Data analysis of metagenomics experiments: 16S

Dr. Natalia Zajac

natalia.zajac@fgcz.ethz.ch

29.03.2023

Microbes are everywhere

We live in a microbial world.

Coprinus comatus

Homo sapiens

Zea mays

Porphyra yezoensis

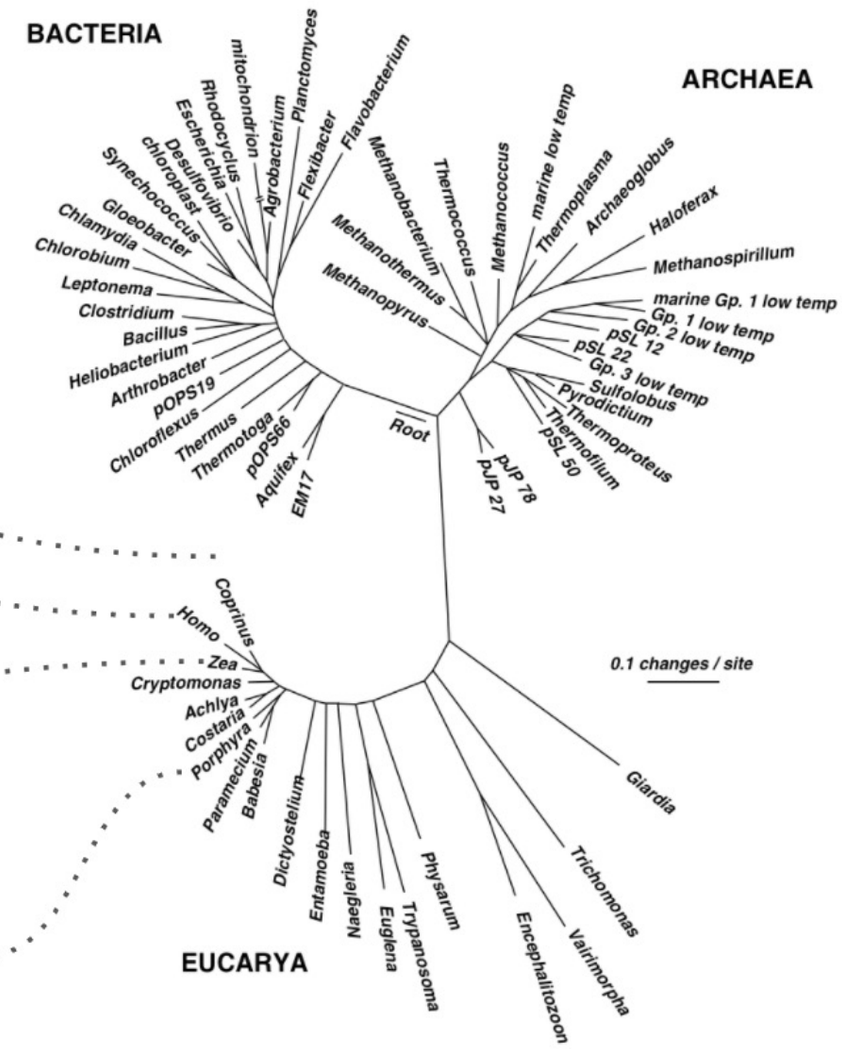


Image sources:
https://en.wikipedia.org/wiki/Homo_sapiens#/media/File:Akha_cropped_hires.JPG
https://en.wikipedia.org/wiki/Coprinus#/media/File:Coprinus_comatus_fresh.jpg
https://en.wikipedia.org/wiki/Maize#/media/File:Corntassel_7095.jpg
https://en.wikipedia.org/wiki/Porphyra#/media/File:Porphyra_yezoensis.jpg

Applications

Understanding microbiomes may lead to new food varieties and more sustainable crops.

Our microbiomes impact the efficacy of medical treatment.

Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors

Bertrand Routy,[1,2,3] Emmanuelle Le Chatelier,[4] Lisa Derosa,[1,2,3] Connie P. M. Duong,[1,2,5] Maryam Tidjani Alou,[1,2,3] Romain Daillère,[1,2,3] Aurélie Fluckiger,[1,2,5] Meriem Messaoudene,[1,2] Conrad Rauber,[1,2,3] Maria P. Roberti,[1,2,5] Marine Fidelle,[1,3,5] Caroline Flament,[1,2,5] Vichnou Poirier-Colame,[1,2,5] Paule Opolon,[6] Christophe Klein,[7] Kristina Iribarren,[8,9,10,11,12] Laura Mondragón,[8,9,10,11,12] Nicolas Jacquelot,[1,2,3] Bo Qu,[1,2,3] Gladys Ferrere,[1,2,3] Céline Clémenson,[1,13] Laura Mezquita,[1,14] Jordi Remon Masip,[1,14] Charles Naltet,[15] Solenn Brosseau,[15] Coureche Kaderbhai,[16] Corentin Richard,[16] Hira Rizvi,[17] Florence Levenez,[4] Nathalie Galleron,[4] Benoit Quinquis,[4] Nicolas Pons,[4] Bernhard Ryffel,[18] Véronique Minard-Colin,[1,19] Patrick Gonin,[1,20] Jean-Charles Soria,[1,14] Eric Deutsch,[1,13] Yohann Loriot,[1,3,14] François Ghiringhelli,[16] Gérard Zalcman,[15] François Goldwasser,[9,21,22] Bernard Escudier,[1,14,23] Matthew D. Hellmann,[24,25] ges,[1,3,14]

ancer immunotherapy (see the consumption is associated with poor rofiled samples from patients with lung

Microbes can help us reduce our impact on the Earth by composting our food waste. And, maybe even by degrading pollution...

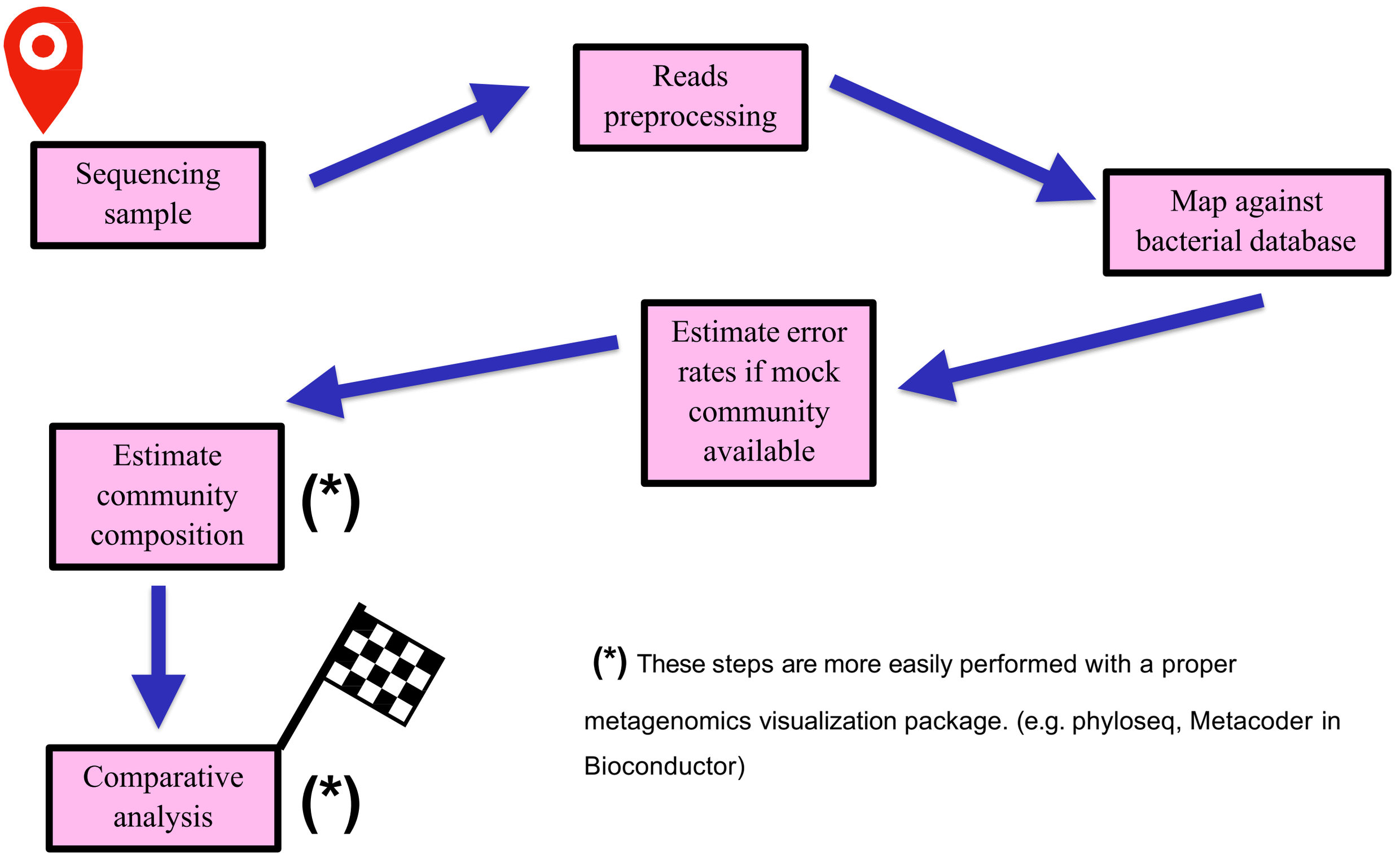
Outline

- 16S-based analysis
  - Main goals
  - Main challenges
  - Approaches
  - Databases
  - File format

Ribosomal RNA

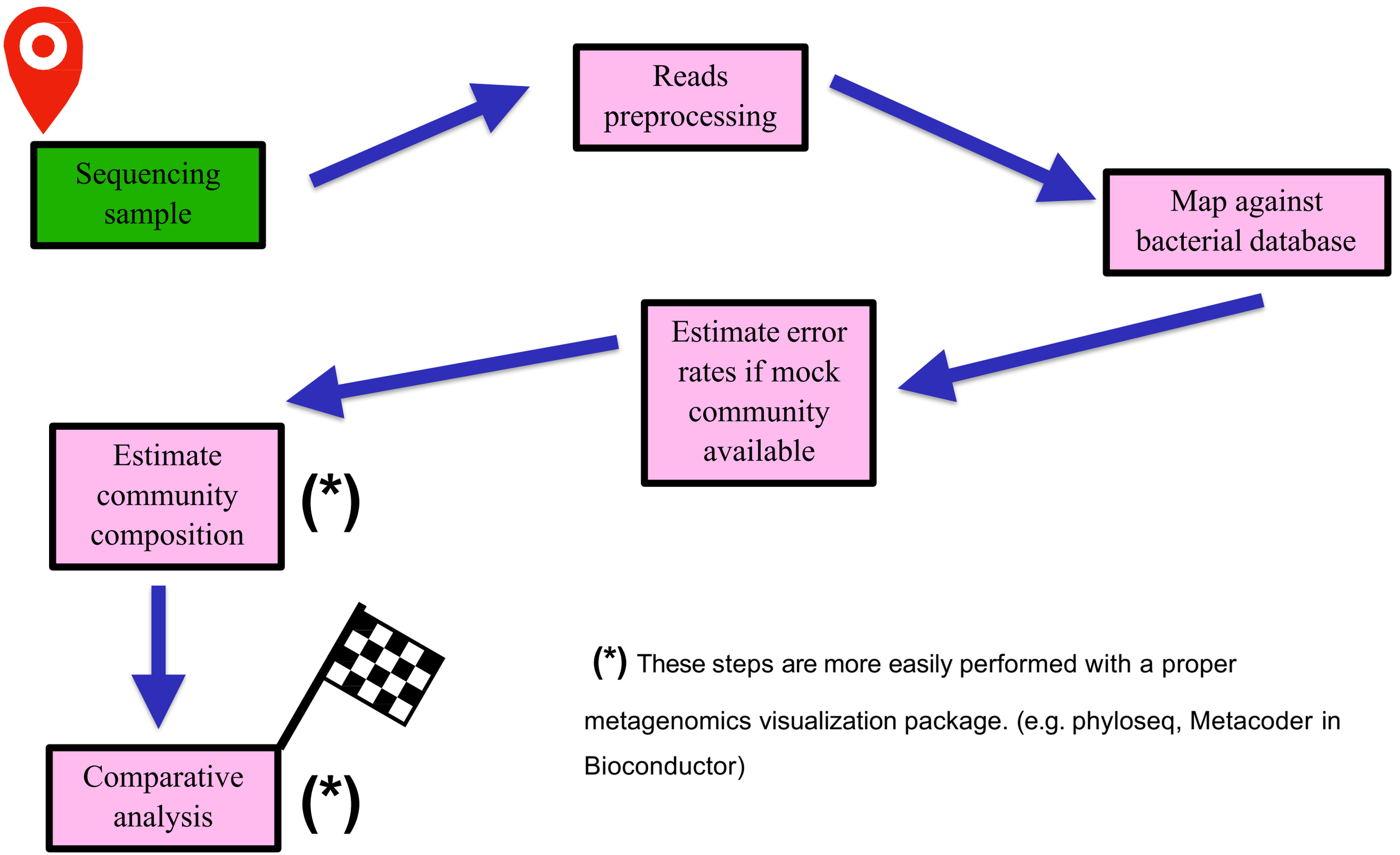
- Present in every organism
- Essential for protein translation
- Relatively conserved
- Most widely used
  - 16S (Bacteria)
  - 18S (Eukaryotes)
  - ITS (Fungi) - Nuclear ribosomal internal transcribed spacer (ITS)
- Multiple markers combination also possible
  - PhyloSift, MetaPhlAn2

