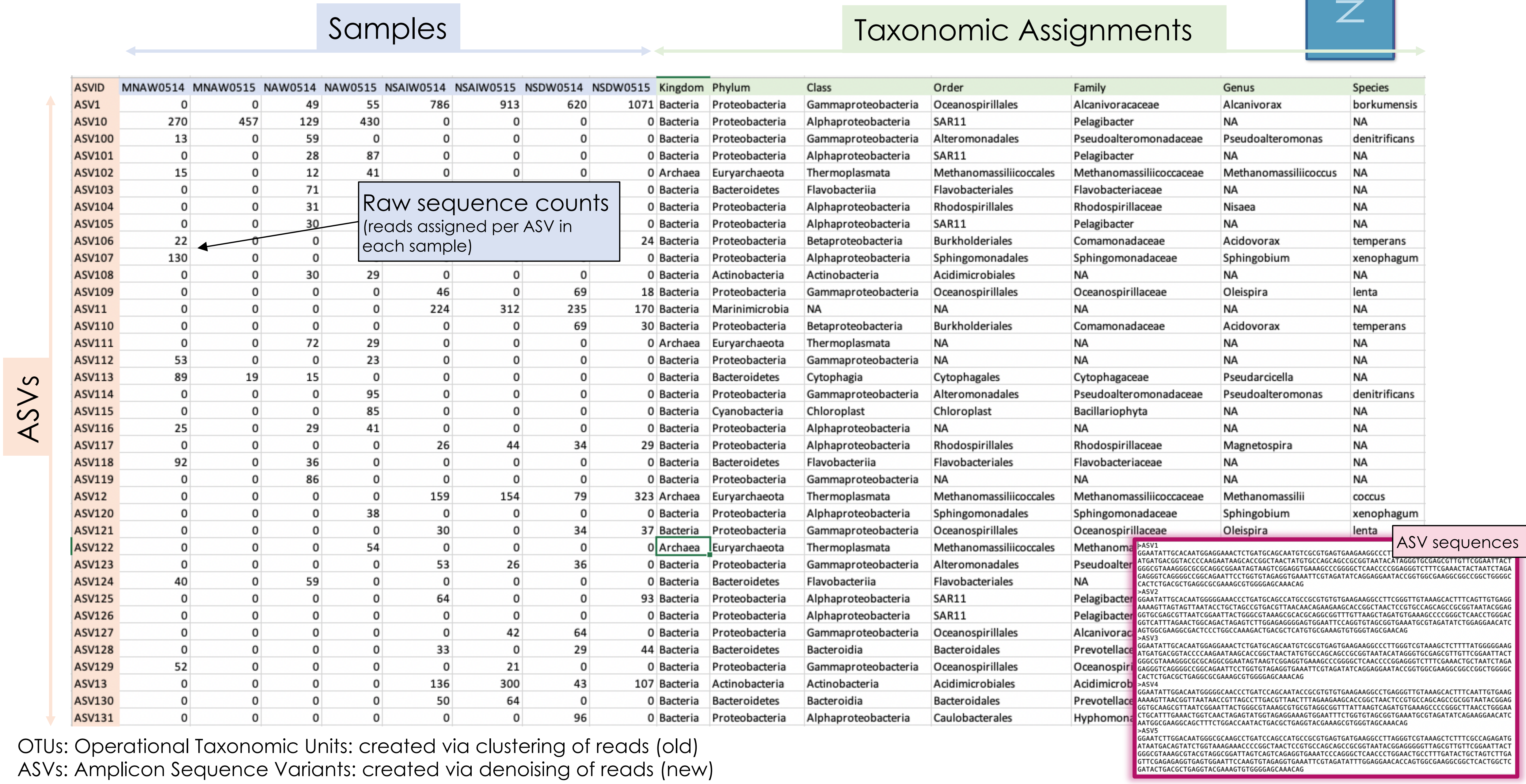
PhytoRef

- Uses *plastidal* 16S rRNA gene to identify photosynthetic micro**eukaryotes**
- 6,490 plastidial 16S rDNA reference sequences
- Latest release: 2015


PhytoRef
A reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy

NIAID

Structure of ASV abundance table



Samples

Taxonomic Assignments

ASVs

Raw sequence counts (reads assigned per ASV in each sample)

