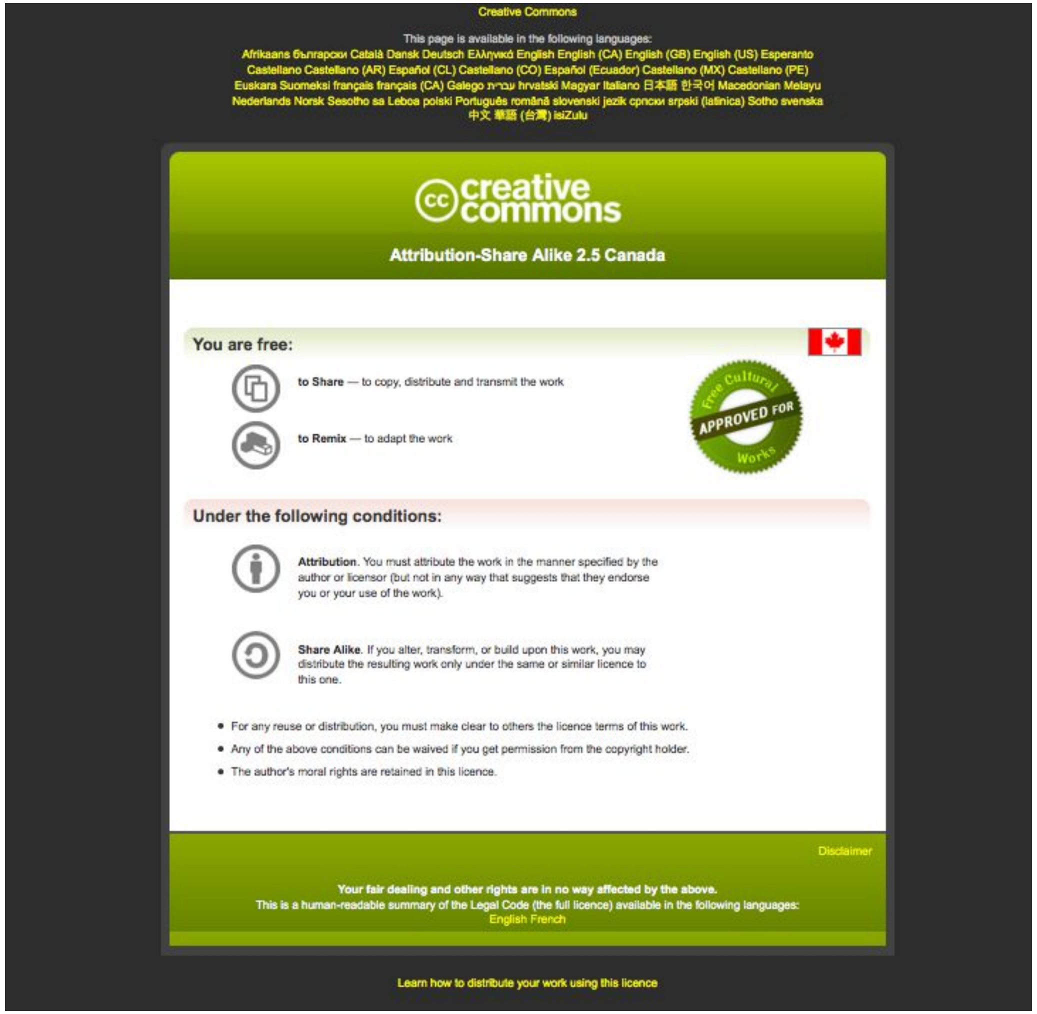
bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io



Creative Commons

This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu

creative commons

Attribution-Share Alike 2.5 Canada

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

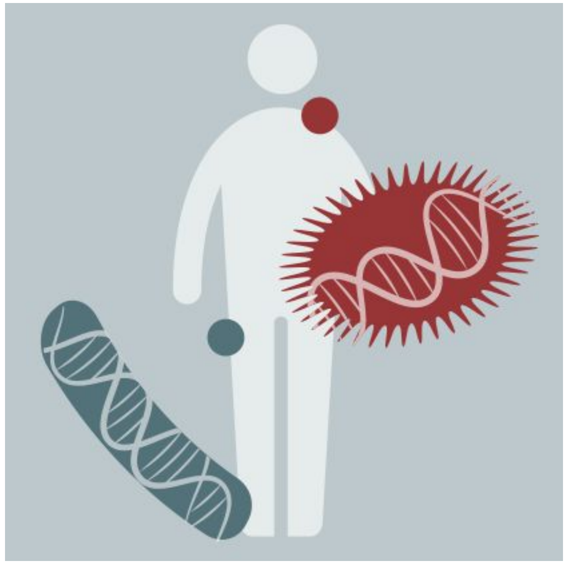
Learn how to distribute your work using this licence

Module 2: Marker Gene Taxonomic Analysis

Morgan Langille

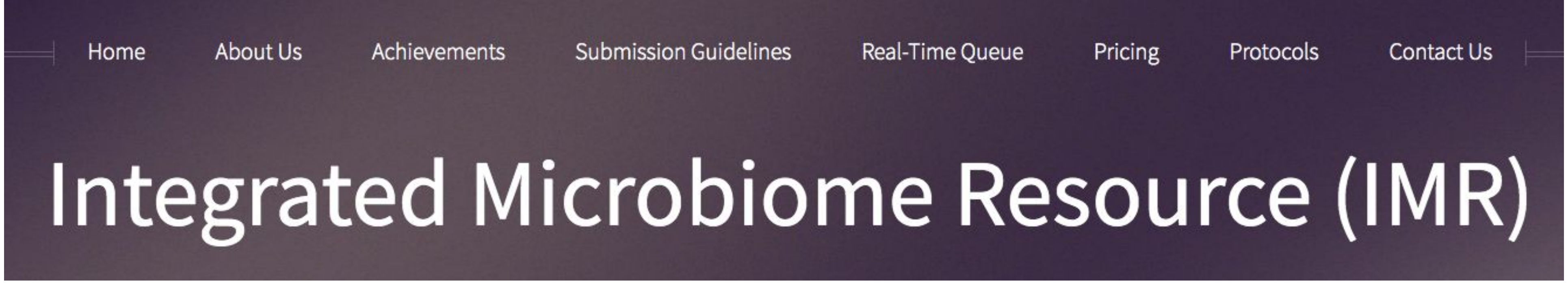
CBW-IMPACTT Microbiome Analysis

July 5-7, 2023



Before we begin…

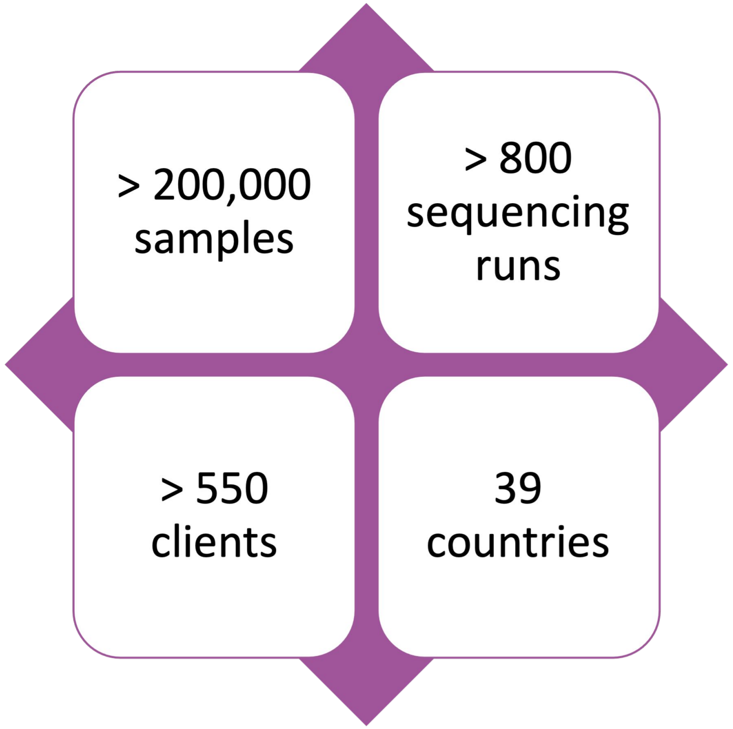
- A bit about research in the Langille Lab.



Integrated Microbiome Resource (IMR)

Sequencing and bioinformatics service for microbiome projects

http://imr.bio

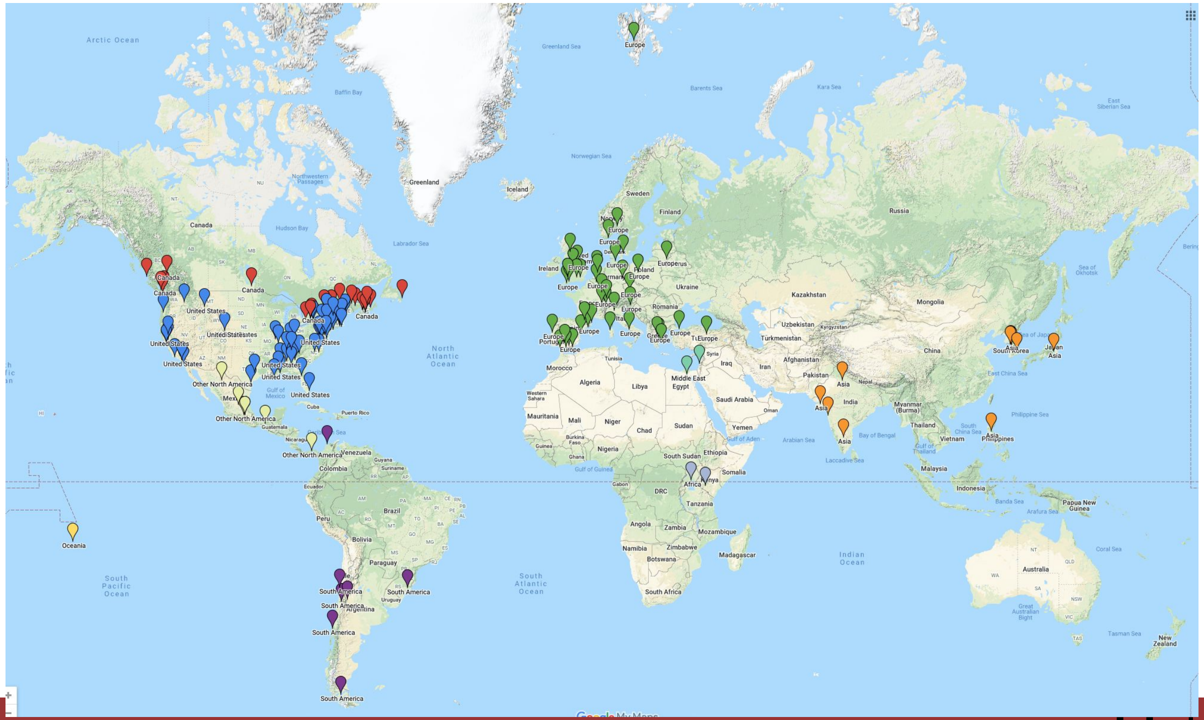


> 200,000 samples

> 800 sequencing runs

> 550 clients

39 countries



Bioinformatic Tool Development

Microbiome Helper

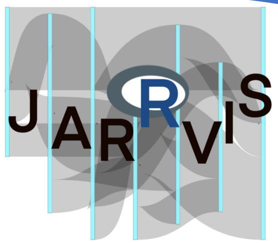
https://github.com/LangilleLab/microbiome_helper/