16S example workflow and associated challenges



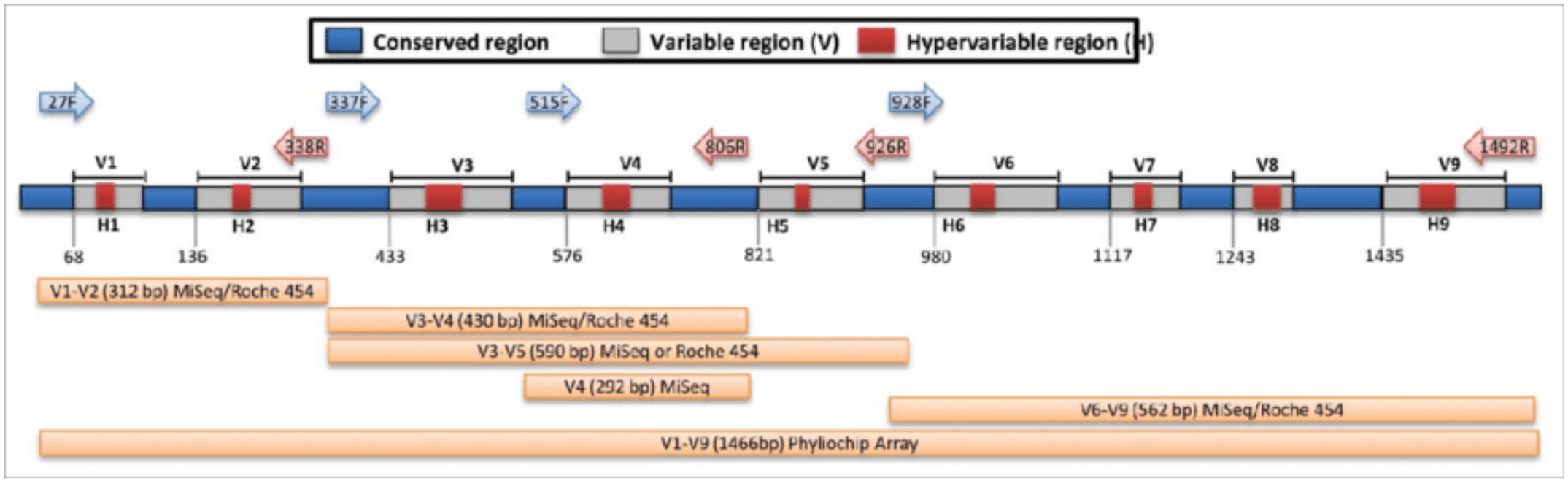
(*) These steps are more easily performed with a proper metagenomics visualization package. (e.g. phyloseq, Metacoder in Bioconductor)

16S example workflow and associated challenges



(*) These steps are more easily performed with a proper metagenomics visualization package. (e.g. phyloseq, Metacoder in Bioconductor)

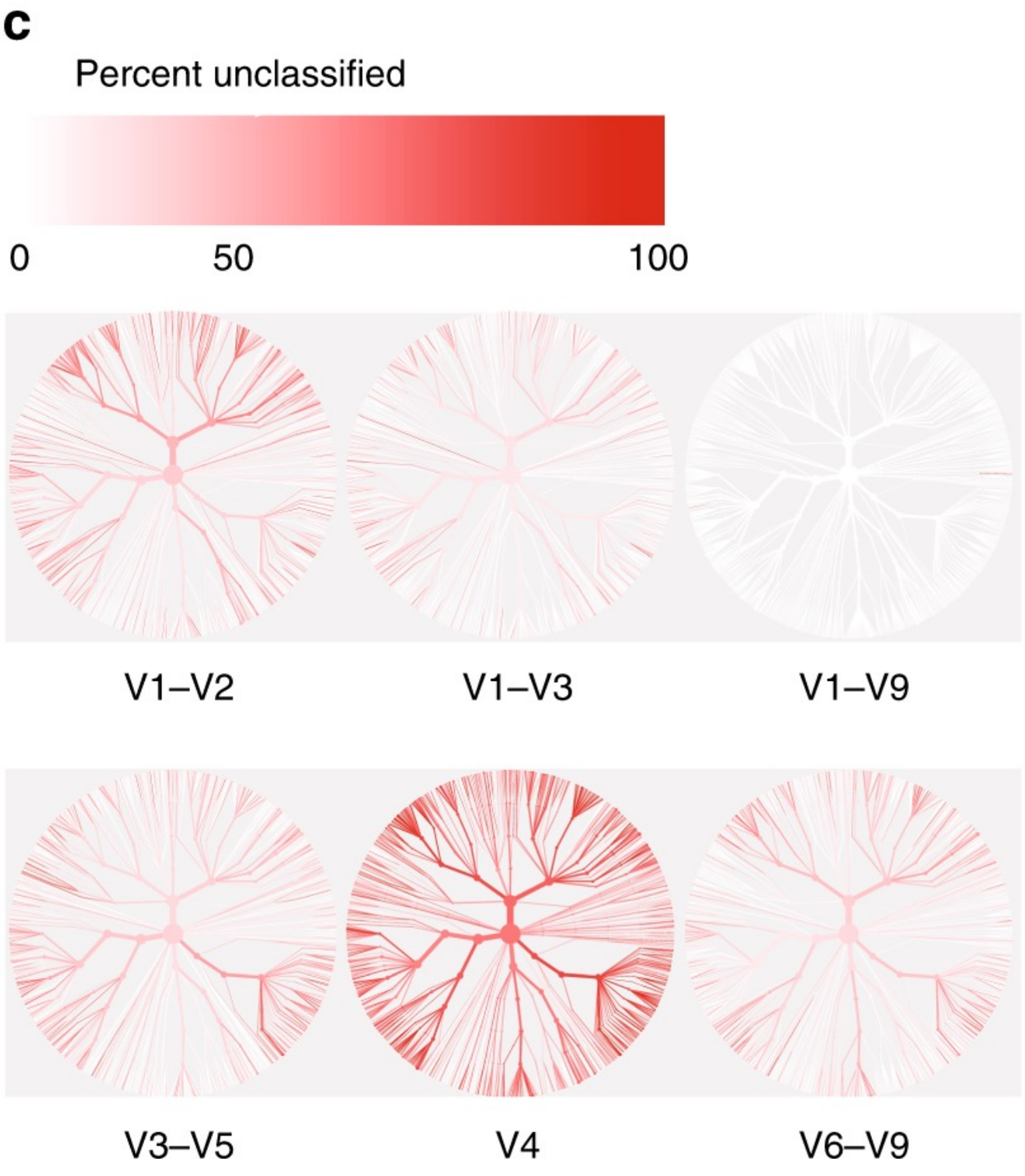
Sequencing sample



- What am I amplifying? (16S, ITS)
- Which regions? (Vx)
- Which primers? (literature has quite a few)

*Choice of primers can lead to potential biases in the representation of the taxonomic units

- How many reads I need? (10K, 50K, more?)
- Which technology is the best? (do I need full length?)
- What databases are available? (Silva or not?)

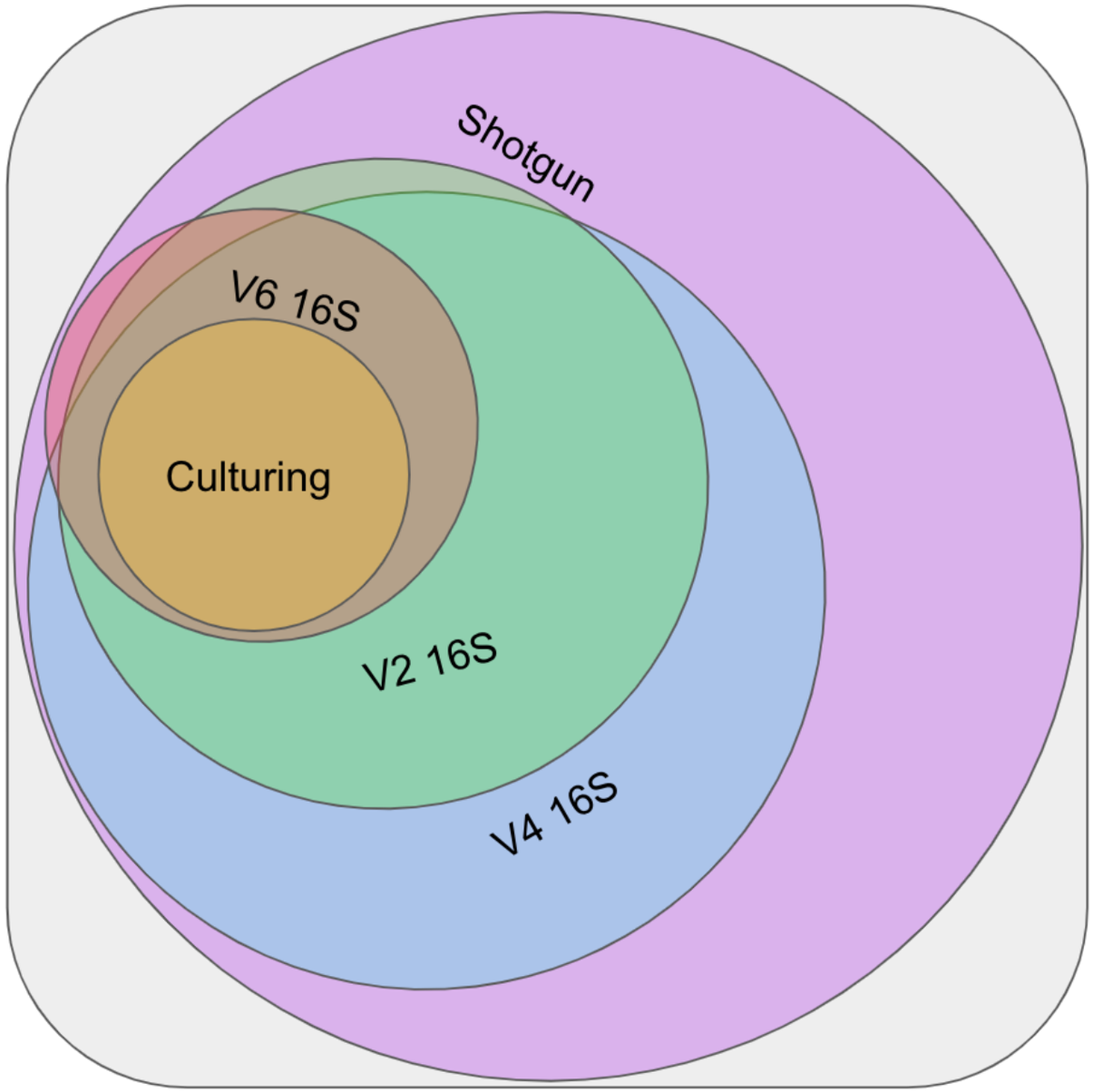


- V1–V2 region performed poorly at classifying sequences belonging to the phylum Proteobacteria
- V3–V5 region performed poorly at classifying sequences belonging to the phylum Actinobacteria but great in identification of *Klebsiella*