ASVID	MNAW0514	MNAW0515	NAW0514	NAW0515	NSAIW0514	NSAIW0515	NSDW0514	NSDW0515	Kingdom	Phylum	Class	Order	Family	Genus	Species
ASV1	0	0	49	55	786	913	620	1071	Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Alcanivoracaceae	Alcanivorax	borkumensis
ASV10	270	457	129	430	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	SAR11	Pelagibacter	NA	NA
ASV100	13	0	59	0	0	0	0	0	Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Pseudoalteromonadaceae	Pseudoalteromonas	denitrificans
ASV101	0	0	28	87	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	SAR11	Pelagibacter	NA	NA
ASV102	15	0	12	41	0	0	0	0	Archaea	Euryarchaeota	Thermoplasmata	Methanomassiliicoccales	Methanomassiliicoccaceae	Methanomassiliicoccus	NA
ASV103	0	0	71					0	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	NA	NA
ASV104	0	0	31					0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Nisaea	NA
ASV105	0	0	30					0	Bacteria	Proteobacteria	Alphaproteobacteria	SAR11	Pelagibacter	NA	NA
ASV106	22	0	0					24	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Acidovorax	temperans
ASV107	130	0	0					0	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	xenophagum
ASV108	0	0	30	29	0	0	0	0	Bacteria	Actinobacteria	Actinobacteria	Acidimicrobiales	NA	NA	NA
ASV109	0	0	0	0	46	0	69	18	Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Oceanospirillaceae	Oleispira	lenta
ASV11	0	0	0	0	224	312	235	170	Bacteria	Marinimicrobia	NA	NA	NA	NA	NA
ASV110	0	0	0	0	0	0	69	30	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Acidovorax	temperans
ASV111	0	0	72	29	0	0	0	0	Archaea	Euryarchaeota	Thermoplasmata	NA	NA	NA	NA
ASV112	53	0	0	23	0	0	0	0	Bacteria	Proteobacteria	Gammaproteobacteria	NA	NA	NA	NA
ASV113	89	19	15	0	0	0	0	0	Bacteria	Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae	Pseudarcicella	NA
ASV114	0	0	0	95	0	0	0	0	Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Pseudoalteromonadaceae	Pseudoalteromonas	denitrificans
ASV115	0	0	0	85	0	0	0	0	Bacteria	Cyanobacteria	Chloroplast	Chloroplast	Bacillariophyta	NA	NA
ASV116	25	0	29	41	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	NA	NA	NA	NA
ASV117	0	0	0	0	26	44	34	29	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Magnetospira	NA
ASV118	92	0	36	0	0	0	0	0	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	NA	NA
ASV119	0	0	86	0	0	0	0	0	Bacteria	Proteobacteria	Gammaproteobacteria	NA	NA	NA	NA
ASV12	0	0	0	0	159	154	79	323	Archaea	Euryarchaeota	Thermoplasmata	Methanomassiliicoccales	Methanomassiliicoccaceae	Methanomassilii	coccus
ASV120	0	0	0	38	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	xenophagum
ASV121	0	0	0	0	30	0	34	37	Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Oceanospirillaceae	Oleispira	lenta
ASV122	0	0	0	54	0	0	0	0	Archaea	Euryarchaeota	Thermoplasmata	Methanomassiliicoccales	Methanoma		
ASV123	0	0	0	0	53	26	36	0	Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Pseudoalter		
ASV124	40	0	59	0	0	0	0	0	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	NA		
ASV125	0	0	0	0	64	0	0	93	Bacteria	Proteobacteria	Alphaproteobacteria	SAR11	Pelagibacter		
ASV126	0	0	0	0	0	0	0	0	Bacteria	Proteobacteria	Alphaproteobacteria	SAR11	Pelagibacter		
ASV127	0	0	0	0	0	42	64	0	Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Alcanivorac		
ASV128	0	0	0	0	33	0	29	44	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellace		
ASV129	52	0	0	0	0	21	0	0	Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Oceanospiri		
ASV13	0	0	0	0	136	300	43	107	Bacteria	Actinobacteria	Actinobacteria	Acidimicrobiales	Acidimicrob		
ASV130	0	0	0	0	50	64	0	0	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellace		
ASV131	0	0	0	0	0	0	96	0	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Hyphomona		