PICRUSt2

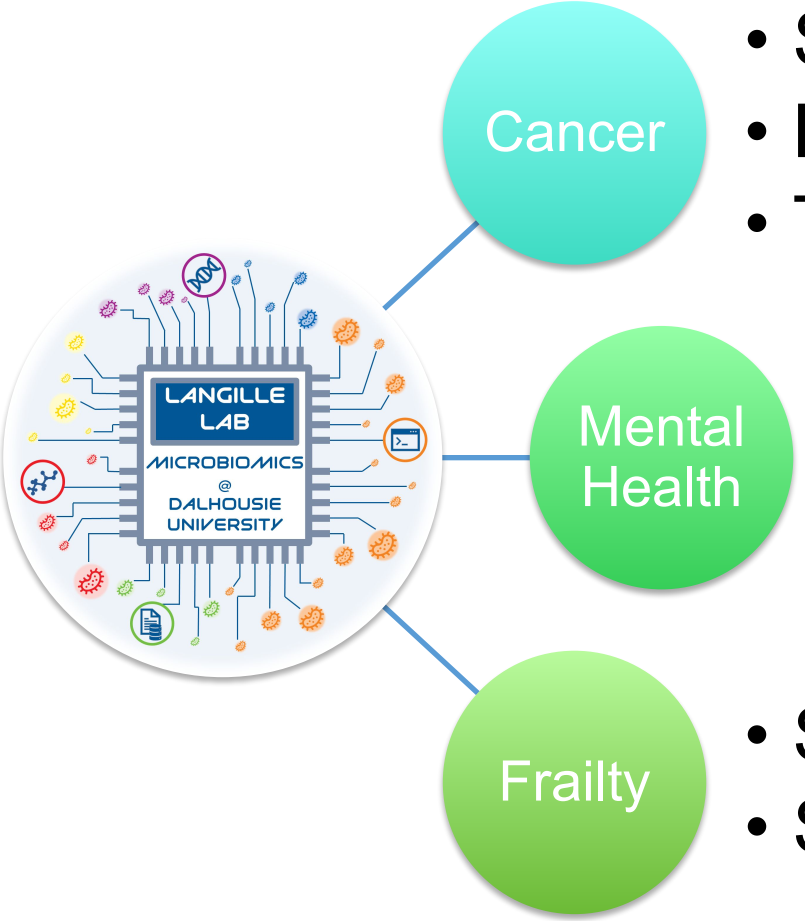
https://github.com/picrust/picrust2/

POMS

https://github.com/gavinmdouglas/POMS



Current Microbiome Research



Cancer

- Saliva
- Blood
- Tumour

Mental Health

- Saliva

Frailty

- Saliva
- Stool

Learning Objectives

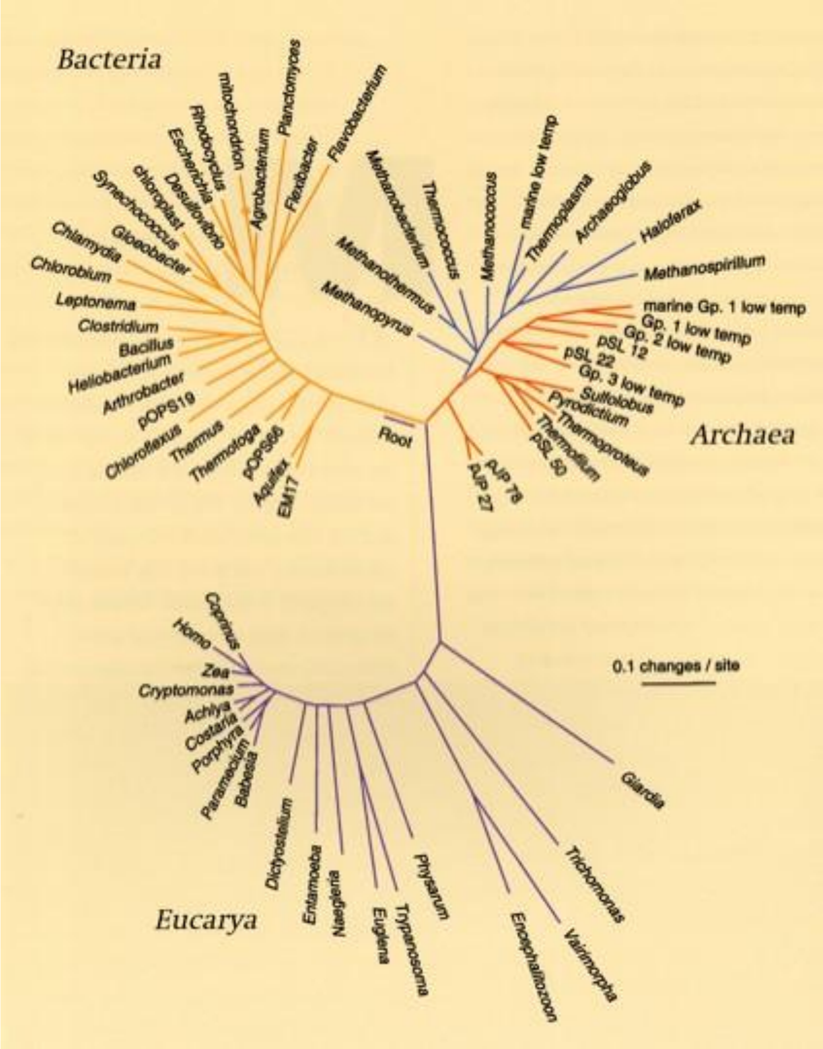
- UNDERSTAND AMPLICON SEQUENCING
- DIFFERENTIATE DIFFERENT AMPLICON TARGETS
- CONTRAST VARIABLE REGIONS WITHIN 16S
- OUTLINE THE MAJOR BIOINFORMATIC STEPS IN 16S DATA ANALYSIS
- UNDERSTAND THE BASIC OUTPUTS FROM PROCESSING 16S DATA

Methods for Studying the Microbiome

- Amplicon-based (16S, 18S, ITS, etc.)
  - “Amplicon” a piece of DNA (or RNA) that is a result of amplification via PCR
  - Sequence a universal gene/barcode
  - Used to identify the of taxa in the sample
    - “Who is there?”
  - Restricted to identifying the organisms targeted by the amplicon primers

rRNAs – the universal phylogenetic markers

- Ribosomal RNAs are present in all living organisms
- rRNAs play critical roles in protein translation
- rRNAs are relatively conserved and thought to be rarely acquired horizontally
- Behave like a molecular clock
  - Useful for phylogenetic analysis
  - Used to build tree-of-life (placing organisms in a single phylogenetic tree)
- 16S rRNA gene most commonly used



Universal phylogenetic tree based on SSU rRNA sequences
-N. Pace, Science, 1997

Other Marker Genes Used

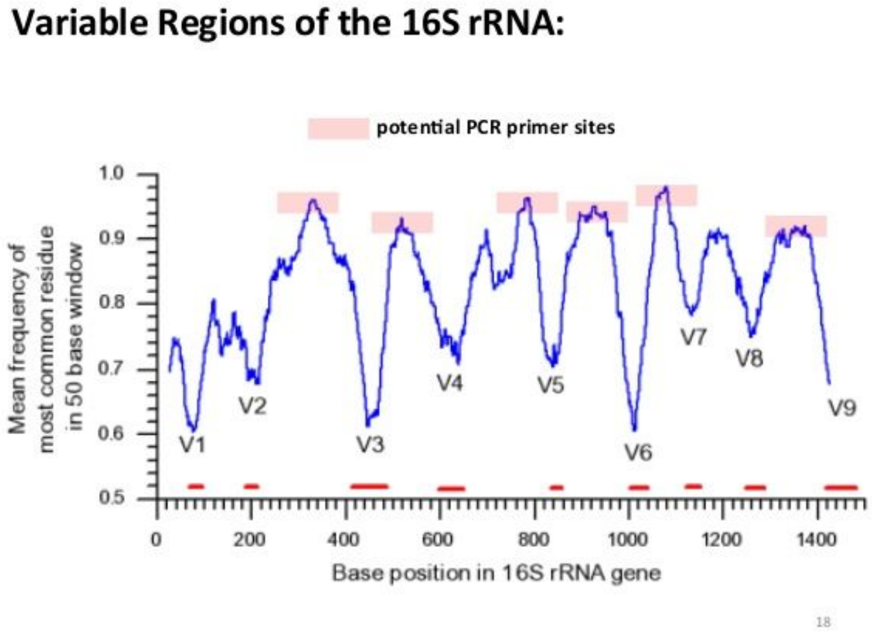
- Bacteria
  - CPN60 (http://www.cpndb.ca/cpnDB/home.php)

- Eukaryotic Organisms (protists, fungi)
  - 18S (http://www.arb-silva.de)
  - ITS (https://unite.ut.ee/)