Johnson et al. 2019

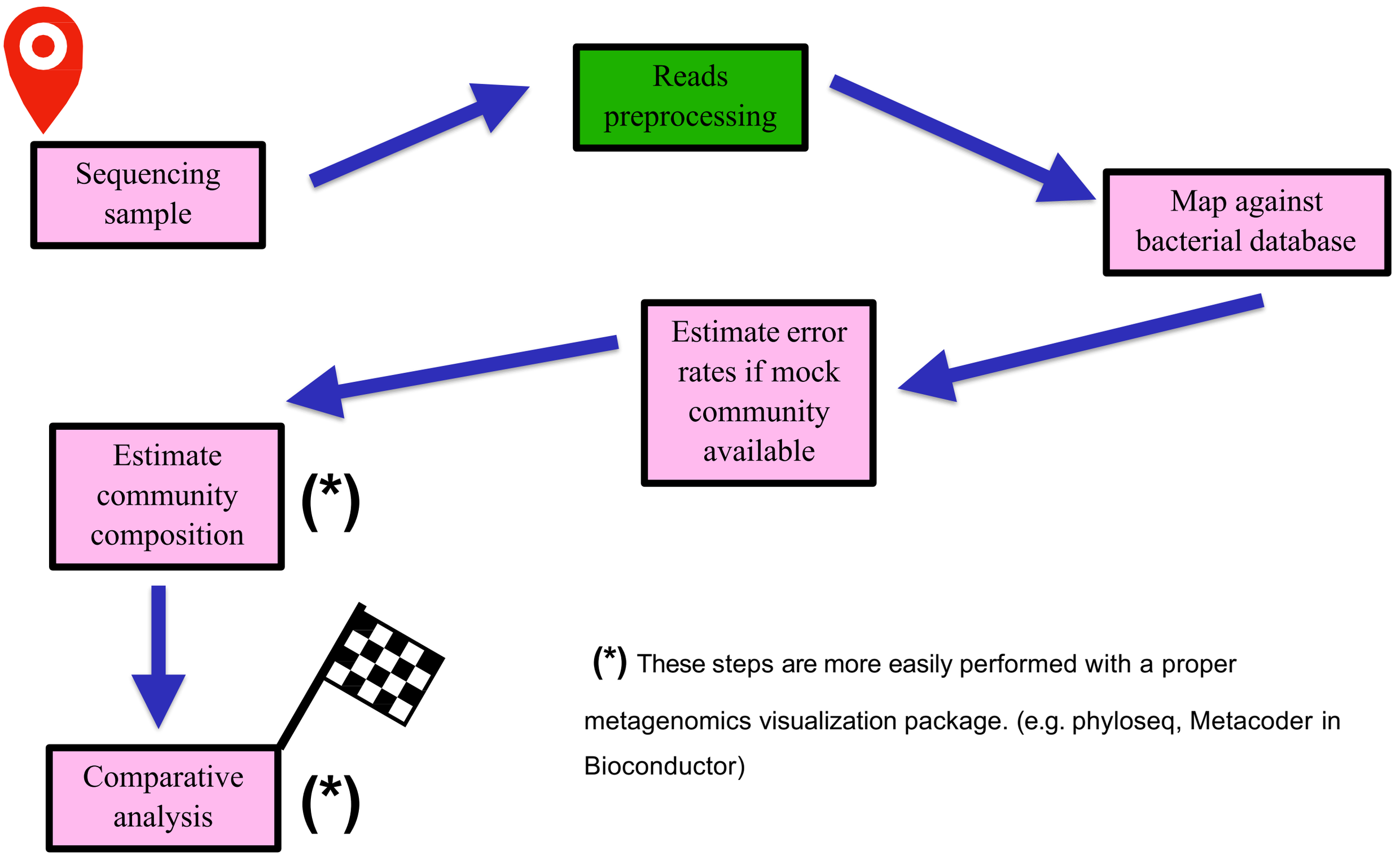
Resolution

We illuminate different regions of the microbial world (represented in grey) with different technologies (represented in other colors).



https://docs.google.com/presentation/d/1lg6nH-M-Myr5KoYRDYDYDTuhl4YMs8_6yQ0xLcx0Shc/edit#slide=id.g4437732396_0_281

16S example workflow and associated challenges



(*) These steps are more easily performed with a proper metagenomics visualization package. (e.g. phyloseq, Metacoder in Bioconductor)

Reads preprocessing

- How many reads I have per sample?
- Are there issues with quality? (trimming by quality, length)
- What filtering do I need? (adapter/other contamination)
- How many reads I have after filtering?
- What is the average accuracy?

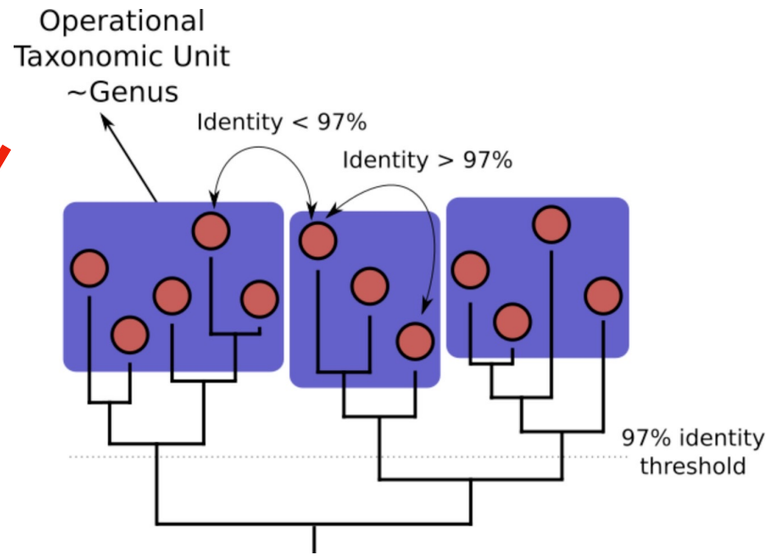
Discard reads that
(1) do not match the primers, or
(2) have ambiguous bases (Ns), or
(3) paired reads that do not have perfect matching overlaps (mismatches or length difference)

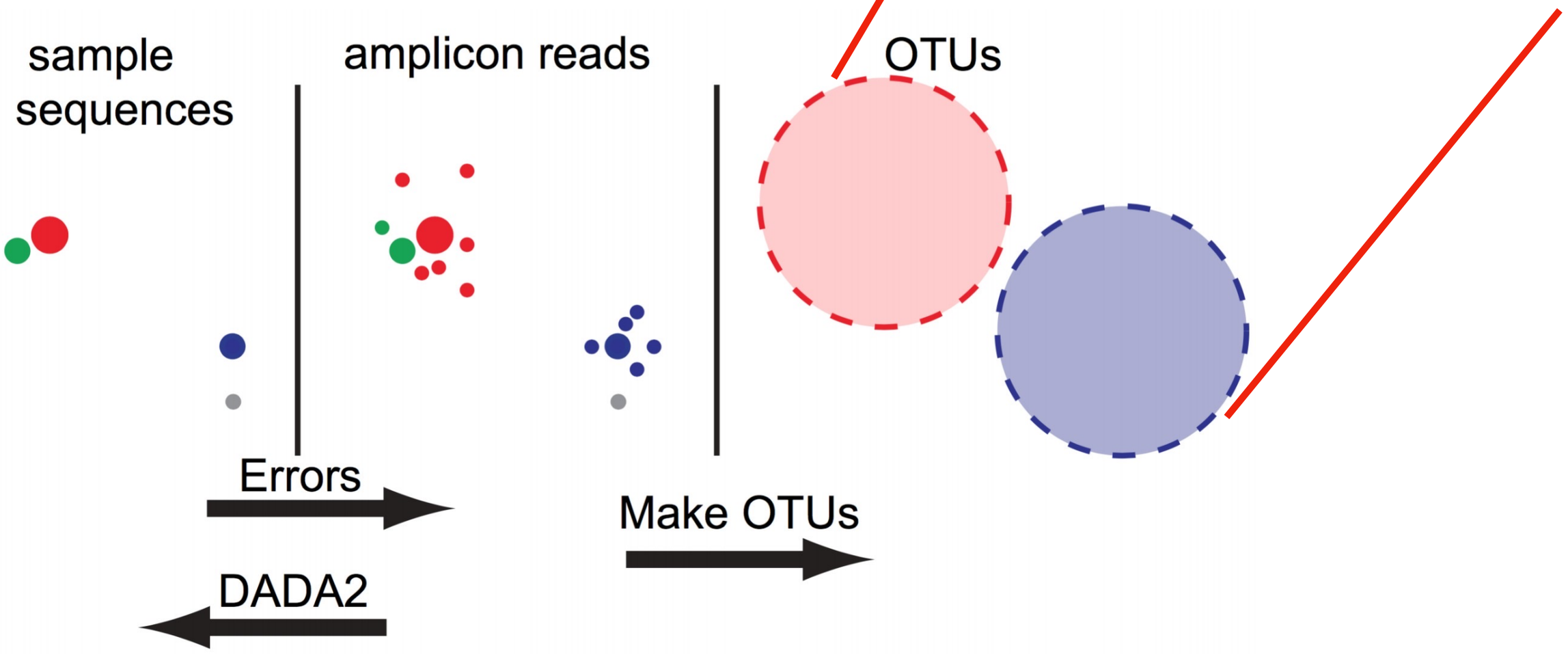
Preprocessing summary

Step	Description
Trim/Merge	Trim front or tail of sequences if low quality, remove adapters Merge paired end reads
Quality filter	Remove sequences by quality
Find unique/dereplicate/sort	Cluster exact same sequences, Cluster at > 99% id, this has the effect of removing most sequences with up to 1% errors, Sort by decreasing abundance. More abundance sequences make better centroids.
Cluster OTUs*	Discard all singletons, which are most likely chimeras, so set cluster size to >2
Make OTU table*	Cluster sequences to the same OTU by 97% sequence identity, using the plus strand

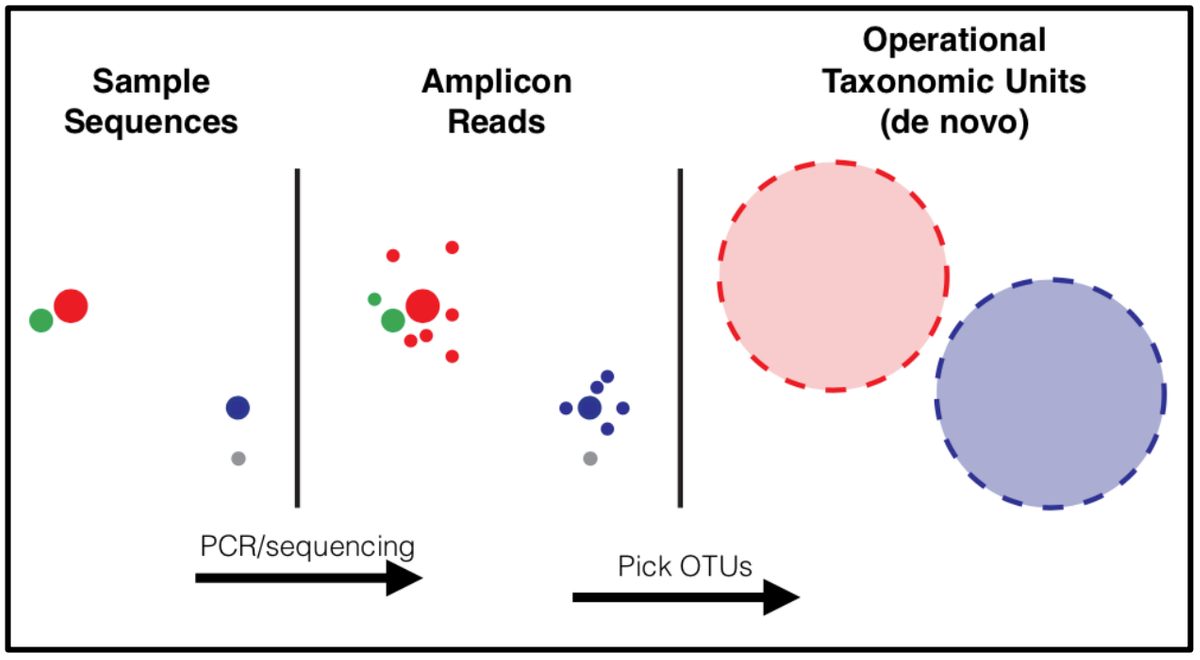
*optional

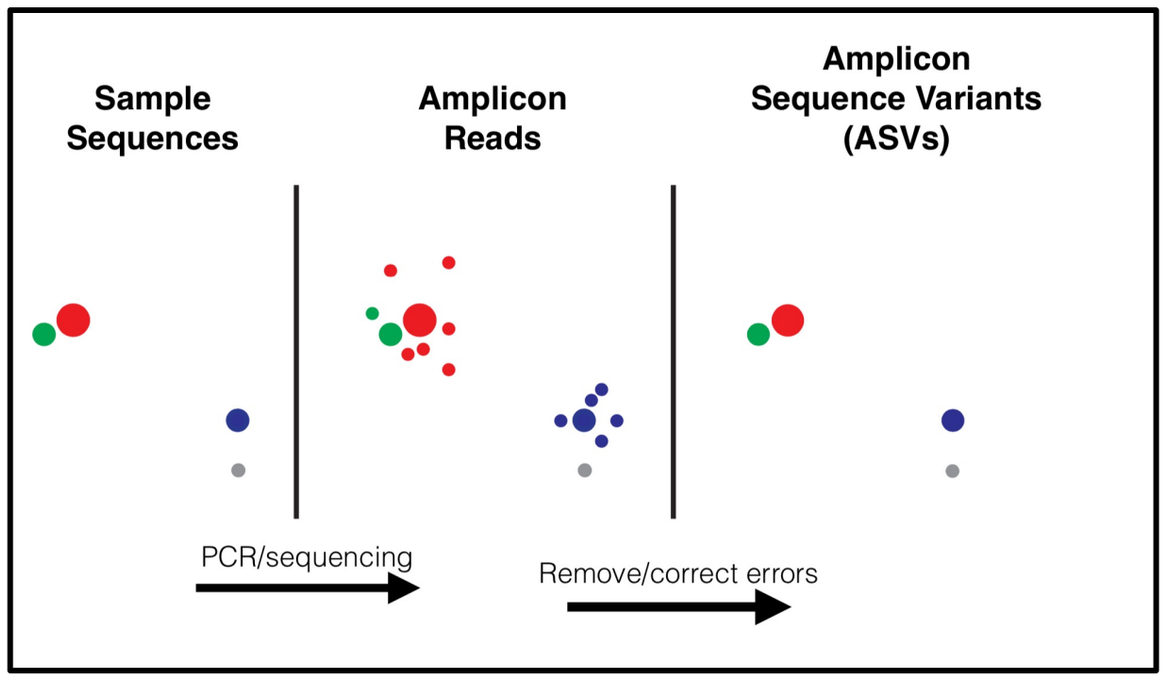
Operational Taxonomic Units (OTUs) vs Amplicon Sequence Variants (ASV)