ASV sequences

```
>ASV1
GGAATATTGCACAATGGAGGAAACTCTGATGCAGCAATGTCGCGTGAGTGAAGAAGGCCCT
ATGATGACGGTACCCCAAGAATAAGCACCGGCTAACTATGTGCCAGCAGCCGCGGTAATACATAGGGTGCGAGCGTTGTTCGGAATTACT
GGGCGTAAAGGGCGCGCAGGCGGAATAGTAAGTCGGAGGTGAAAGCCCGGGGCTCAACCCCGGAGGGTCTTTCGAAACTACTAATCTAGA
GAGGGTCAGGGGCCGGCAGAATTCCTGGTGTAGAGGTGAAATTCGTAGATATCAGGAGGAATACCGGTGGCGAAGGCGGCCGGCTGGGGC
CACTCTGACGCTGAGGCGCGAAAGCGTGGGGAGCAAACAG
>ASV2
GGAATATTGCACAATGGGGGAAACCCTGATGCAGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTCAGTTGTGAGG
AAAAGTTAGTAGTTAATACCTGCTAGCCGTGACGTTAACAACAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAG
GGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGCTAGATGTGAAAGCCCCGGGCTCAACCTGGGAC
GGTCATTTAGAACTGGCAGACTAGAGTCTTGGAGAGGGGAGTGGAATTCCAGGTGTAGCGGTGAAATGCGTAGATATCTGGAGGAACATC
AGTGGCGAAGGCGACTCCCTGGCCAAAGACTGACGCTCATGTGCGAAAGTGTGGGTAGCGAACAG
>ASV3
GGAATATTGCACAATGGAGGAAACTCTGATGCAGCAATGTCGCGTGAGTGAAGAAGGCCCTTGGGTCGTAAAGCTCTTTTATGGGGGAAG
ATGATGACGGTACCCCAAGAATAAGCACCGGCTAACTATGTGCCAGCAGCCGCGGTAATACATAGGGTGCGAGCGTTGTTCGGAATTACT
GGGCGTAAAGGGCGCGCAGGCGGAATAGTAAGTCGGAGGTGAAAGCCCGGGGCTCAACCCCGGAGGGTCTTTCGAAACTGCTAATCTAGA
GAGGGTCAGGGGCCGGCAGAATTCCTGGTGTAGAGGTGAAATTCGTAGATATCAGGAGGAATACCGGTGGCGAAGGCGGCCGGCTGGGGC
CACTCTGACGCTGAGGCGCGAAAGCGTGGGGAGCAAACAG
>ASV4
GGAATATTGGACAATGGGGGCAACCCTGATCCAGCAATACCGCGTGTGTGAAGAAGGCCTGAGGGTTGTAAAGCACTTTCAATTGTGAAG
AAAAGTTAACGGTTAATAACCGTTAGCCTTGACGTTAACTTTAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAG
GGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGTAGGCGGTTTATTAAGTCAGATGTGAAAGCCCCGGGCTTAACCTGGGAA
CTGCATTTGAAACTGGTCAACTAGAGTATGGTAGAGGAAAGTGGAATTTCTGGTGTAGCGGTGAAATGCGTAGATATCAGAAGGAACATC
AATGGCGAAGGCAGCTTTCTGGACCAATACTGACGCTGAGGTACGAAAGCGTGGGTAGCAAACAG
>ASV5
GGAATCTTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGAGTGATGAAGGCCTTAGGGTCGTAAAGCTCTTTCGCCAGAGATG
ATAATGACAGTATCTGGTAAAGAAACCCCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGGGTTAGCGTTGTTCGGAATTACT
GGGCGTAAAGCGTACGTAGGCGGATTAGTCAGTCAGAGGTGAAATCCCAGGGCTCAACCCTGGAACTGCCTTTGATACTGCTAGTCTTGA
GTTCGAGAGAGGTGAGTGGAATTCCAAGTGTAGAGGTGAAATTCGTAGATATTTGGAGGAACACCAGTGGCGAAGGCGGCTCACTGGCTC
GATACTGACGCTGAGGTACGAAAGTGTGGGGAGCAAACAG
```