- Viruses
  - No universal marker!
  - Metagenomics is better approach

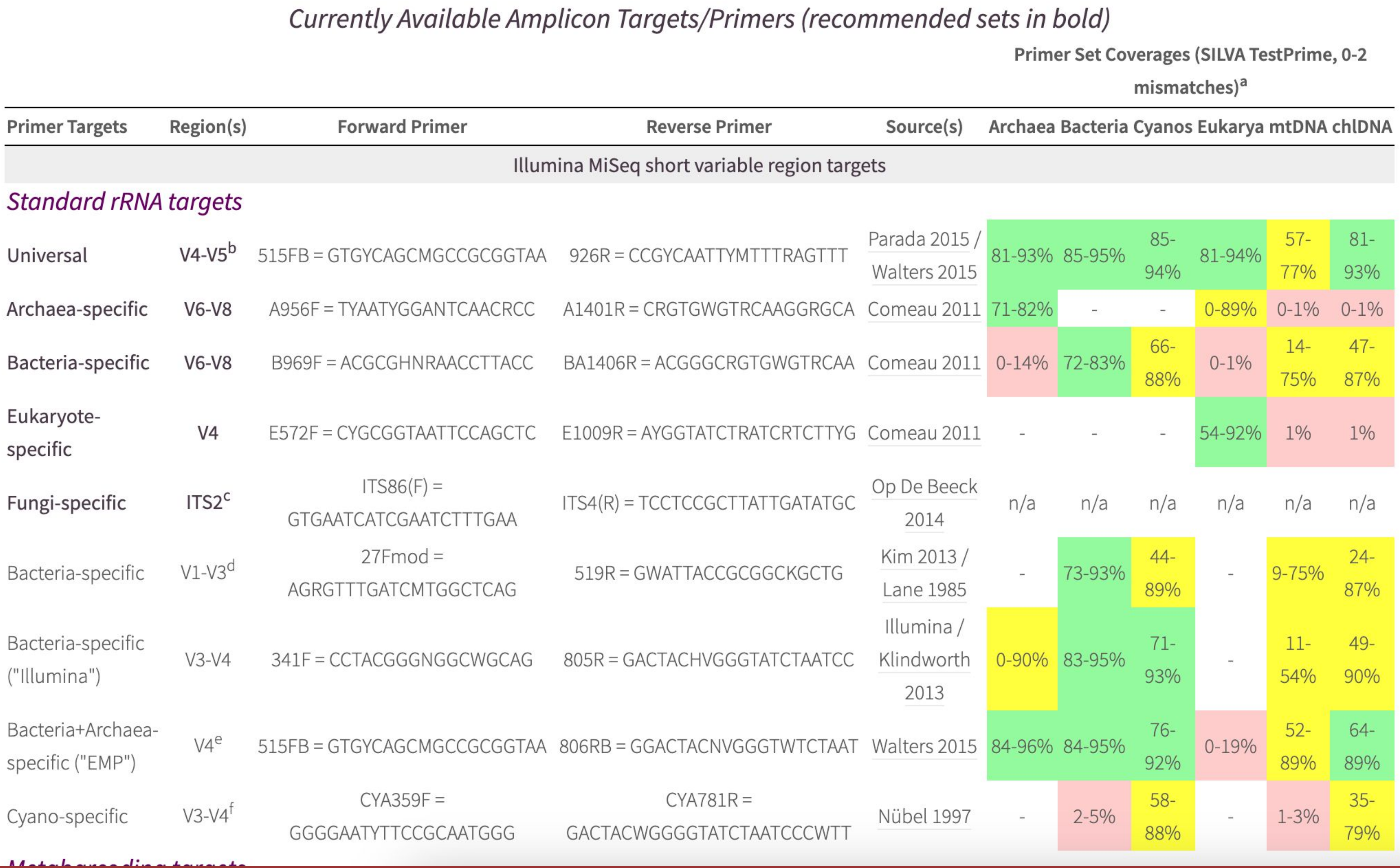
Target Selection and Bias

- 16S rRNA contains nine hypervariable regions (V1-V9)
- Different V regions have different phylogenetic resolutions and bias
- Giving rise to slightly different community composition results



http://themicrobiome.com/media/16S_viewer.cfm

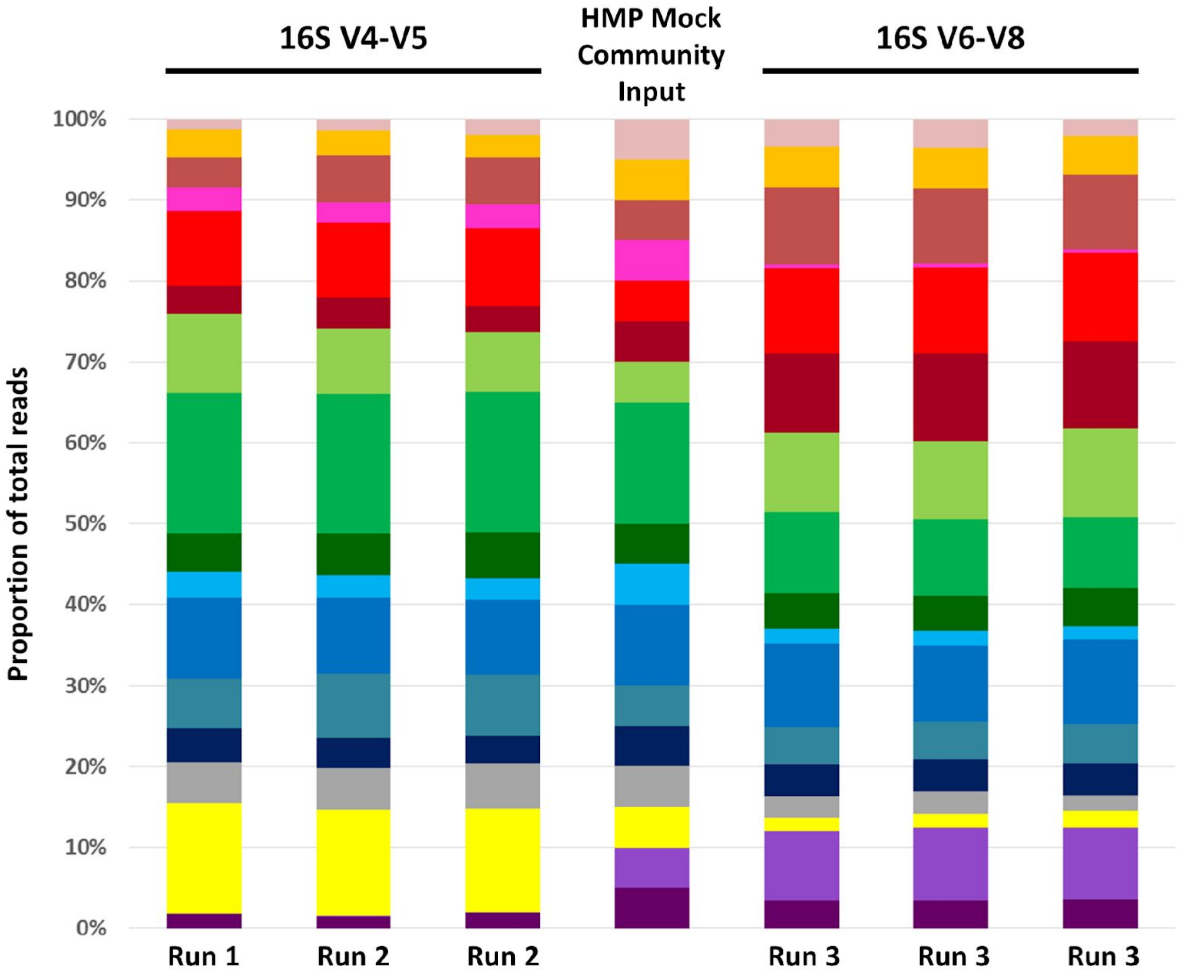
16S Variable Regions



Currently Available Amplicon Targets/Primers (recommended sets in bold)

					Primer Set Coverages (SILVA TestPrime, 0-2 mismatches)[a]					
Primer Targets	**Region(s)**	**Forward Primer**	**Reverse Primer**	**Source(s)**	**Archaea**	**Bacteria**	**Cyanos**	**Eukarya**	**mtDNA**	**chlDNA**
Illumina MiSeq short variable region targets										
Standard rRNA targets										
Universal	**V4-V5**[b]	515FB = GTGYCAGCMGCCGCGGTAA	926R = CCGYCAATTYMTTTRAGTTT	Parada 2015 / Walters 2015	81-93%	85-95%	85-94%	81-94%	57-77%	81-93%
Archaea-specific	**V6-V8**	A956F = TYAATYGGANTCAACRCC	A1401R = CRGTGWGTRCAAGGRGCA	Comeau 2011	71-82%	-	-	0-89%	0-1%	0-1%
Bacteria-specific	**V6-V8**	B969F = ACGCGHNRAACCTTACC	BA1406R = ACGGGCRGTGWGTRCAA	Comeau 2011	0-14%	72-83%	66-88%	0-1%	14-75%	47-87%
Eukaryote-specific	**V4**	E572F = CYGCGGTAATTCCAGCTC	E1009R = AYGGTATCTRATCRTCTTYG	Comeau 2011	-	-	-	54-92%	1%	1%
Fungi-specific	**ITS2**[c]	ITS86(F) = GTGAATCATCGAATCTTTGAA	ITS4(R) = TCCTCCGCTTATTGATATGC	Op De Beeck 2014	n/a	n/a	n/a	n/a	n/a	n/a
Bacteria-specific	V1-V3[d]	27Fmod = AGRGTTTGATCMTGGCTCAG	519R = GWATTACCGCGGCKGCTG	Kim 2013 / Lane 1985	-	73-93%	44-89%	-	9-75%	24-87%
Bacteria-specific ("Illumina")	V3-V4	341F = CCTACGGGNGGCWGCAG	805R = GACTACHVGGGTATCTAATCC	Illumina / Klindworth 2013	0-90%	83-95%	71-93%	-	11-54%	49-90%
Bacteria+Archaea-specific ("EMP")	V4[e]	515FB = GTGYCAGCMGCCGCGGTAA	806RB = GGACTACNVGGGTWTCTAAT	Walters 2015	84-96%	84-95%	76-92%	0-19%	52-89%	64-89%
Cyano-specific	V3-V4[f]	CYA359F = GGGGAATYTTCCGCAATGGG	CYA781R = GACTACWGGGGTATCTAATCCCWTT	Nübel 1997	-	2-5%	58-88%	-	1-3%	35-79%
Metabarcoding targets										