Operational Taxonomic Units (OTUs) vs Amplicon Sequence Variants (ASV)





Operational Taxonomic Units (OTUs) vs Amplicon Sequence Variants (ASV)

OTUs (e.g., Mothur)	ASV (e.g., DADA2)
Result of sequence clustering (typically at 97% identity)	Only identical variants are clustered
Representative sequence obtained	Each cluster is a unique variant
Cannot distinguish very subtle strains	Can distinguish complex communities
Needs fewer reads	Requires more reads
Can get away with imperfect data	Needs virtually error-free sequences

Clustering

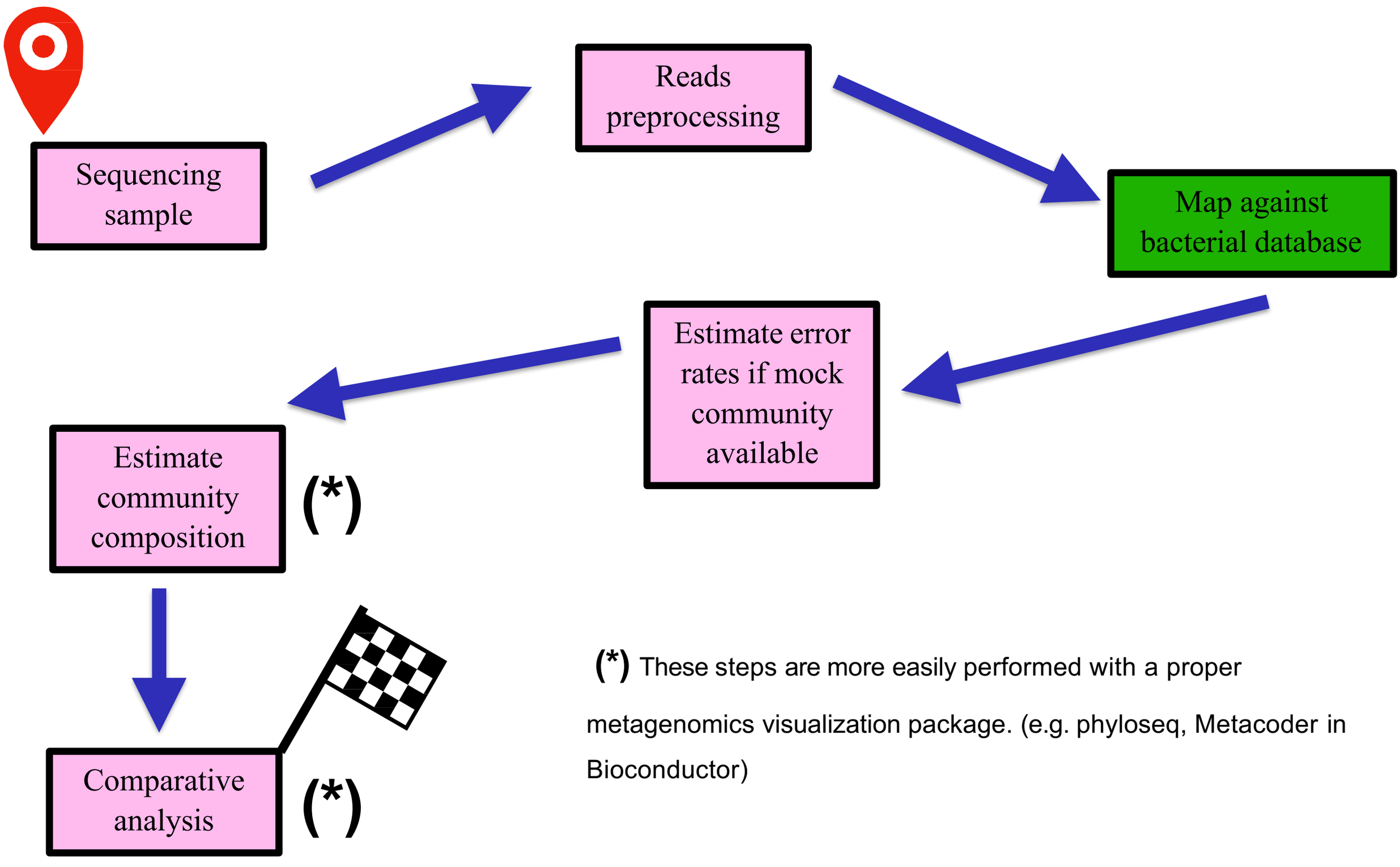
- Reference-free - no reference database and creates the clusters entirely from observed sequences

- Reference-based:
  - Closed-reference: uses a reference database of target gene sequences from known taxa and compares discovered sequences to them

  - Open-reference: open-reference clustering was developed, where sequences that can be quickly clustered to a reference database are clustered in a manner similar to closed-reference and remaining sequences are clustered in a manner similar to *de novo*

https://www.zymoresearch.com/blogs/

16S example workflow and associated challenges



(*) These steps are more easily performed with a proper metagenomics visualization package. (e.g. phyloseq, Metacoder in Bioconductor)

Map against bacterial database

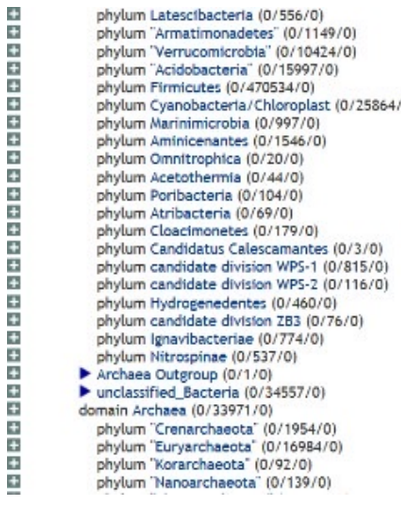
- Am I attempting species/strain classification?
- What is known about the available databases?
  - Is it biased towards certain kingdoms or ranks?
- Is there anything specific for my experiment?
  - E.g., very particular soil samples with large presence of certain organisms

16S databases

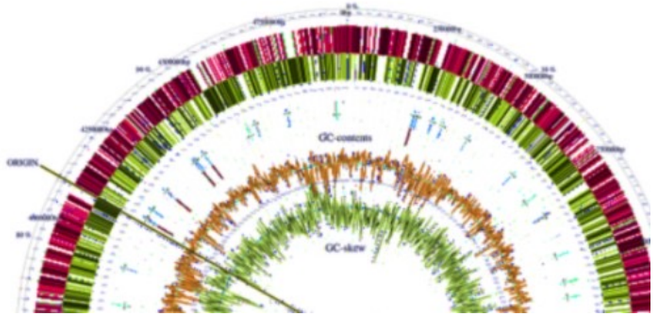
Genomics Inform. 2018 Dec; 16(4): e24.
Published online 2018 Dec 28. doi: 10.5808/GI.2018.16.4.e24
PMCID: PMC6440677
PMID: 30602085

Evaluation of 16S rRNA Databases for Taxonomic Assignments Using a Mock Community

Sang-Cheol Park[1] and Sungho Won[1,2,3,*]



RDP II: Cole et al.
NAR (2013)

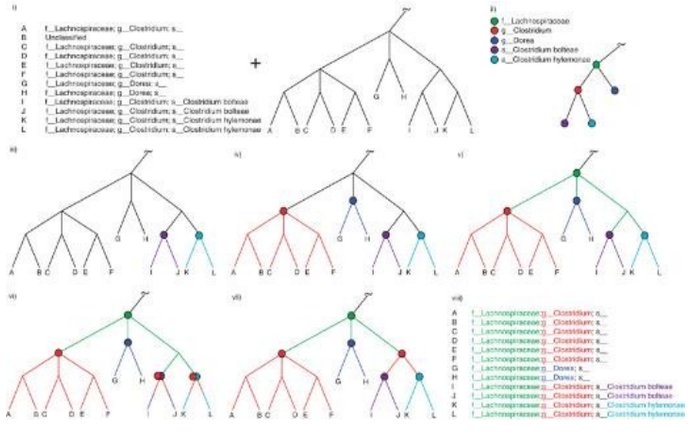


rrnDB: Stoddard et al.
NAR (2014)

EzBioCloud: Yoon et al.
PubMed (2017)

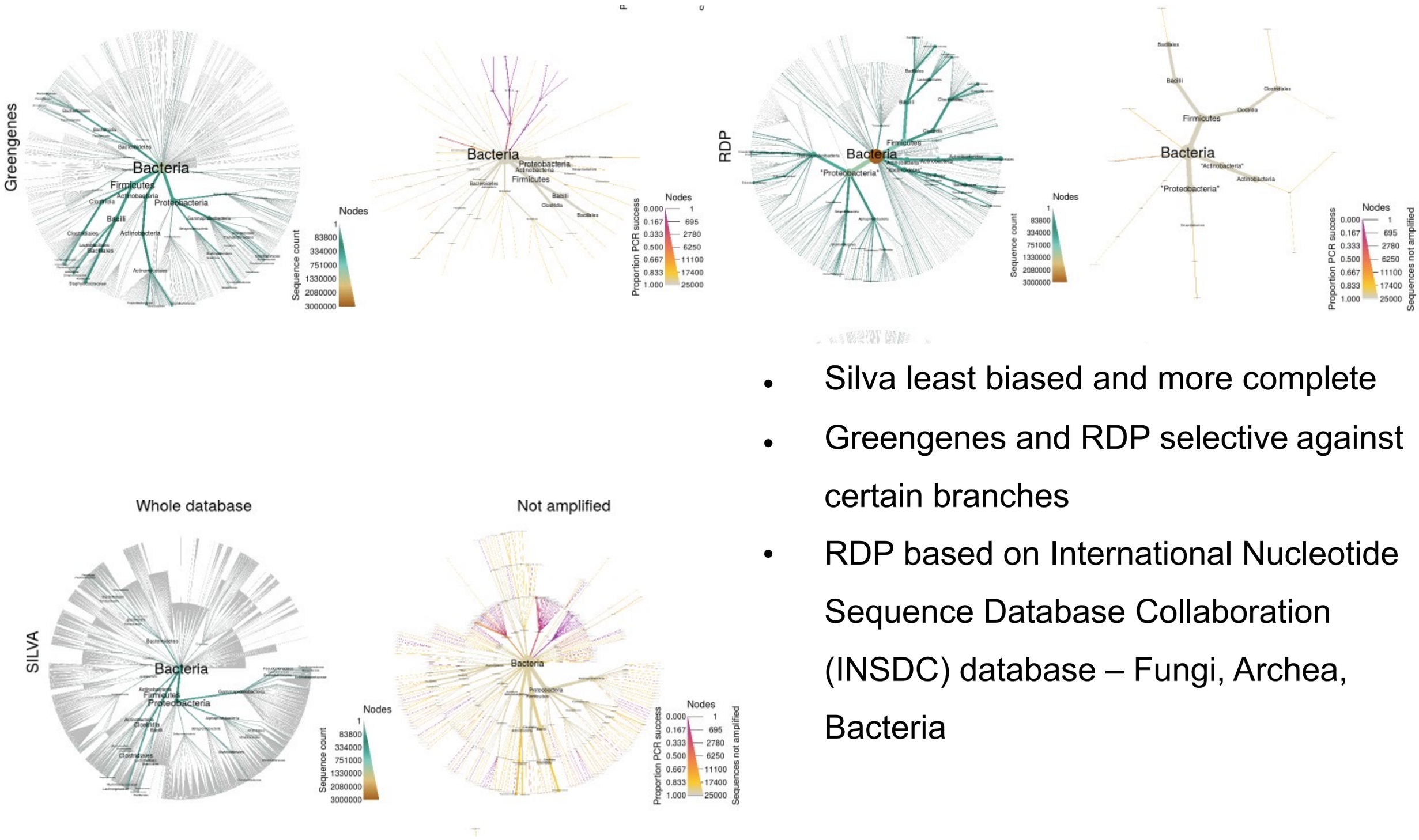
silva
high quality ribosomal RNA databases

SILVA: Quast et al.
NAR (2013)