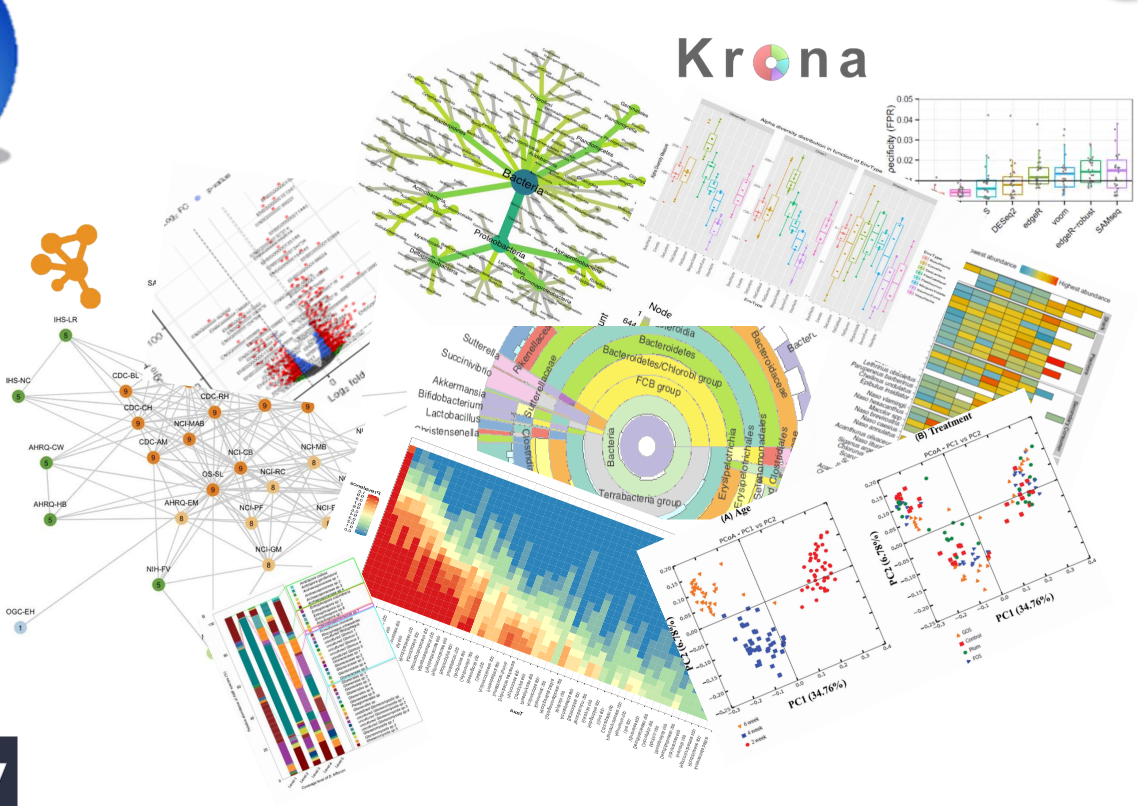
OTUs: Operational Taxonomic Units: created via clustering of reads (old)
ASVs: Amplicon Sequence Variants: created via denoising of reads (new)

Community Analysis

Calypso

R

Krona



MEGAN6
Community Edition

Prism 8
Analyze, graph and present your scientific work.

STAMP

Galaxy
PROJECT

MG-RAST
metagenomics analysis server
version 4.0.3
cite us

PRIMER-E
Multivariate Statistics for Ecologists

Questions addressable by metataxonomics

- What organisms are present in each microbiome and in what proportion? (community structure)
- What is the natural variation of organisms within each microbiome (diversity of organisms)?
- How different is community 1 from community 2 in its composition?
- Which organisms are different and which the same between 2 microbiomes?
- Which microbiome is our organism of interest, more abundant in?
- What is the natural non-variable faction of organisms the microbiome (core microbiome)?
- How is the core microbiome different in community 1 vs community 2?
- How does the diversity of community change depending on factors (e.g. treatment, time)?
- Which organisms responds to factors 1, factor 2, etc.?

Questions **NOT addressable** by 16S amplicon sequencing

- What is the functional capacity of the microbiome as a whole or for each individual organism?
- What is the finer scale (strains-level) and larger scale (cross-kingdom) diversity of a microbiome?
- How does a community adjust to factor 1 in terms of its functionally?