Variable Region Comparison

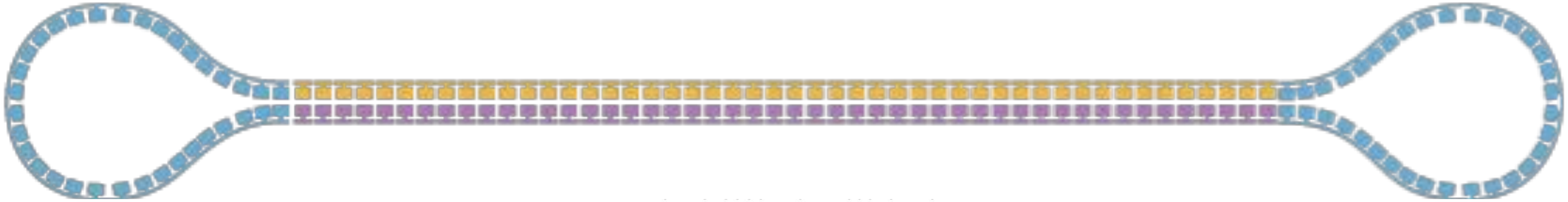


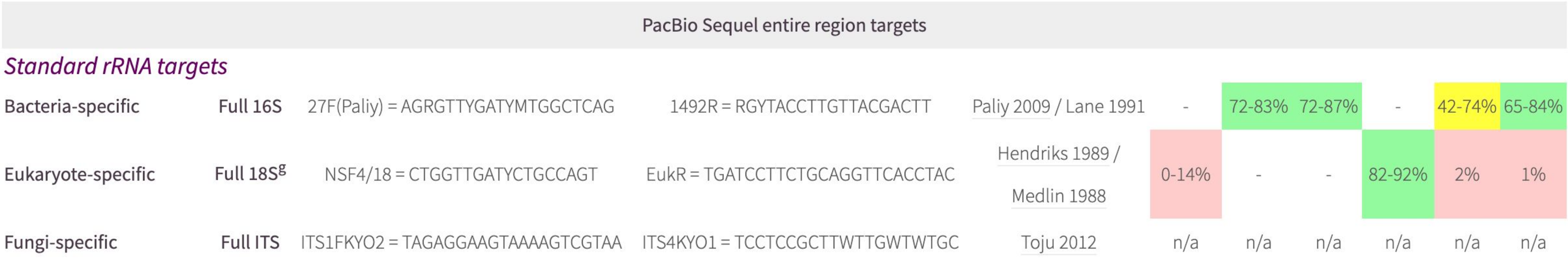
		# of OTUs		
		input	V4-V5	V6-V8
Proteobacteria	■ *Pseudomonas*	1	1	2
	■ *Acinetobacter*	1	2	2
	■ *Escherichia/Shigella*	1	2	3
	■ *Helicobacter*	1	1	1
	■ *Neisseria*	1	2	2
	■ *Rhodobacter*	1	4	3
Firmicutes	■ *Clostridium*	1	1	3
	■ *Streptococcus*	3	6	4
	■ *Lactobacillus*	1	1	4
	■ *Enterococcus*	1	1	1
	■ *Staphylococcus*	2	4	11
	■ *Listeria*	1	2	3
	■ *Bacillus*	1	3	3
	■ *Deinococcus*	1	1	1
	■ *Bacteroides*	1	3	5
Actinobacteria	■ *Propionibacterium*	1	1	4
	■ *Actinomyces*	1	2	2
	Total =	20	37	54

Comeau et al., mSystems, 2017

Full length 16S sequencing

- Pacific Biosystems (PacBio)
    - Long read sequencing (multiple kb)
    - Bad: Accuracy 85-90%
    - Good: Circular Consensus Sequence leads to much higher accuracy (99.0-99.9%) -> “HiFi” reads





PacBio Sequel entire region targets

Standard rRNA targets

Bacteria-specific	Full 16S	27F(Paliy) = AGRGTTYGATYMTGGCTCAG	1492R = RGYTACCTTGTTACGACTT	Paliy 2009 / Lane 1991	-	72-83%	72-87%	-	42-74%	65-84%
Eukaryote-specific	Full 18S^g	NSF4/18 = CTGGTTGATYCTGCCAGT	EukR = TGATCCTTCTGCAGGTTCACCTAC	Hendriks 1989 / Medlin 1988	0-14%	-	-	82-92%	2%	1%
Fungi-specific	Full ITS	ITS1FKYO2 = TAGAGGAAGTAAAAGTCGTAA	ITS4KYO1 = TCCTCCGCTTWTTGWTWTGC	Toju 2012	n/a	n/a	n/a	n/a	n/a	n/a

Available Marker Gene Analysis Platforms

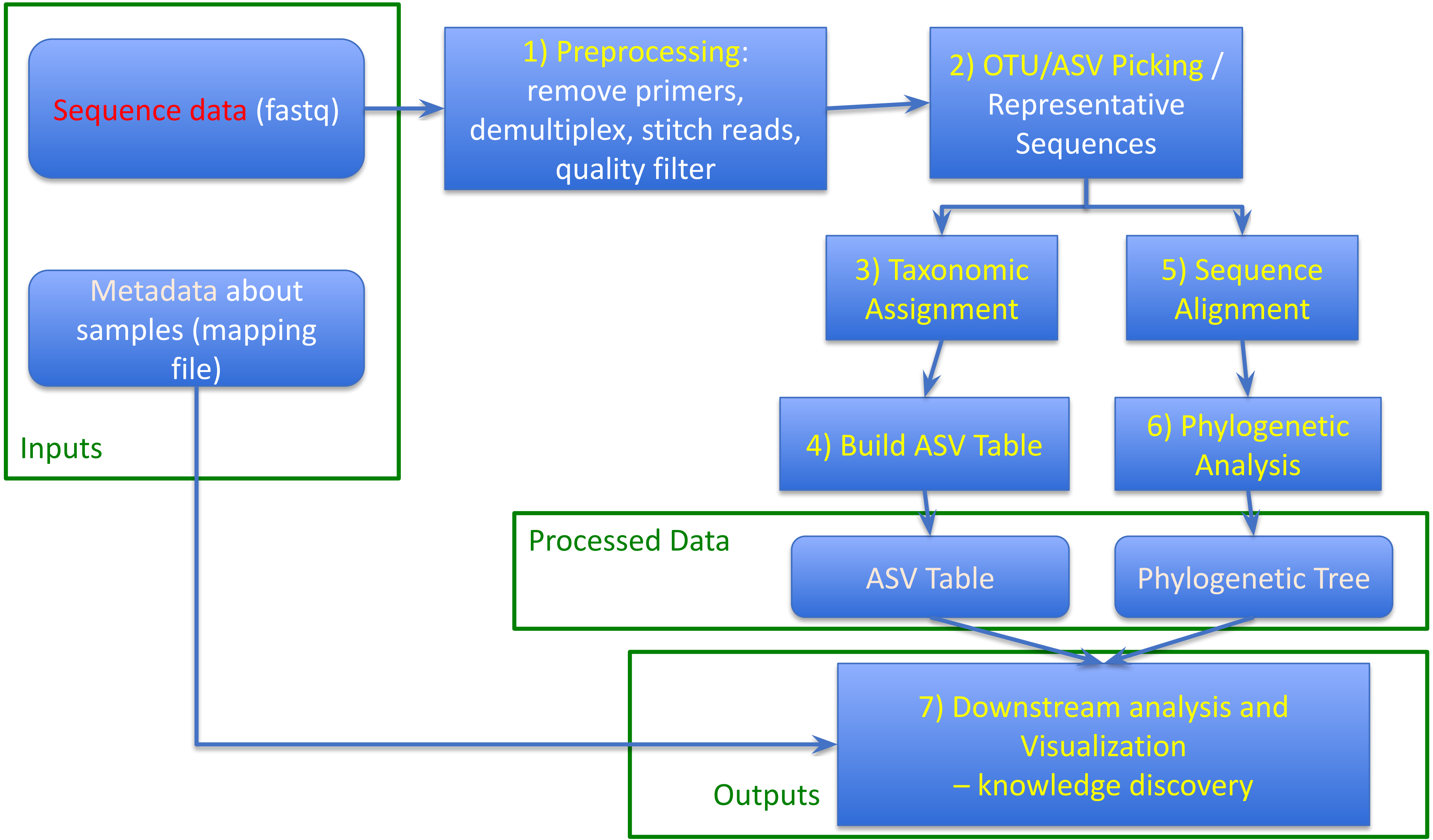
- QIIME (http://qiime.org)

qiime Quantitative Insights Into Microbial Ecology

- Mothur (http://www.mothur.org)

mothur

Overall Bioinformatics Workflow



Sequence data (fastq)

Metadata about samples (mapping file)