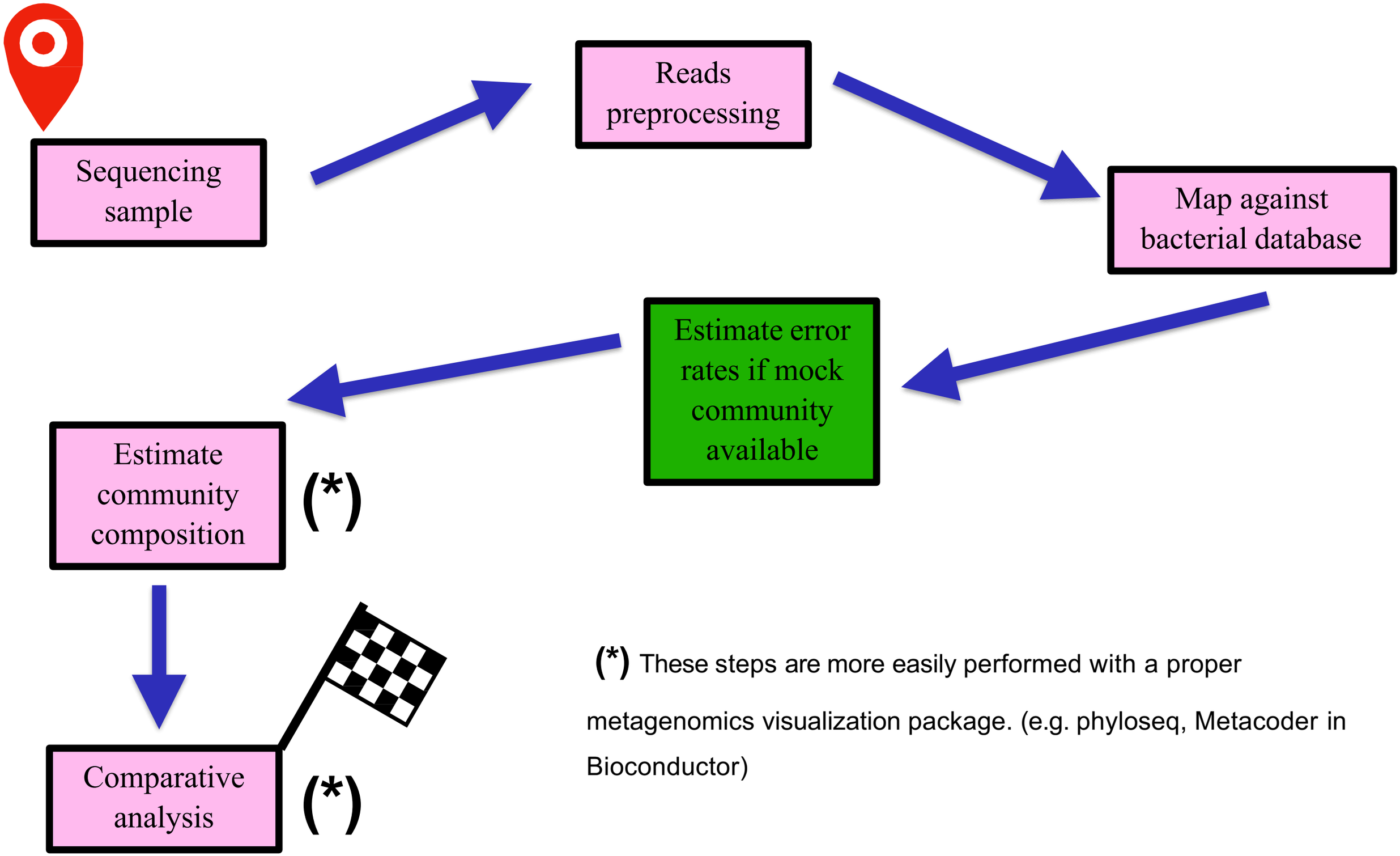
GreenGenes:
MacDonald et al. *ISME J* (2012)

16S databases



- Silva least biased and more complete
- Greengenes and RDP selective against certain branches
- RDP based on International Nucleotide Sequence Database Collaboration (INSDC) database – Fungi, Archea, Bacteria

16S example workflow and associated challenges

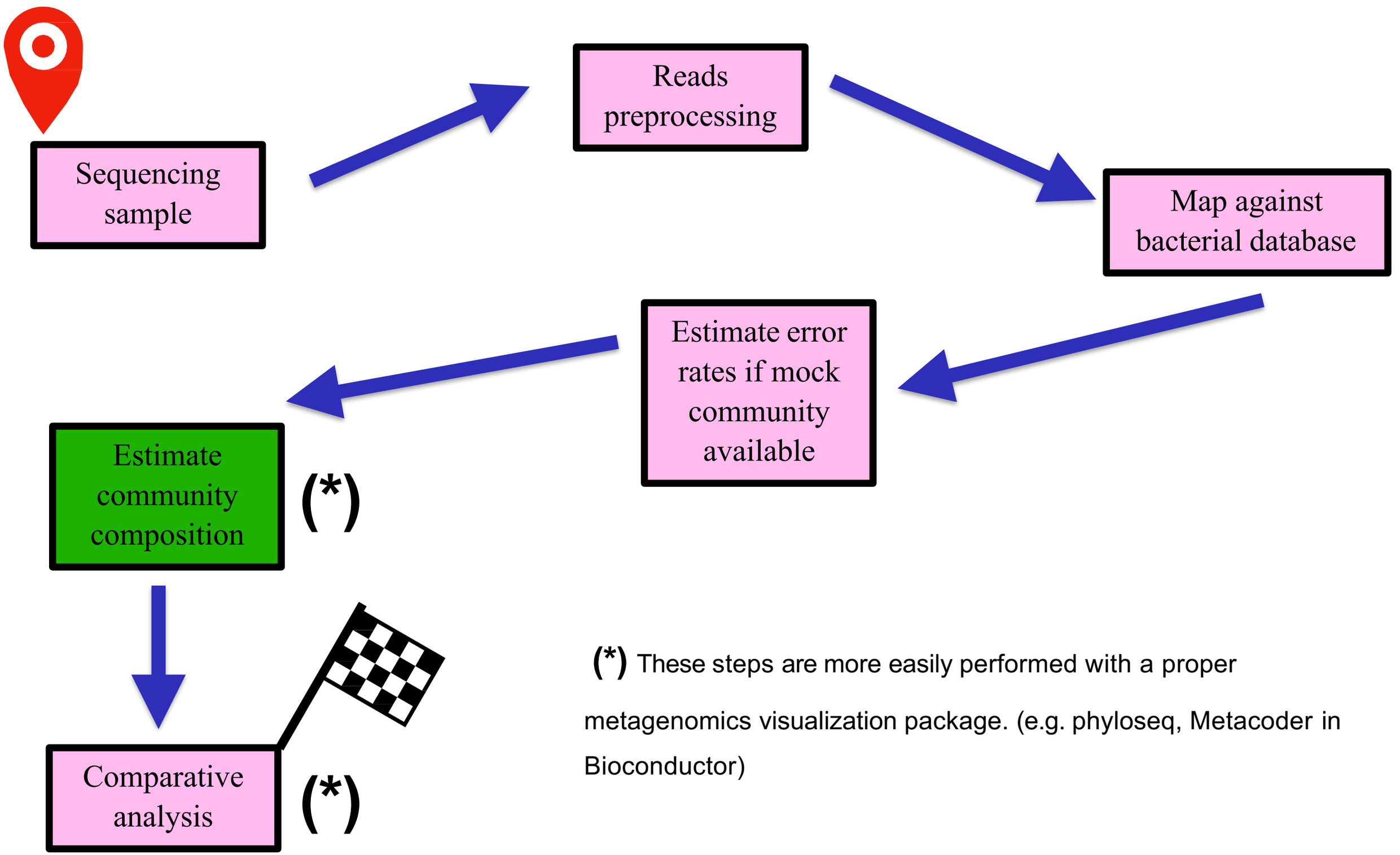


(*) These steps are more easily performed with a proper metagenomics visualization package. (e.g. phyloseq, Metacoder in Bioconductor)

Estimate error rates if mock community available

- Do I have a group truth available?
- Do I need it?
- How reliable the sequences are?
- What is the error rate of my data?
- What rank would be reasonable to achieve with such an error rate?

16S example workflow and associated challenges

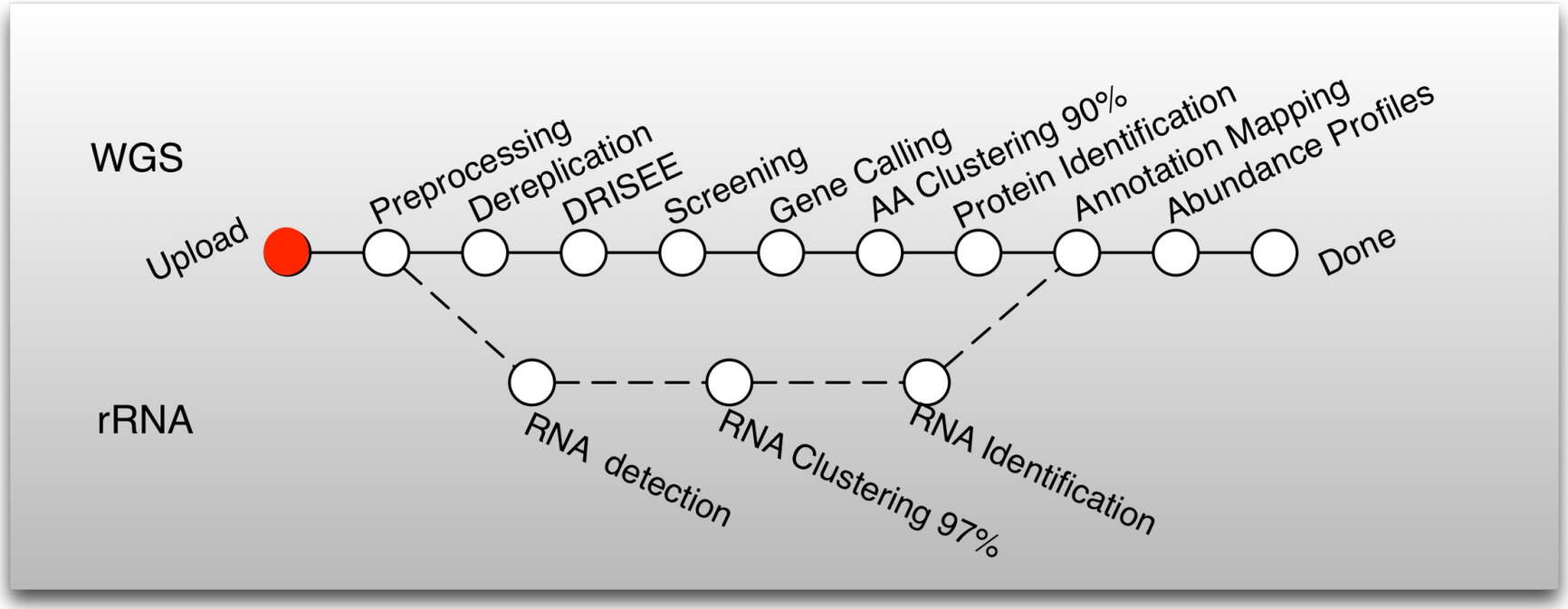


(*) These steps are more easily performed with a proper metagenomics visualization package. (e.g. phyloseq, Metacoder in Bioconductor)

Estimate community composition

- Which method will I use (e.g., ASV or OTU)?
- Which tool?
- Which filtering on the abundance data?
- Is my community complete?
- What is the diversity?
- Do I need more data?

MG-RAST pipeline



All analysis is done on a web-server. You need to set up an account.

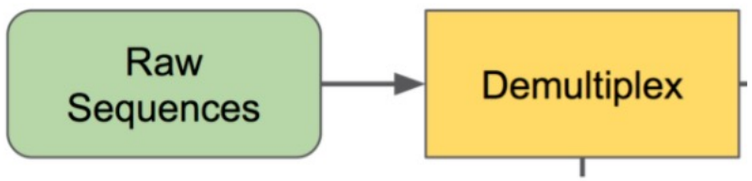
rRNA identification

A BLAT similarity search for the longest cluster representative is performed against the M5rna database which integrates SILVA(Pruesse et al. 2007), Greengenes(DeSantis et al. 2006), and RDP(Cole et al. 2003).