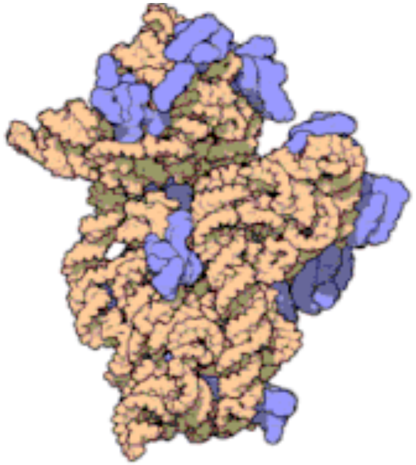
NIH National Institute of Allergy and Infectious Diseases

NIAID

Limitations of 16S rRNA gene

The 16S rRNA gene has become the most sequenced taxonomic marker and the cornerstone for current systematic classification of bacteria and archaea.

Restrictions, caveats, limitations:



- No single variable region captures the variability of the full gene
- Different variable regions have different capacity to differentiate taxonomies
- Due to inherent biases of this method, abundance estimates can be askew
- 16S rRNA gene copy number issues can distort true abundances
  - Gene number can vary from species to species, creating distorted profiles
  - Gene sequence can vary between copies even within the same organism
- Inferring *true* phylogenic relations from a single gene can be risky!
  - Even full length 16S gene cannot absolutely resolve the diversification of closely related organisms (species or strains)
  - Some evolutionarily distant organisms have similar 16S rRNA gene copies, which can cluster closely in a phylogenic tree

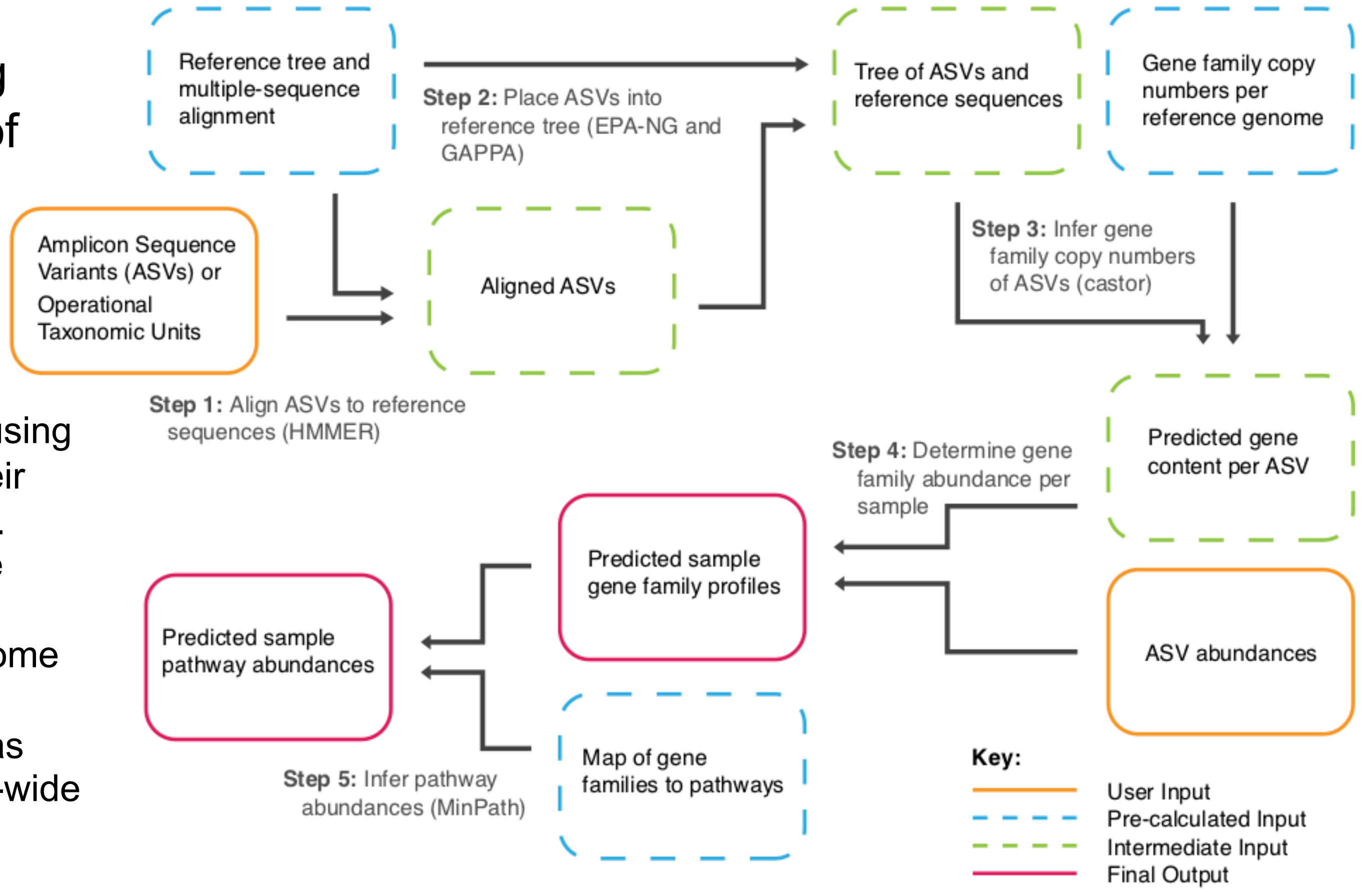
Front Microbiol. Mar 2017; 8: 365. doi: 10.3389/fmicb.2017.00365
Nat Rev Microboil, 2014. https://doi.org/10.1038/nrmicro3330