Inputs

1) Preprocessing: remove primers, demultiplex, stitch reads, quality filter

2) OTU/ASV Picking / Representative Sequences

3) Taxonomic Assignment

5) Sequence Alignment

4) Build ASV Table

6) Phylogenetic Analysis

Processed Data

ASV Table

Phylogenetic Tree

7) Downstream analysis and Visualization – knowledge discovery

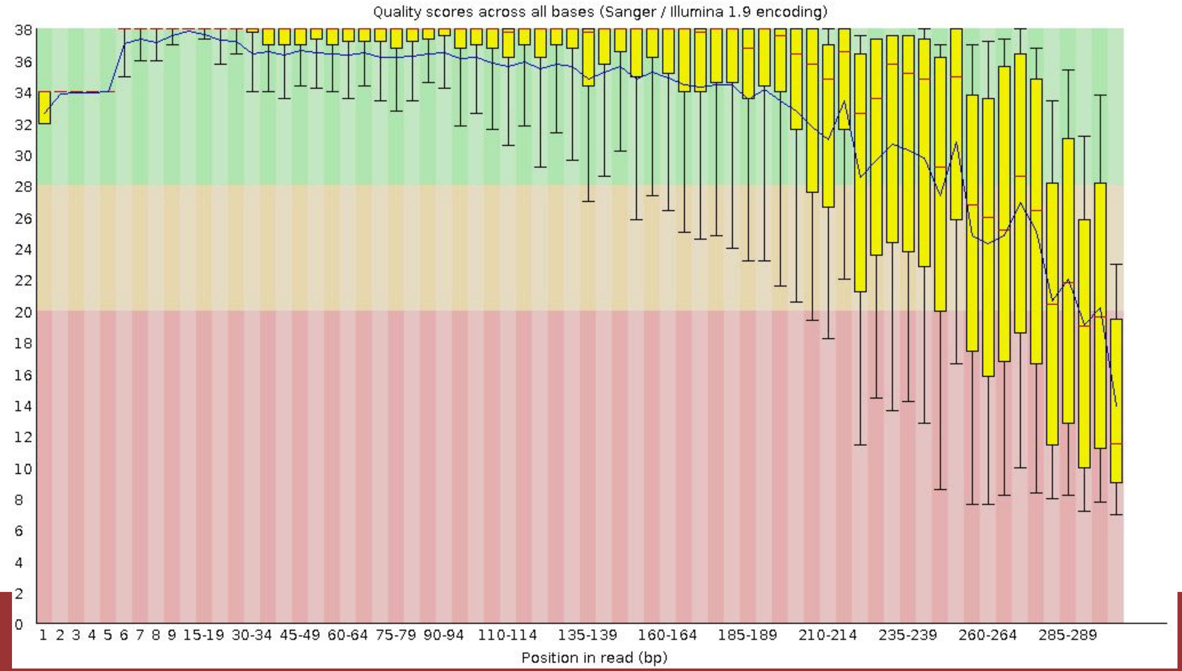
Outputs

Sample de-multiplexing

- “Multiplexing”: combining samples on the same run
- Unique DNA barcodes can be incorporated into your amplicons to differentiate samples
- Reads need to be linked back to the samples they came from using the unique barcodes
- De-multiplexing separates reads into individual sample files based on their barcodes
- Some sequencers will demultiplex for you

Quality Filtering

- Reads can be filtered (i.e. removed) using various criteria
- Reads can also be ”trimmed” to remove the lower quality part of the read
- Requires FastQ files as input



Quality Filtering

- Various methods for filtering

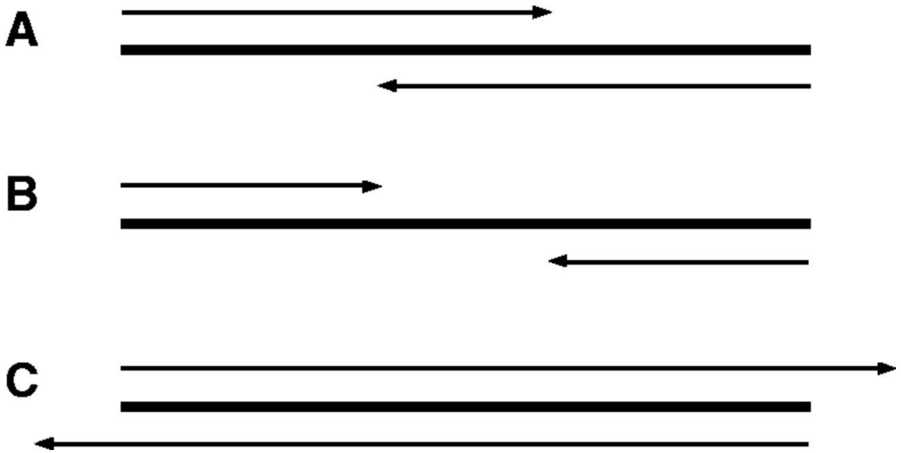
- Throw away read method:
  - Minimum base quality (e.g. q > 30)
  - Minimum percentage of high-quality bases (as % of total read length) (e.g. 90%)
  - Maximal number of ambiguous bases (N's)
  - Minimum read length

- Keep only reads with primer (and trim primer off)

- Other quality filtering tools available for "trimming"
  - Cutadapt (https://github.com/marcelm/cutadapt)
  - Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic)
  - Sickle (https://github.com/najoshi/sickle)

Read Joining/Stitching

- Paired-end reads result in a forward and reverse read from the same sequence

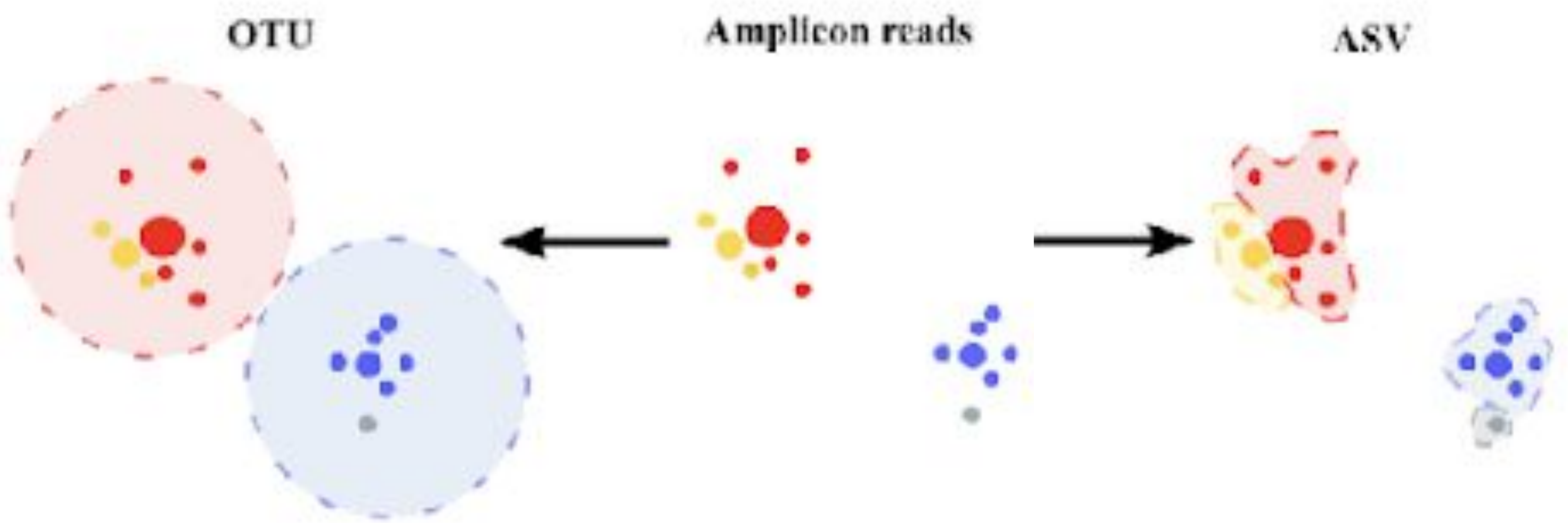


- VSEARCH can be run within QIIME
- PEAR is another alternative

Denoising

- Option 1: collapse based on sequence identity (i.e. 97%)
  - Operational taxonomic units (OTUs)

- Option 2: collapse by modelling and correcting sequencing errors
  - Amplicon sequence variants (ASVs)



OTU Picking

- OTUs: formed arbitrarily based on sequence identity
  - 97% of sequence similarity ≈ species

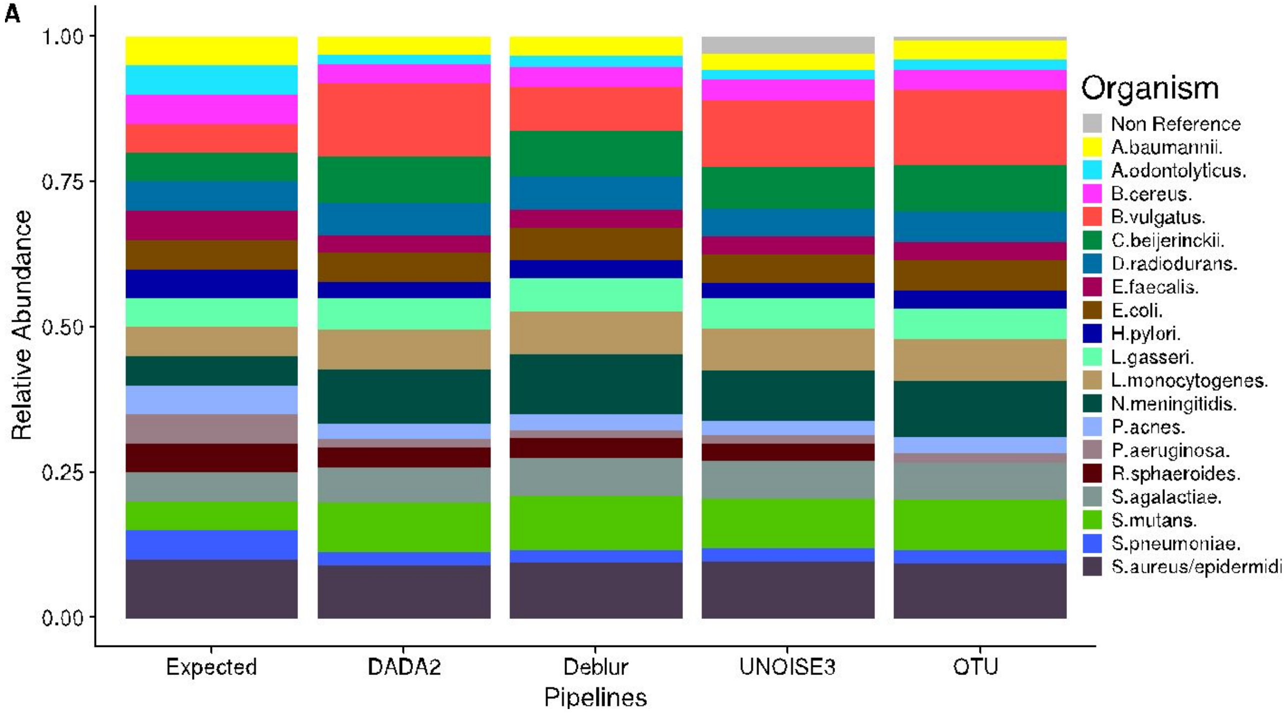
- Major approaches
  - De novo clustering
  - Closed-reference
  - Open-reference