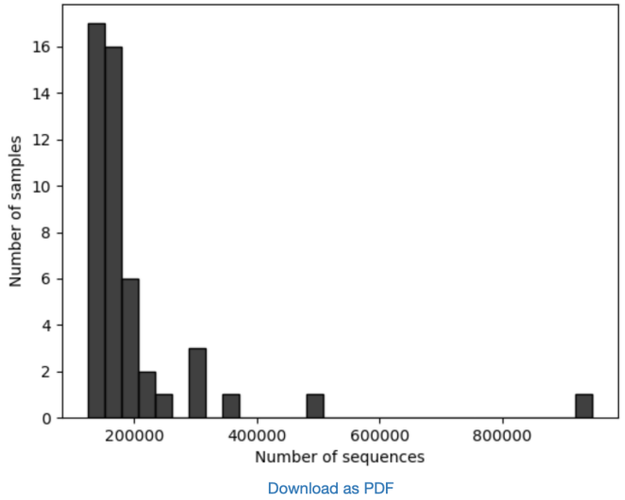
Conceptual overview of QIIME 2

Now that we have read the glossary and key, let us examine a conceptual overview of the various possible workflows for examining amplicon sequence data:

https://docs.qiime2.org/2022.2/tutorials/overview/



1. Per sample read counts
2. Read frequency histograms

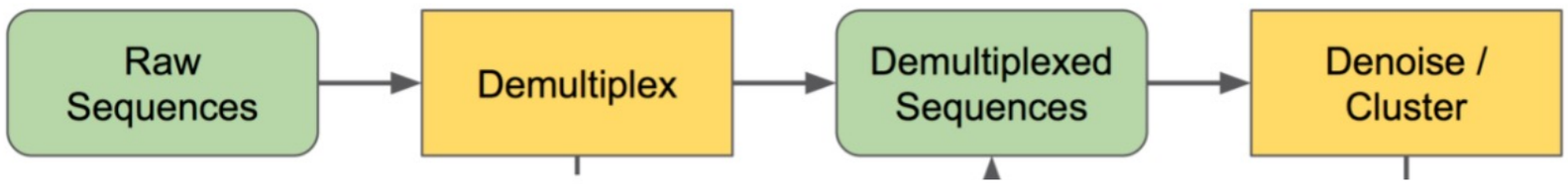


Requires installation but can be installed on a linux, windows or mac.

Conceptual overview of QIIME 2

Now that we have read the glossary and key, let us examine a conceptual overview of the various possible workflows for examining amplicon sequence data:

https://docs.qiime2.org/2022.2/tutorials/overview/



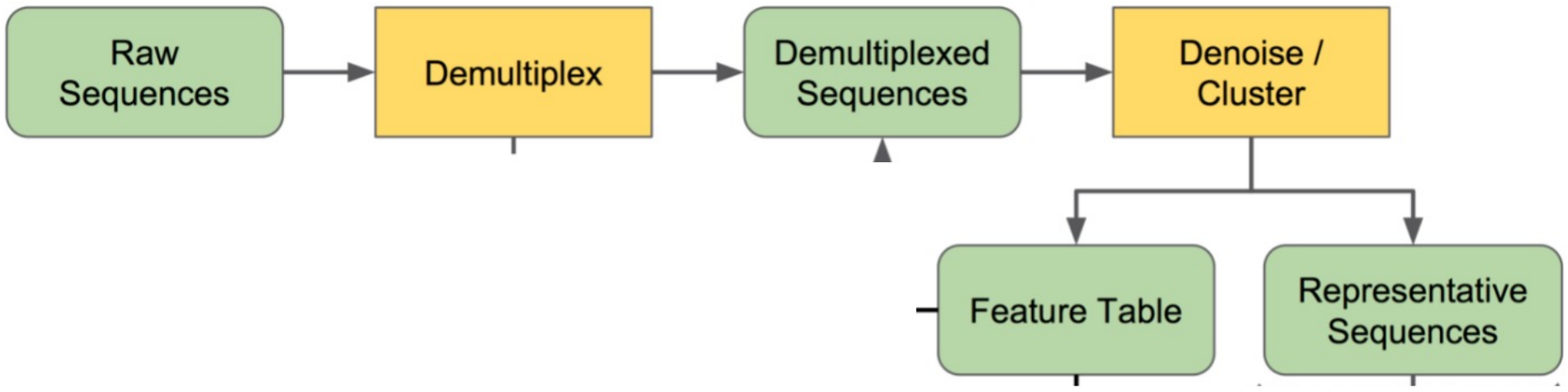
1. Performed with DADA2 or Deblur
2. Output: ASVs
3. Percentage of denoised, non-chimeric reads

Requires installation but can be installed on a linux, windows or mac.

Conceptual overview of QIIME 2

Now that we have read the glossary and key, let us examine a conceptual overview of the various possible workflows for examining amplicon sequence data:

https://docs.qiime2.org/2022.2/tutorials/overview/



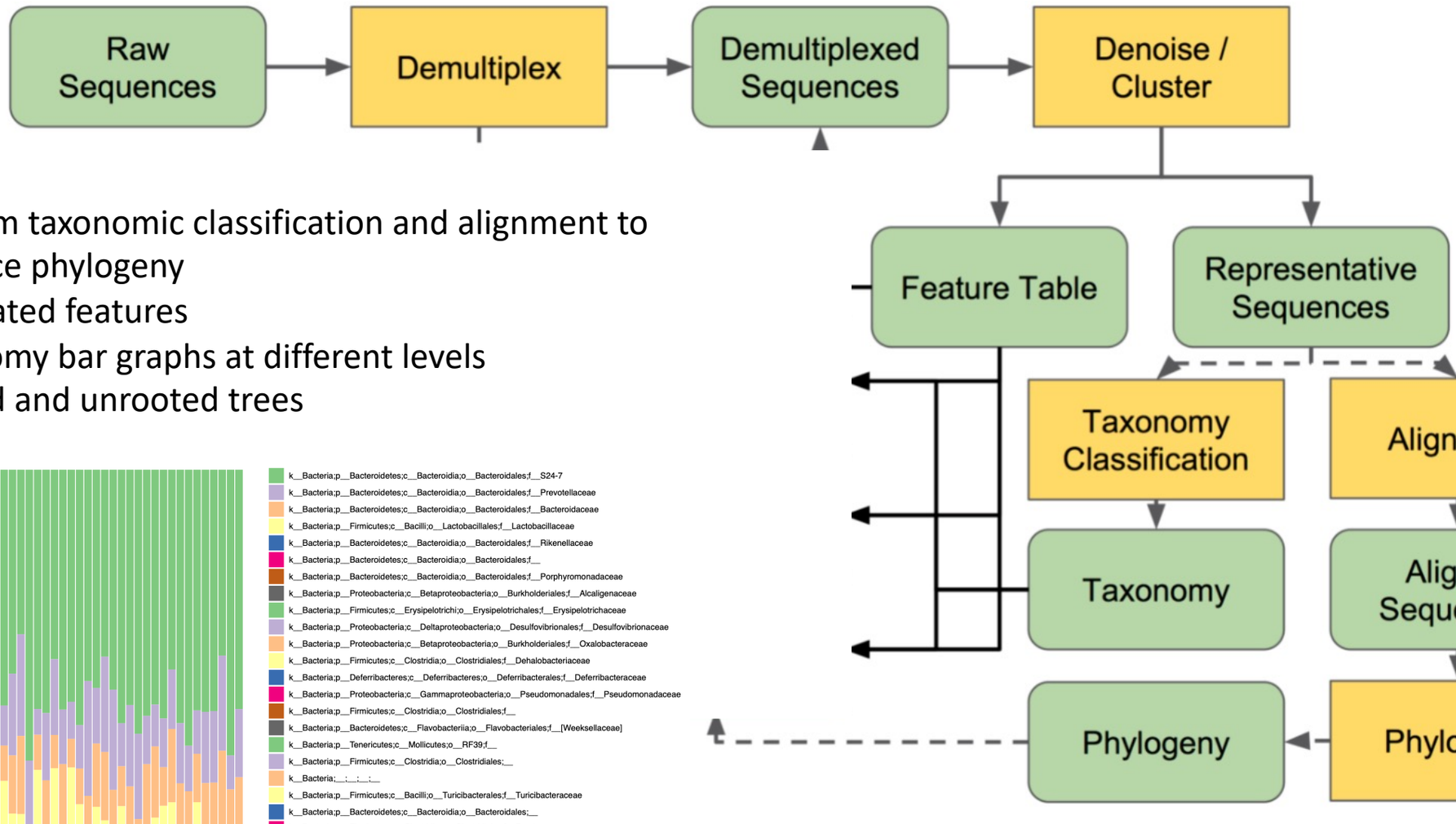
1. Use ASVs or classify the data onto OTUs with vsearch

Requires installation but can be installed on a linux, windows or mac.

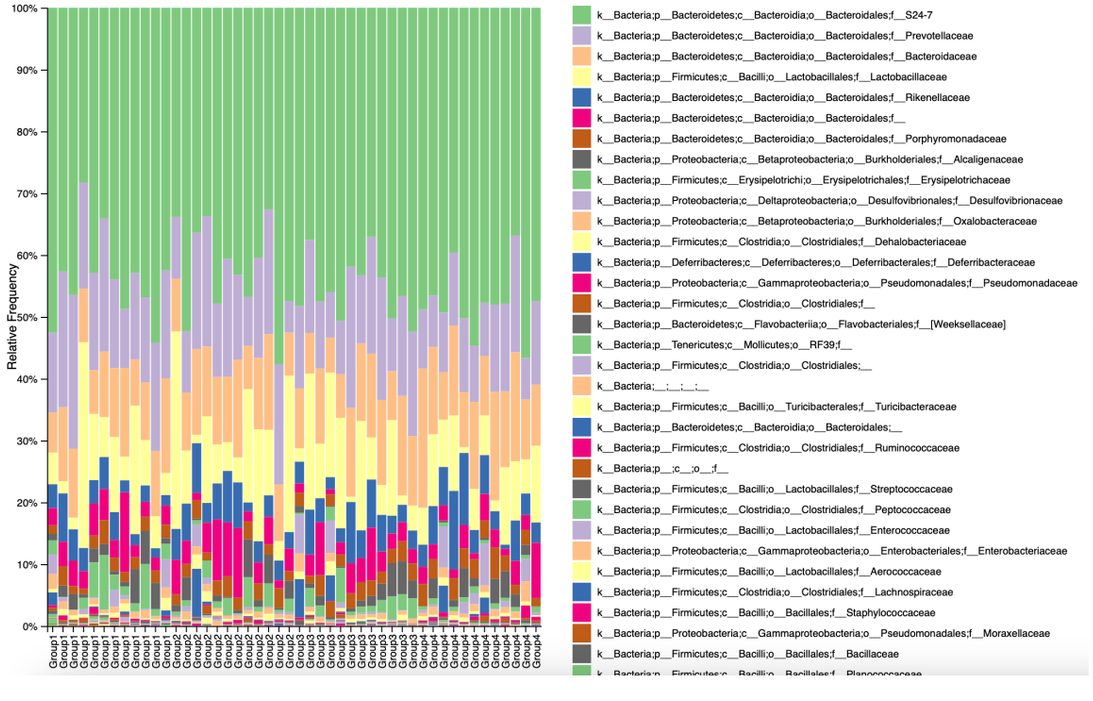
Conceptual overview of QIIME 2

Now that we have read the glossary and key, let us examine a conceptual overview of the various possible workflows for examining amplicon sequence data:

https://docs.qiime2.org/2022.2/tutorials/overview/



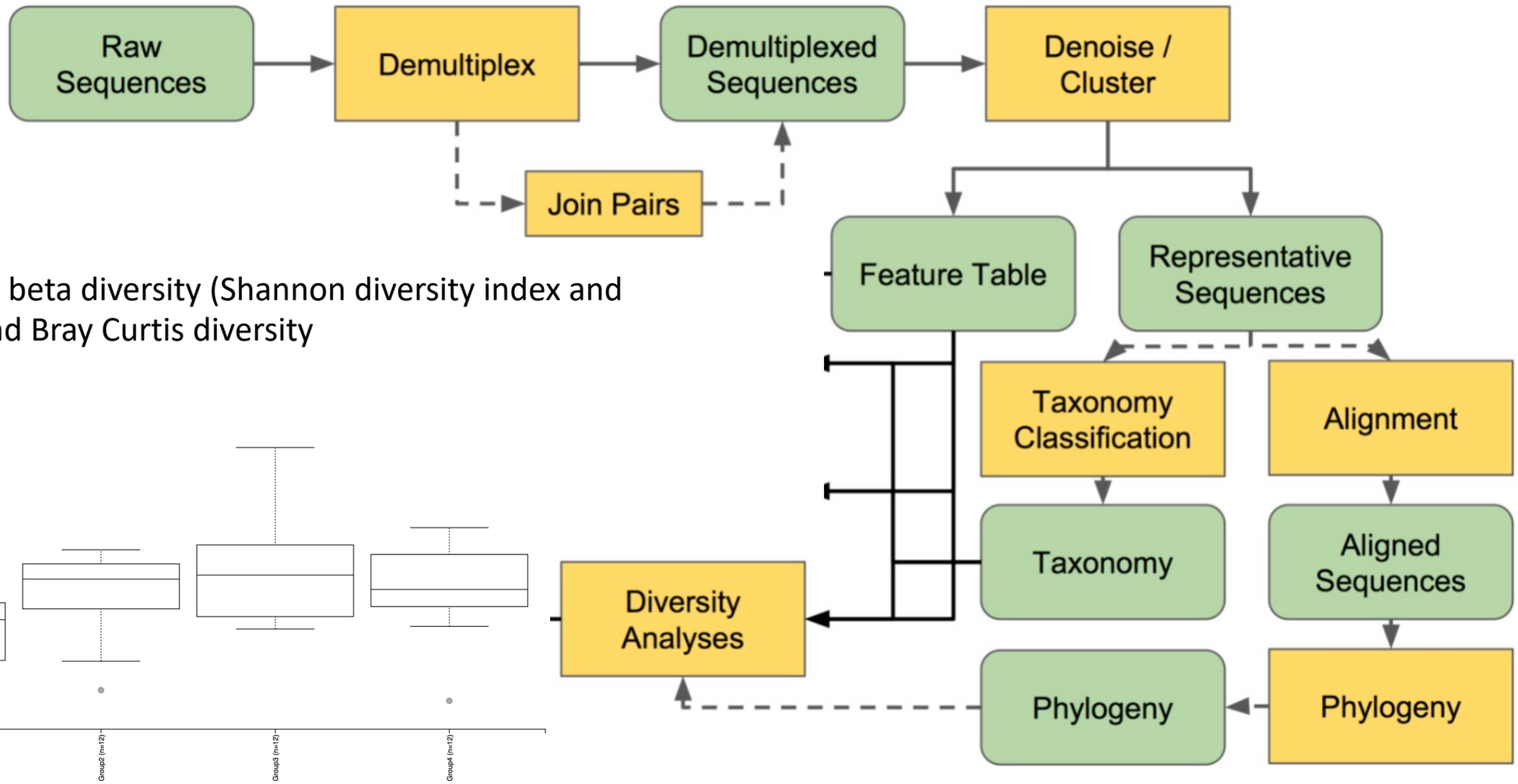
1. Perform taxonomic classification and alignment to produce phylogeny
2. Annotated features
3. Taxonomy bar graphs at different levels
4. Rooted and unrooted trees