PICRUSt2

Phylogenic Investigation of Communities by Reconstruction of Unobserved States … 2

An *in silico* approach to **predicting** the functional **potential** of metagenome using marker gene data (16S) and a database of annotated reference genome.

- The genetic content of each sample is *estimated* using the taxonomic identities of the input ASVs, and their most closely related genomes from IMG database.
- Gene abundances are approximated based on the ASV abundances and are corrected for gene copy numbers, again estimates from the reference genome for each taxa.
- The genetic content and abundances (presented as KO & EC numbers) are translated into community-wide pathways abundances, derived from the MetaCyc database for metabolic pathways.



Reference tree and multiple-sequence alignment

Amplicon Sequence Variants (ASVs) or Operational Taxonomic Units

Aligned ASVs

Step 1: Align ASVs to reference sequences (HMMER)

Step 2: Place ASVs into reference tree (EPA-NG and GAPPA)

Tree of ASVs and reference sequences

Gene family copy numbers per reference genome

Step 3: Infer gene family copy numbers of ASVs (castor)

Predicted gene content per ASV

Step 4: Determine gene family abundance per sample

ASV abundances

Predicted sample gene family profiles

Map of gene families to pathways

Step 5: Infer pathway abundances (MinPath)

Predicted sample pathway abundances

Key:

━━	User Input
- - -	Pre-calculated Input
- - -	Intermediate Input
━━	Final Output