- OTUs are not used as much in recent years

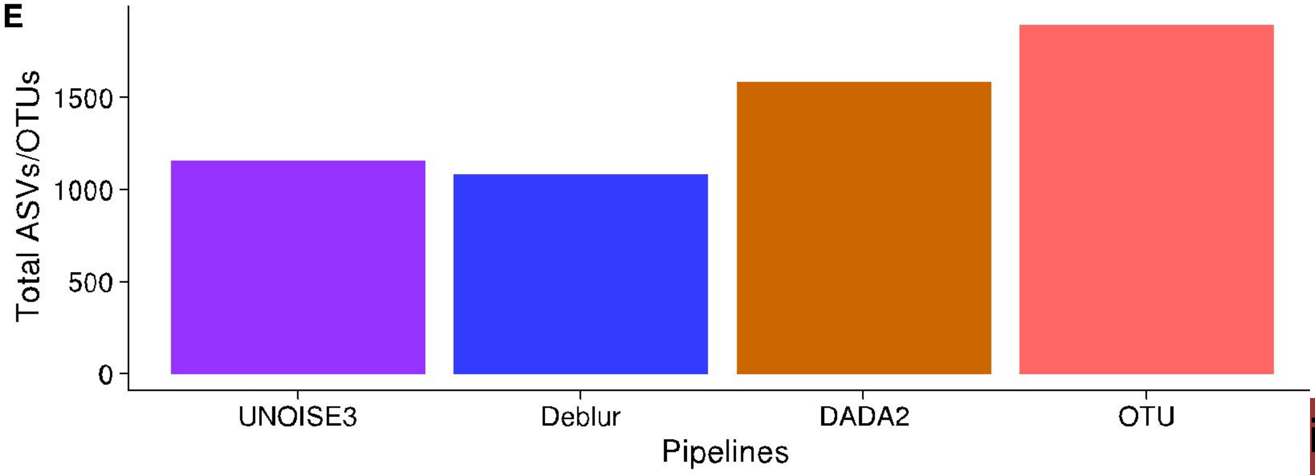
No OTU picking!

- Collapsing sequences at 97% removes information
- However, collapsing to 100% identity would allow a single nucleotide error to result in a spurious taxa
- Alternative, attempt to model errors and collapse to correct sequence (e.g. “denoising”)
- Instead of OTUs, called ASVs/sOTUs/etc.
- Current methods for sequence correction:
  - Dada2 (Susan Holmes)
  - Deblur (Rob Knight)
  - UNOISE2 (Robert Edgar)

Denoising the denoisers



(Nearing et al. 2018)

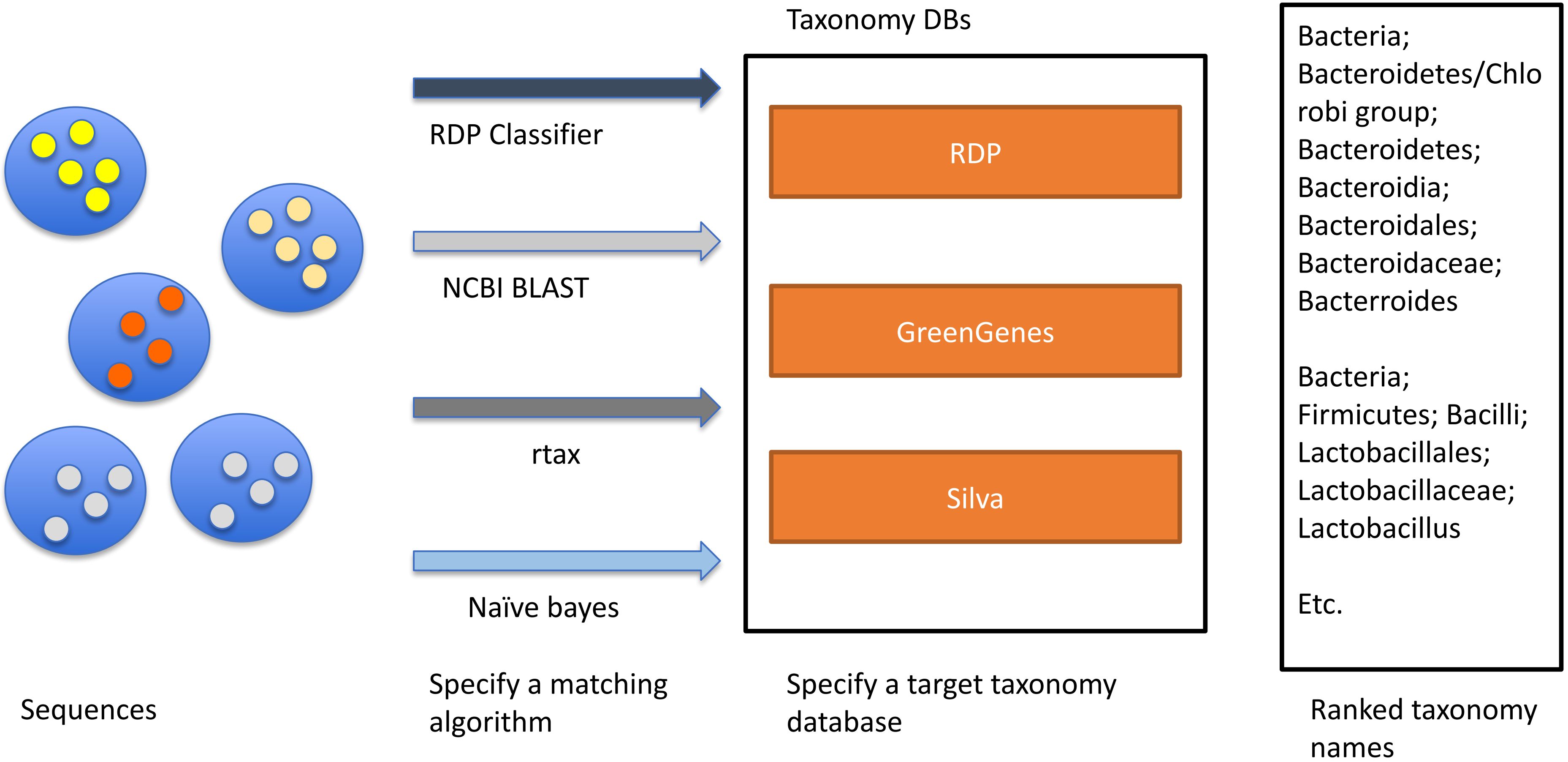


Taxonomic Assignment

- OTUs/ASVs can be analyzed without additional labels
  - ASVs are often simply reduced to a random string of characters representing their md5sum

  - However, taxonomic labels have advantages
    - Much easier to remember <-> easier communication
    - Taxa become known for their functions
    - Taxonomy allows grouping of related organisms at different resolutions
      - Collapse at Genus, Family, or even Phylum level

Taxonomic Assignment



Taxonomy DBs

RDP Classifier

NCBI BLAST

rtax

Naïve bayes

RDP

GreenGenes

Silva

Bacteria; Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacterroides

Bacteria; Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Lactobacillus

Etc.

Sequences

Specify a matching algorithm

Specify a target taxonomy database

Ranked taxonomy names

Taxonomy Databases

- RDP (Cole et al 2009)
  - Most similar to NCBI Taxonomy
  - Has a rapid classification tool (RDP-Classifier)

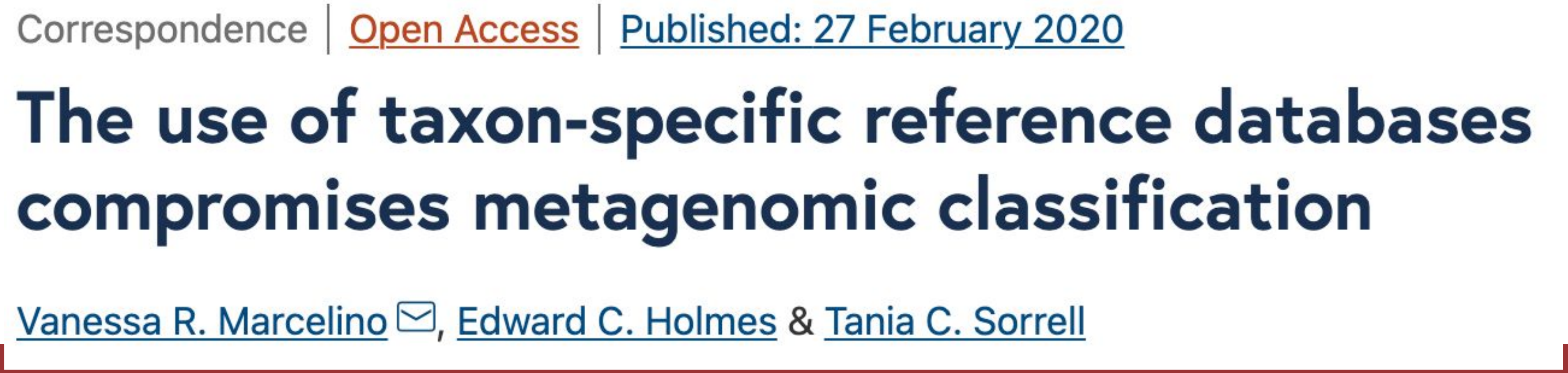
- Silva (Quast et al. 2013)
  - Preferred by Mothur in early days
  - Became preferred choice in recent years

- GreenGenes (McDonald et al 2012)
  - Once was preferred by QIIME but updates were lacking
  - However, GreenGenes2 is in preprint!

Special Taxonomy Databases

- Specific focused databases may provide better curated datasets and may provide more taxonomic resolution.

- However, overly focused (i.e. non-comprehensive) databases may lead to false positives



Correspondence | Open Access | Published: 27 February 2020

The use of taxon-specific reference databases compromises metagenomic classification

Vanessa R. Marcelino ✉, Edward C. Holmes & Tania C. Sorrell

OTU/ASV Table

- OTU/ASV table is a sample-by-observation matrix

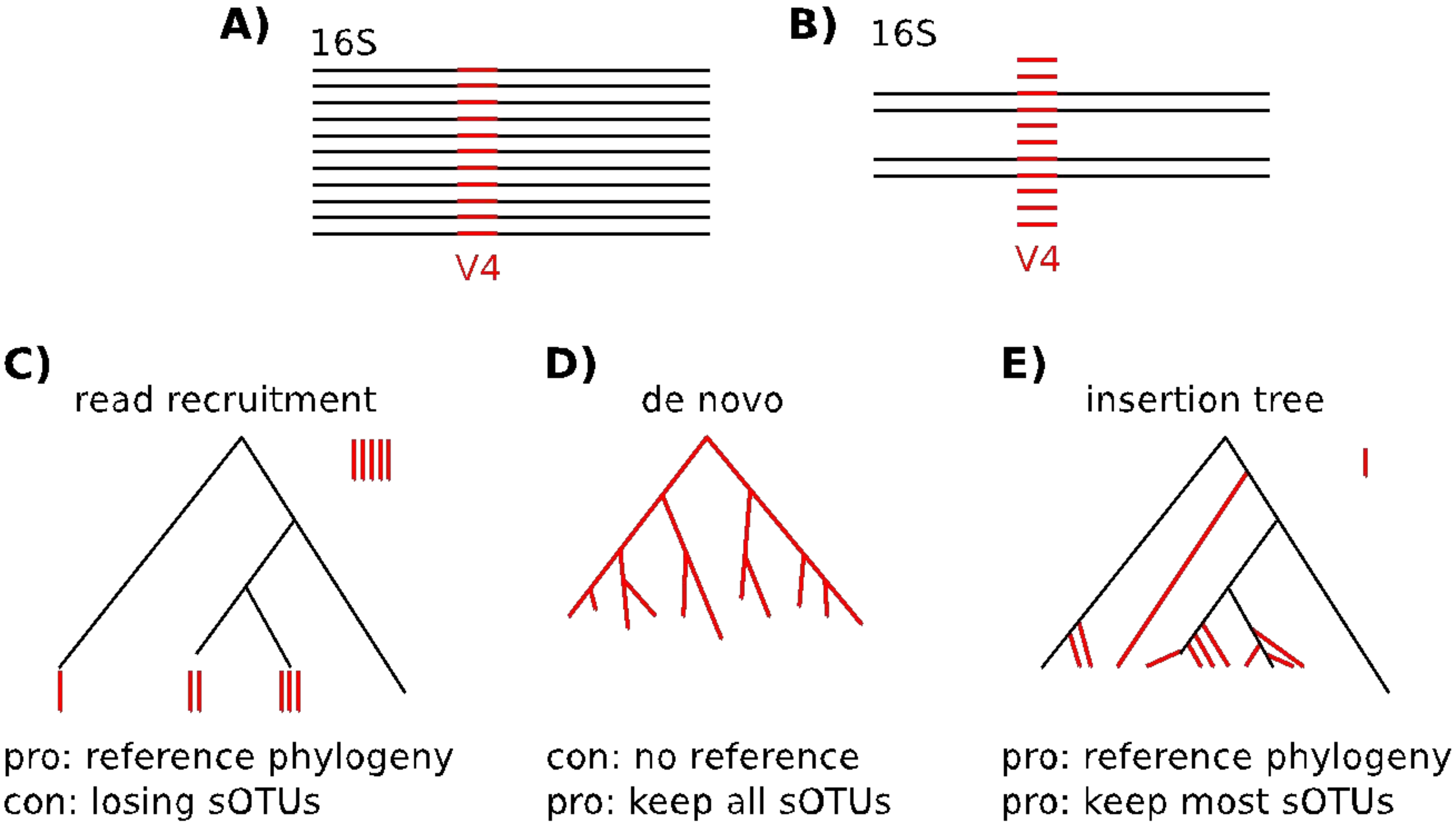
	Sample1	Sample2	Sample3	Sample4
OTU1	10	14	0	33
OTU2	5	0	54	2
OTU3…	5	3	7	9

- Table can be in multiple file formats
  - .tsv, .csv, .qza, .biom

More Filtering!

- Bleed-through ASVs: based on Illumina reporting (0.1% of mean sample depth)
- “Contaminant” ASVs: mitochondria, chloroplasts, etc.
- Other filtering criteria:
  - removing samples with low sequencing depth (<1000 reads)
  - prevalence filtering (present in <10% samples)