National Institute of Allergy and Infectious Diseases

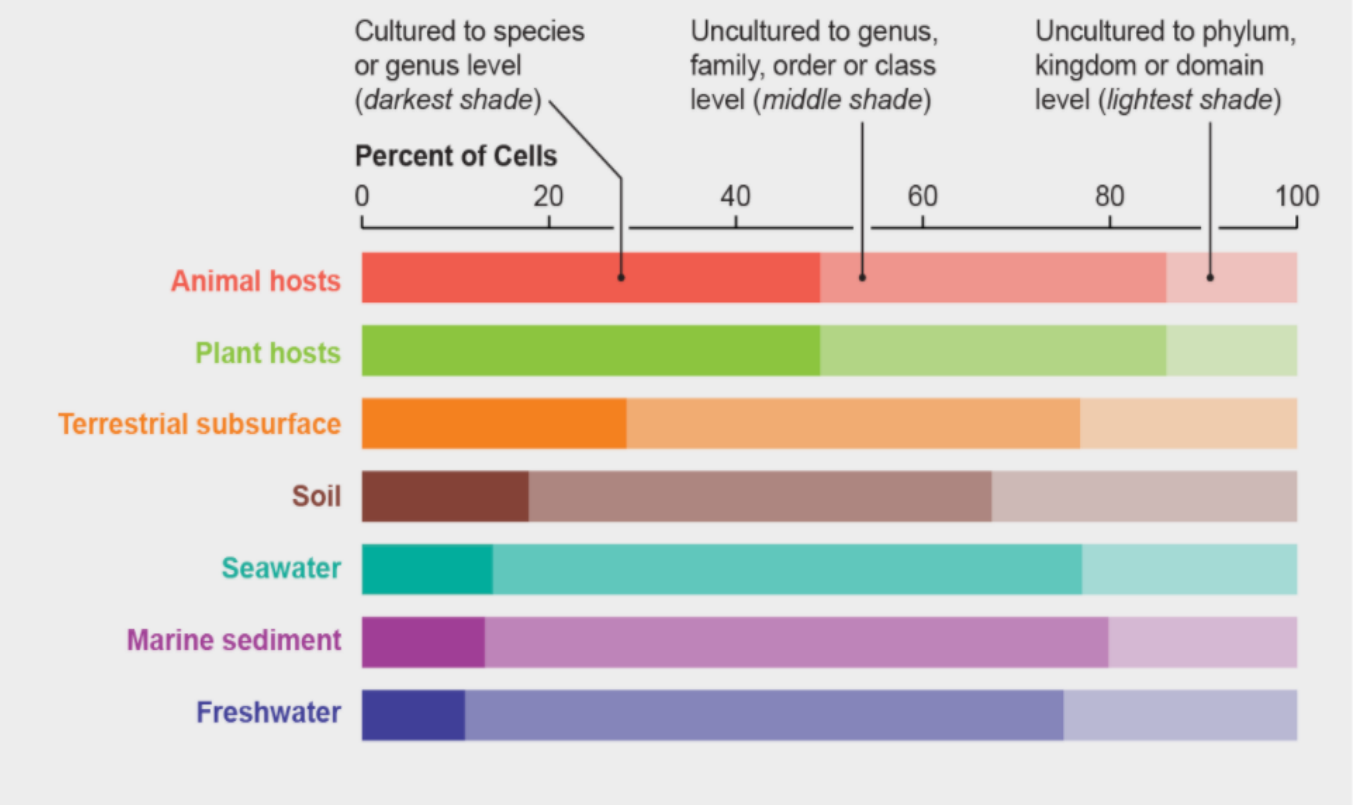
PICRUSt2 tutorial
Douglas *et al.* 2020. *Nat Biotechnol* **38.** https://doi.org/10.1038/s41587-020-0548-6

Data resource explosion

NGS has produced data explosion, causing new opportunities for exploration, but also challenges in the scalability and compatibility of data analysis.

- Opportunities: obtaining insight into the microbial “dark matter” problem without culturing

- Challenges: Vast quantities of available data pose problems in its analytical reproducibility, compatibility and comparability



Credit: Amanda Montañez; Source: “Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes,” by Karen G. Lloyd et al., in *mSystems*, Vol. 3, No. 5; September/October 2018

National Institute of Allergy and Infectious Diseases

https://www.scientificamerican.com/article/most-microbial-species-are-dark-matter/

Dudhagara (2015), Genomics, Proteomics & Bioinformatics. 13:296

Data availability & quality control

- The wide variation in biological questions, coupled with technical variability during sample processing and bioinformatic tools and pipelines for analysis, makes it impossible create a single best protocol for all studies.
- Huge variations are observed between the results of analyses from studies using various DNA extraction methods, 16S primer selection, or bioinformatics pipelines, even when utilized on the same samples.
- the Microbiome Quality Control Project (MBQC) attempts to evaluate and standardize technologies and computational methods for assessing (at least the) human-associated microbial communities.

Sinha et al (2015), Genome Biol. 16:276

Metadata is just as important as the data itself!

- Metadata considered ***critical*** *to data interpretation & reproducibility.*
- Needs to be recorded & provided as accurate and concise as possible

∴ **Community Driven Metadata Standards** are being implemented (e.g. NCBI BioSample db)!

…to promote international standardization of (meta)genome quality **and accompanying metadata** (e.g. vocabulary/ontology, informational fields)

…to promote data discoverability, comparability and reproducibility within and across studies.

Checklists of **Minimum Information about any (x) Sequence (MIxS)** available to implement informational requirements (required metadata) for different types of studies (e.g. host-associated vs soil vs water-associated samples)

- Checklists for metaxtaxonomic (marker gene) studies: **MIMARKS** (**Minimal Information about a Marker Sequence)**
- Checklists for (meta)genomic studies: MIGS & MIMS (**Minimal Information about a (Meta)Genomic Sequence)**

These and other standardization checklists available at: https://gensc.org/mixs/

Thank you

Questions

bioinformatics@niaid.nih.gov

NIH National Institute of Allergy and Infectious Diseases

Bioinformatics and Computational Biosciences Branch
Office of Cyber Infrastructure and Computational Biology

NIAID