Phylogenetic Tree Reconstruction

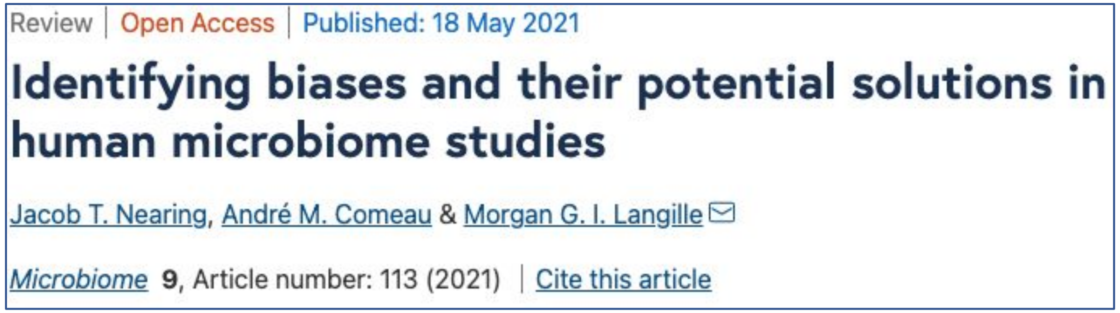


pro: reference phylogeny
con: losing sOTUs

con: no reference
pro: keep all sOTUs

pro: reference phylogeny
pro: keep most sOTUs

Janssen, et al. 2018

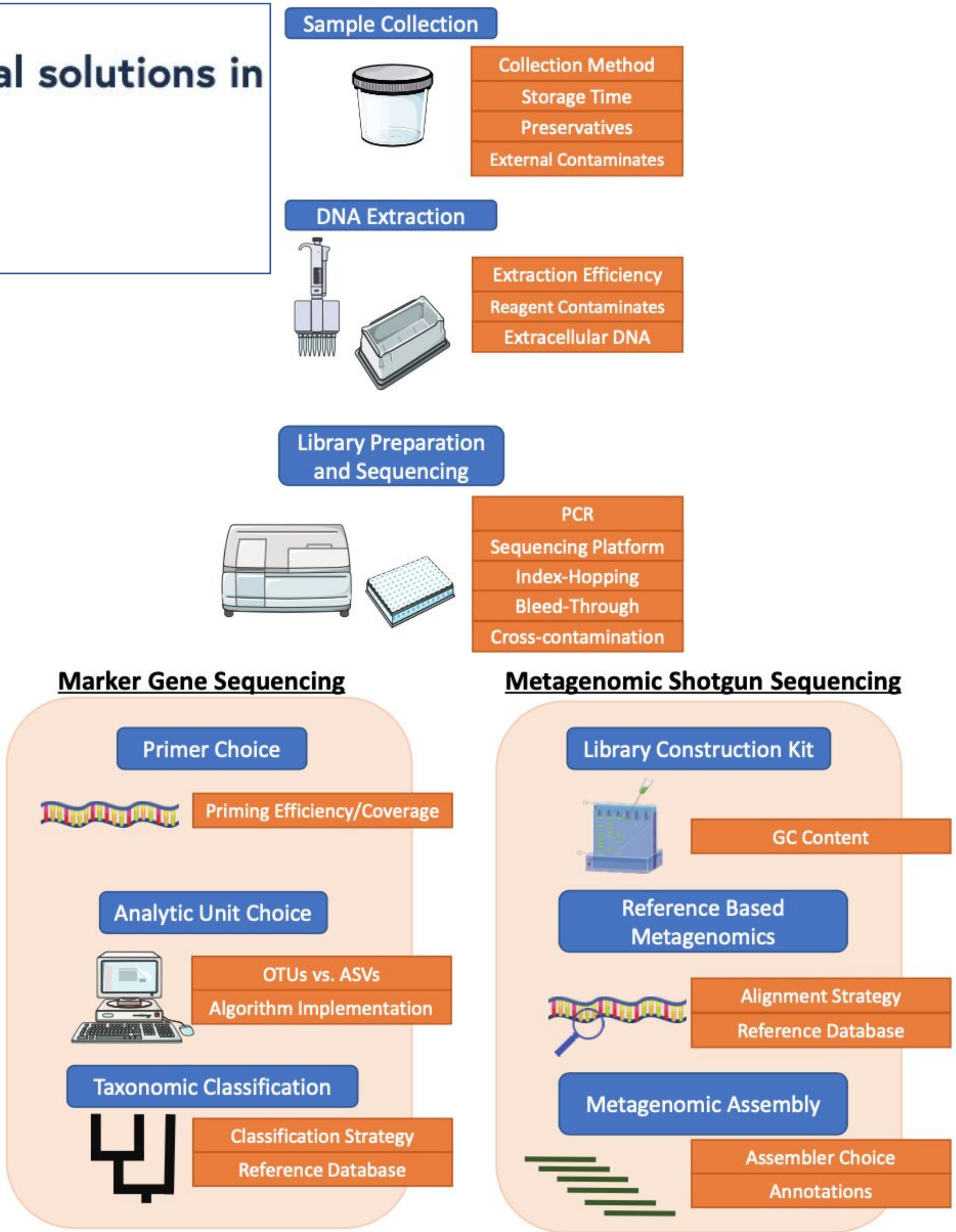


Review | Open Access | Published: 18 May 2021

Identifying biases and their potential solutions in human microbiome studies

Jacob T. Nearing, André M. Comeau & Morgan G. I. Langille

Microbiome **9**, Article number: 113 (2021) | Cite this article



Sample Collection

Collection Method
Storage Time
Preservatives
External Contaminates

DNA Extraction

Extraction Efficiency
Reagent Contaminates
Extracellular DNA

Library Preparation and Sequencing

PCR
Sequencing Platform
Index-Hopping
Bleed-Through
Cross-contamination

Marker Gene Sequencing

Primer Choice

Priming Efficiency/Coverage

Analytic Unit Choice

OTUs vs. ASVs
Algorithm Implementation

Taxonomic Classification

Classification Strategy
Reference Database

Metagenomic Shotgun Sequencing

Library Construction Kit

GC Content

Reference Based Metagenomics

Alignment Strategy
Reference Database

Metagenomic Assembly

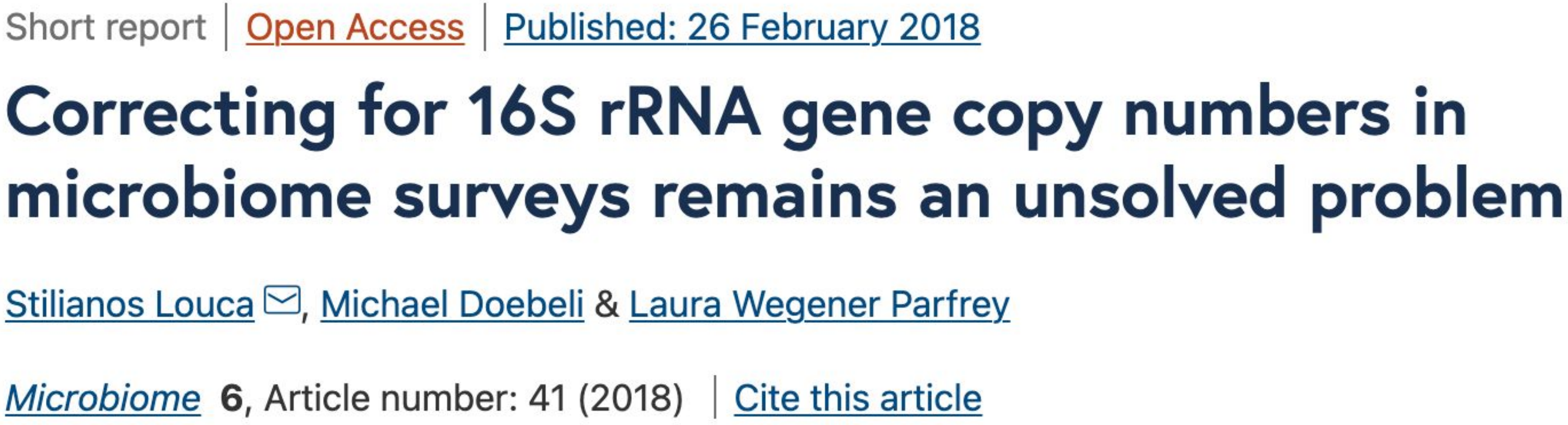
Assembler Choice
Annotations

16S copy number

- Some bacteria and archaea can have more than one copy of the 16S gene in its genome

- A few tools attempt to correct for this bias
    - PICRUSt, CopyRighter, PAPRICA

- Correcting for 16S copy number is not routine



Short report | Open Access | Published: 26 February 2018

Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem

Stilianos Louca, Michael Doebeli & Laura Wegener Parfrey

Microbiome **6**, Article number: 41 (2018) | Cite this article

QIIME2

- Start-to-finish microbiome analysis
  - Built on user-made plugins
  - Tracks workflow within file (provenance)

- Utilizes two file formats:
  - QZA: artifact file for analysis
  - QZV: visualization file

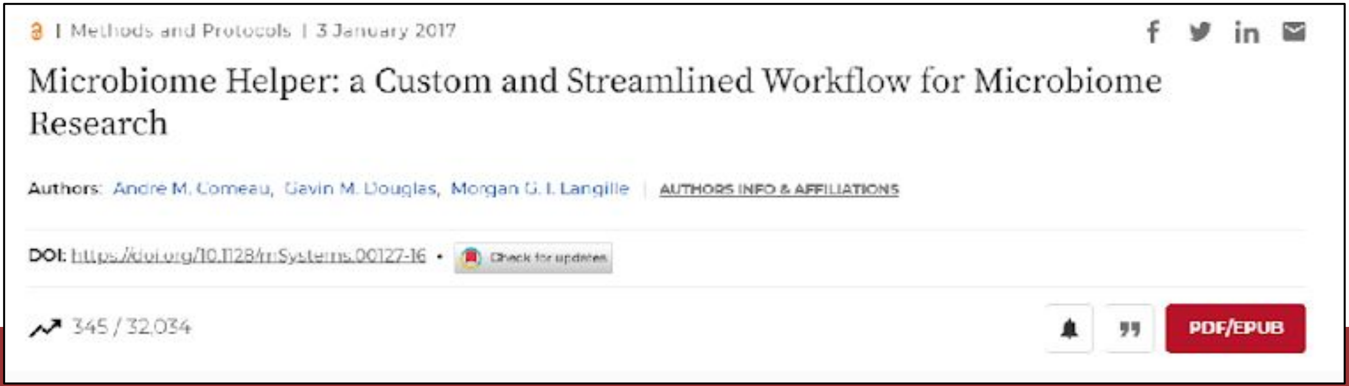
qiime2docs

https://docs.qiime2.org/2023.2/

- Core concepts
- Tutorials
- Plugin documentation
- Etc.

qiime2view

https://view.qiime2.org/



| Methods and Protocols | 3 January 2017

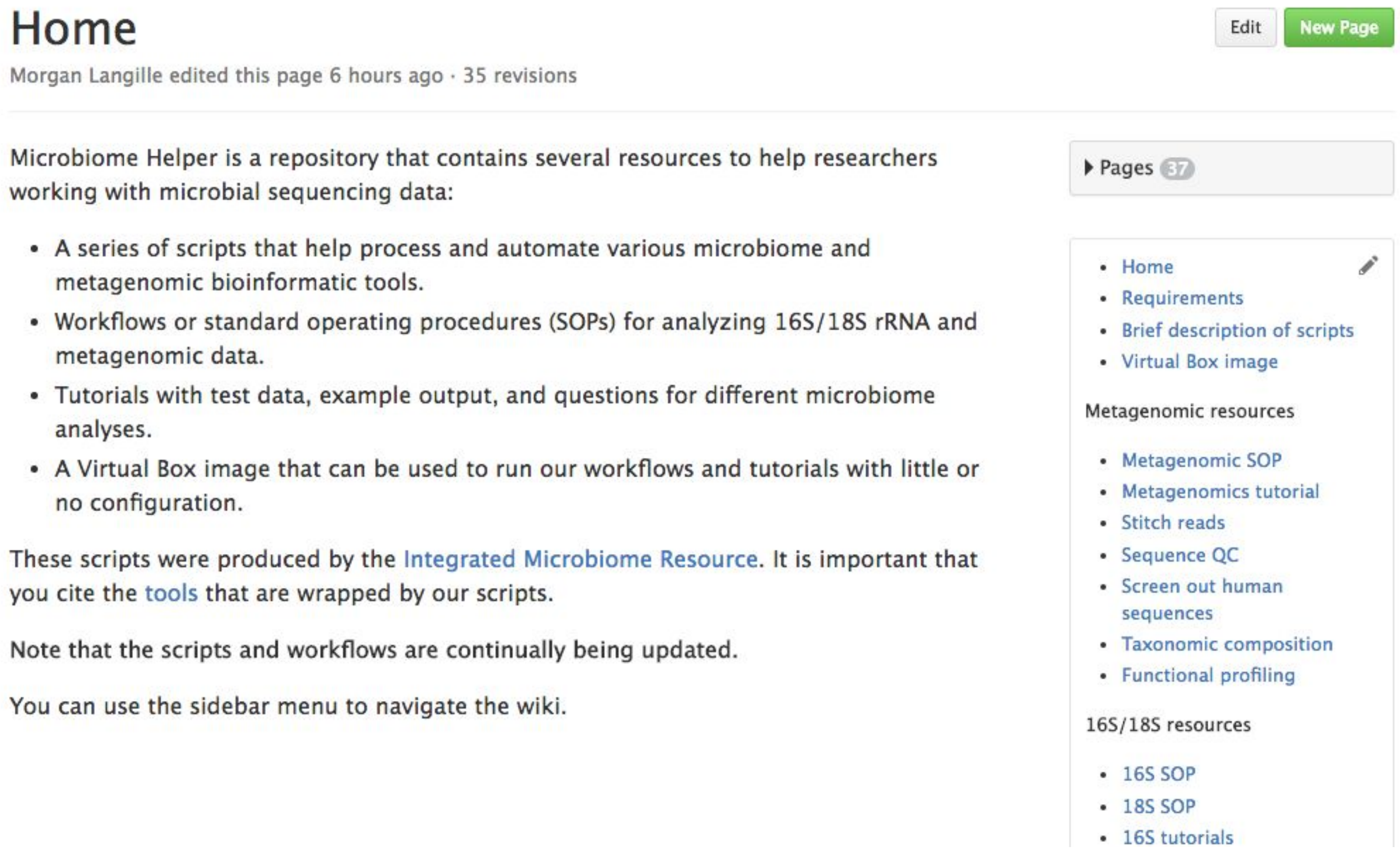
Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research

Authors: Andre M. Comeau, Gavin M. Douglas, Morgan G. I. Langille | AUTHORS INFO & AFFILIATIONS

DOI: https://doi.org/10.1128/mSystems.00127-16 • Check for updates

345 / 32,034

Microbiome Helper Wiki



Home

Morgan Langille edited this page 6 hours ago · 35 revisions

Microbiome Helper is a repository that contains several resources to help researchers working with microbial sequencing data:

- A series of scripts that help process and automate various microbiome and metagenomic bioinformatic tools.
- Workflows or standard operating procedures (SOPs) for analyzing 16S/18S rRNA and metagenomic data.
- Tutorials with test data, example output, and questions for different microbiome analyses.
- A Virtual Box image that can be used to run our workflows and tutorials with little or no configuration.

These scripts were produced by the Integrated Microbiome Resource. It is important that you cite the tools that are wrapped by our scripts.

Note that the scripts and workflows are continually being updated.

You can use the sidebar menu to navigate the wiki.

https://github.com/mlangill/microbiome_helper/wiki

Questions?

We are on a Coffee Break & Networking Session

Workshop Sponsors:

Canadian Centre for Computational Genomics

HPC4Health

OICR Ontario Institute for Cancer Research

GenomeCanada

IMPACTT