Conceptual overview of QIIME 2

Now that we have read the glossary and key, let us examine a conceptual overview of the various possible workflows for examining amplicon sequence data:

https://docs.qiime2.org/2022.2/tutorials/overview/

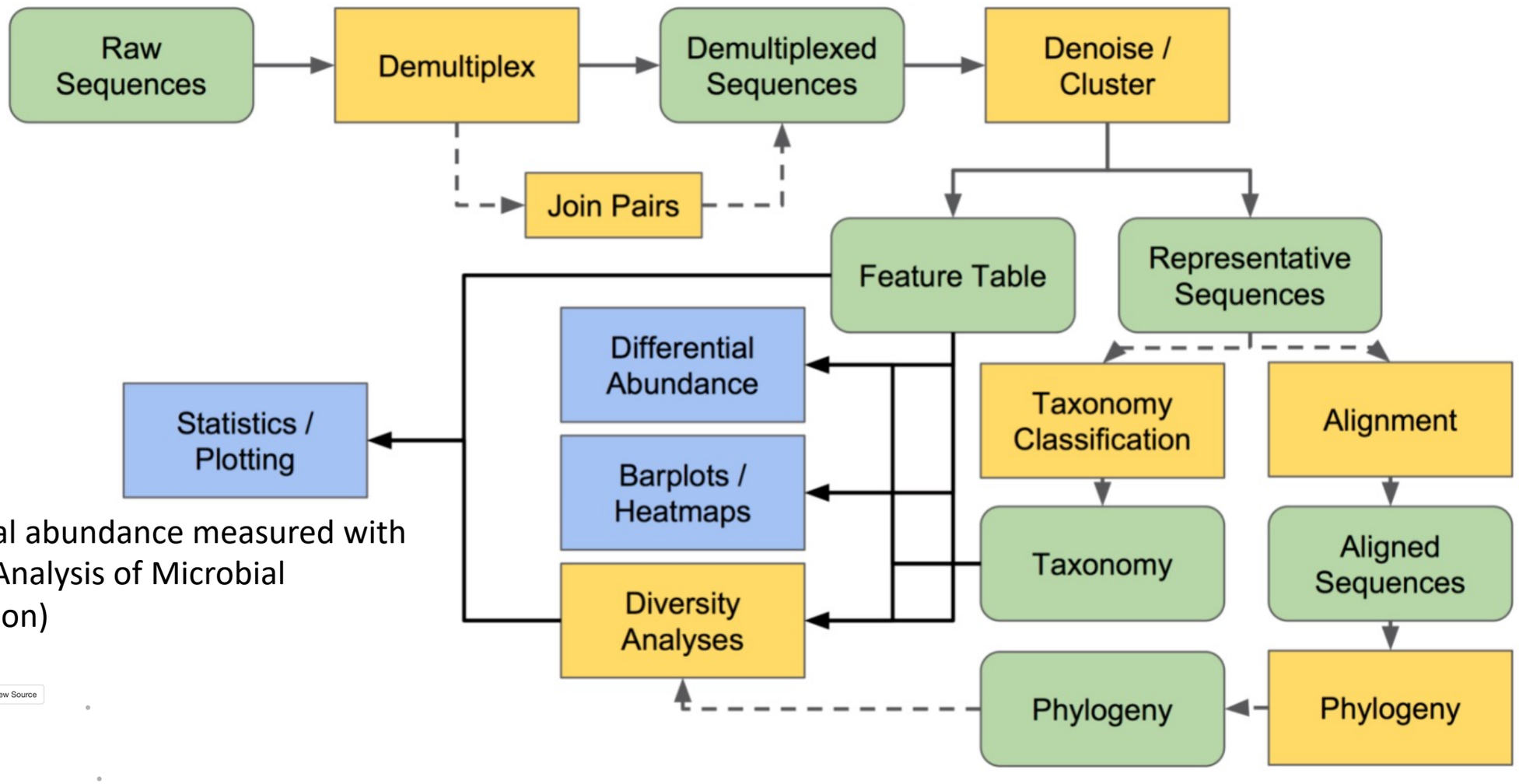


1. Alpha and beta diversity (Shannon diversity index and Jaccard and Bray Curtis diversity

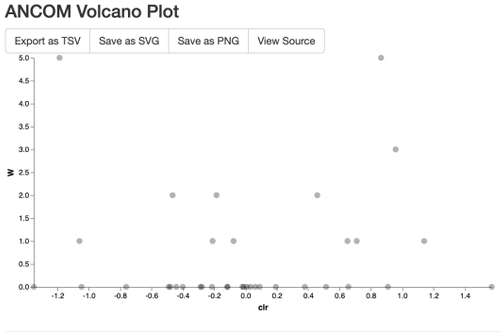
Conceptual overview of QIIME 2

Now that we have read the glossary and key, let us examine a conceptual overview of the various possible workflows for examining amplicon sequence data:

https://docs.qiime2.org/2022.2/tutorials/overview/



1. Differential abundance measured with ANCOM (Analysis of Microbial Composition)



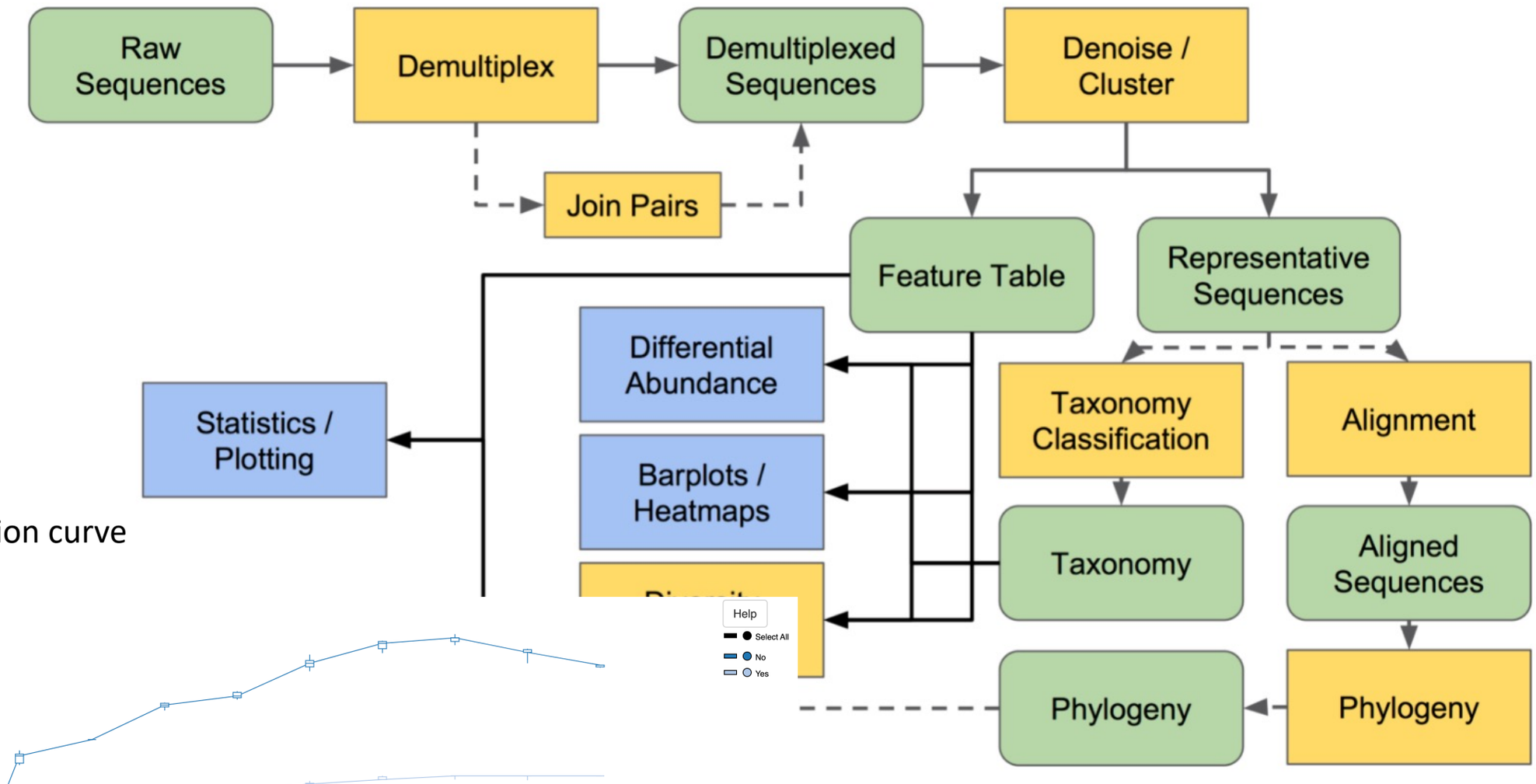
ANCOM statistical results

	W
62dc76a09326c359be5946507e82e9fa	5

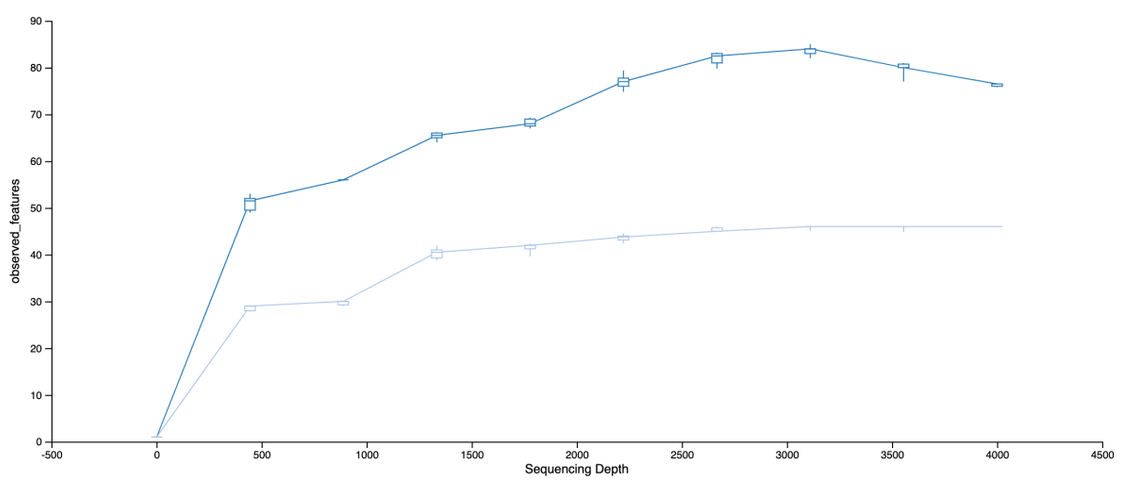
Conceptual overview of QIIME 2

Now that we have read the glossary and key, let us examine a conceptual overview of the various possible workflows for examining amplicon sequence data:

https://docs.qiime2.org/2022.2/tutorials/overview/



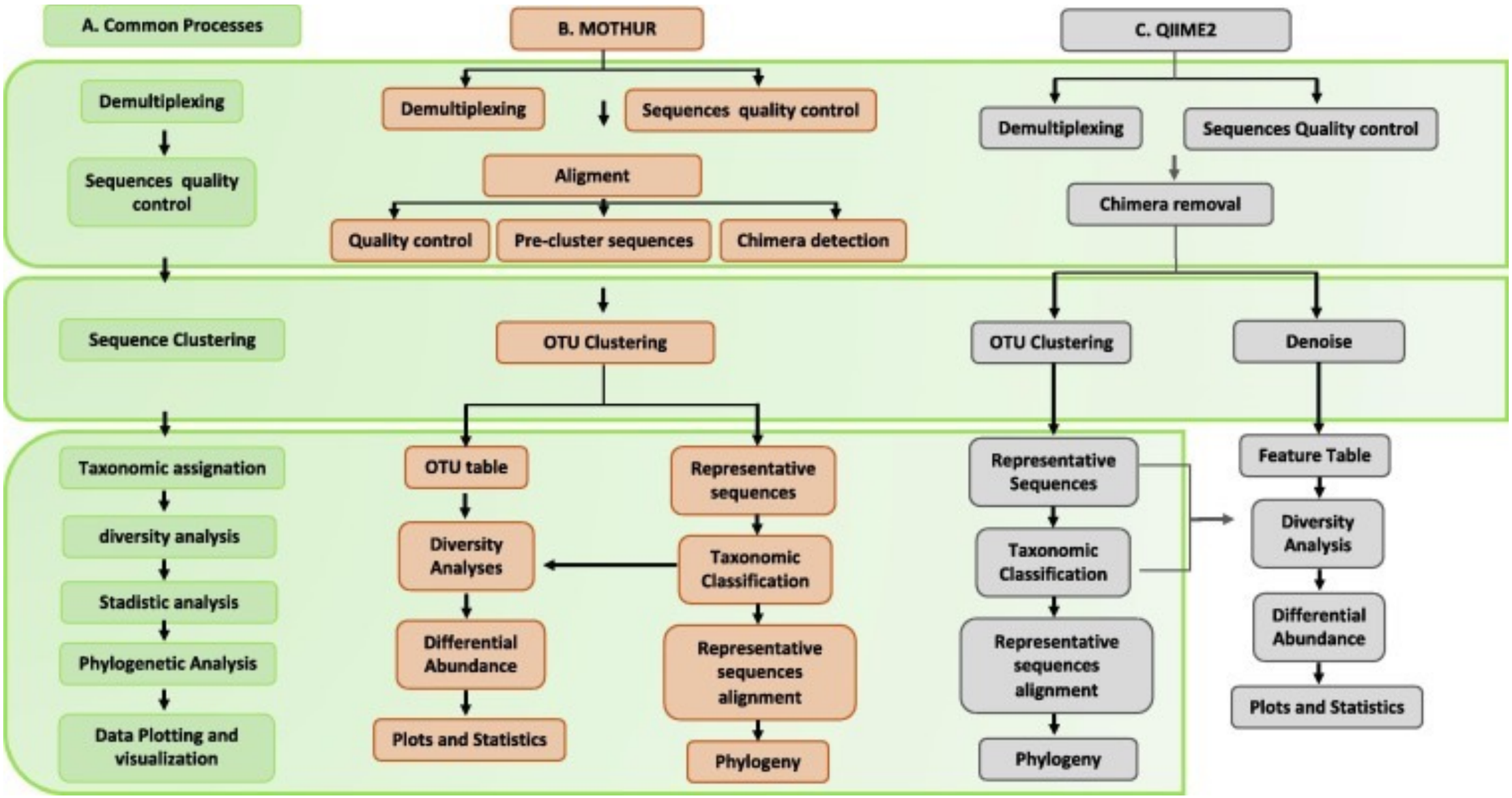
1. Rarefaction curve



A quick outlook at some other tools

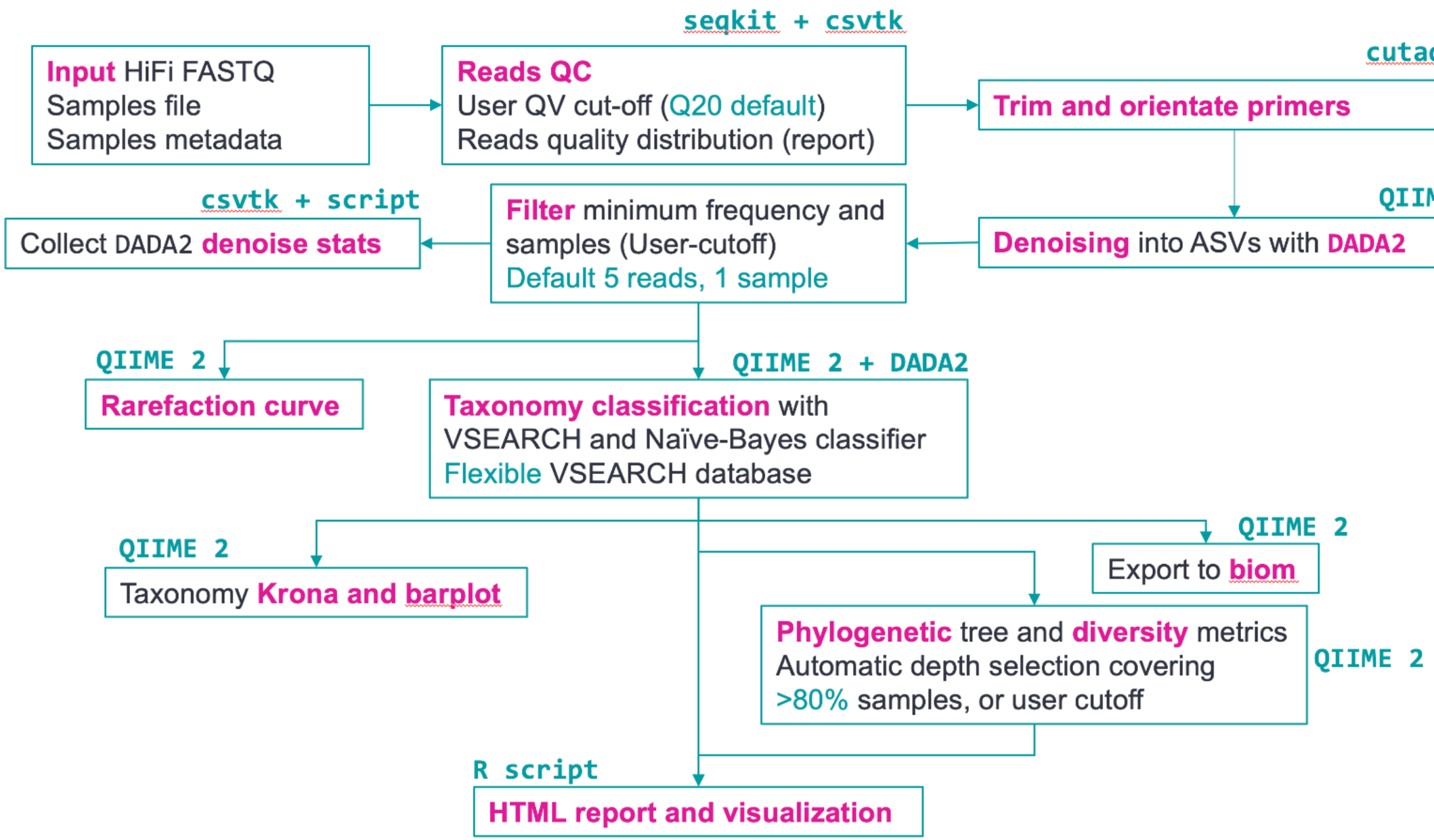
USEARCH/UPARSE	Mothur	DADA2 within QIIME2 Divisive Amplicon Denoising Algorithm
Just a binary, super easy	Single program with minimal dependencies (read easy to install and setup)	Full R-based; easy to proceed with downstream analysis
Main clustering algorithm also an implementation of vsearch	Reimplementation of tried-and-tested algorithms (e.g., mapping, clustering)	Fast
High memory, multi-core server	High memory, multi-core server	Low memory, potentially parallelizable
Scalable (simple binary)	Not so much scalable (slows down with sample size)	Heavily affected by errors; requires more reads
https://www.drive5.com/usearch/manual/uparse_pipeline.html	https://www.mothur.org/wiki/MiSeq_SOP	https://benjjneb.github.io/dada2/tutorial.html

A quick outlook at some other tools



Santos et al. 2020 Computational methods for 16S metabarcoding studies using Nanopore sequencing data

16S long read sequencing



Example of statistical analysis – diversity estimates

Alpha Diversity

Shannon diversity index
measures the diversity of species in a community.

The formula: $H = -\sum(Pi) \times ln(Pi)$

where Pi = the proportion of individuals in each species

Simpson diversity index
measures the diversity of species in a community

$$D = \frac{N(N-1)}{\sum n(n-1)}$$

N = Total number of organisms

n = Population of each individual species

Beta Diversity

Bray–Curtis dissimilarity
- based on abundance/read count data
- differences in microbial abundances between two samples (e.g., at species level)
values are from 0 to 1

Jaccard distance
- based on presence or absence of species
- difference in microbial composition between two samples
0 means both samples share exact the same species
1 means both samples have no species in common

Example of statistical analysis – diversity estimates

Alpha Diversity

Shannon diversity index
measures the diversity of species in a community.

The formula: $H = -\sum(Pi) \times ln(Pi)$

where *Pi* = the proportion of individuals in each species

Pielou's evenness – also known as equitability

Shannon diversity divided by the logarithm of number of taxa

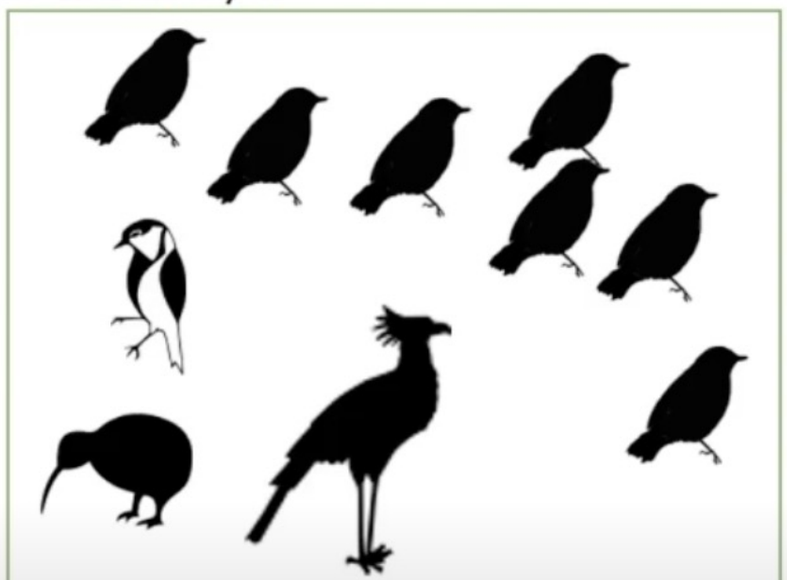
Beta Diversity

Bray–Curtis dissimilarity
- based on abundance/read count data
- differences in microbial abundances between two samples (e.g., at species level)
values are from 0 to 1

Jaccard distance
- based on presence or absence of species
- difference in microbial composition between two samples
0 means both samples share exact the same species
1 means both samples have no species in common

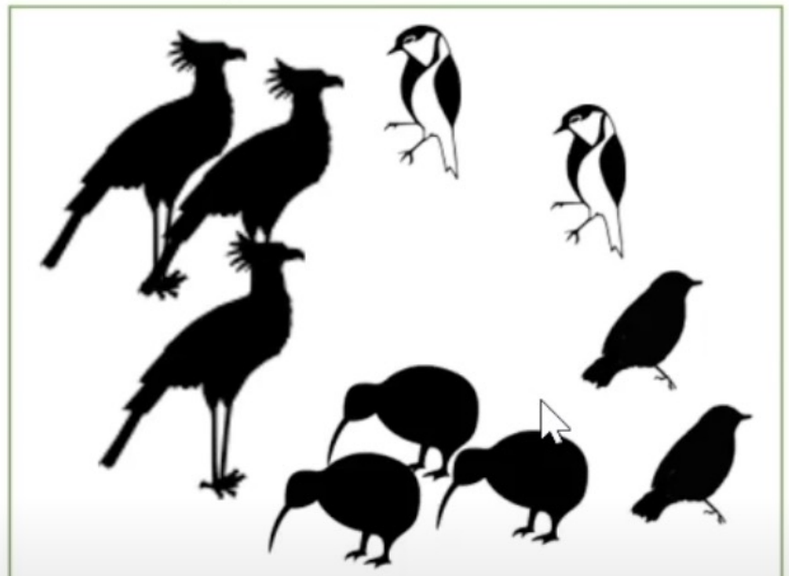
Simpson vs Shannon diversity index

Community A



Abundance = 10
Species Richness = 4
Diversity = ?

Community B



Abundance = 10
Species Richness = 4
Diversity =

Species	n	n-1	n(n-1)
A			
B			
C			
D			
Total	10		

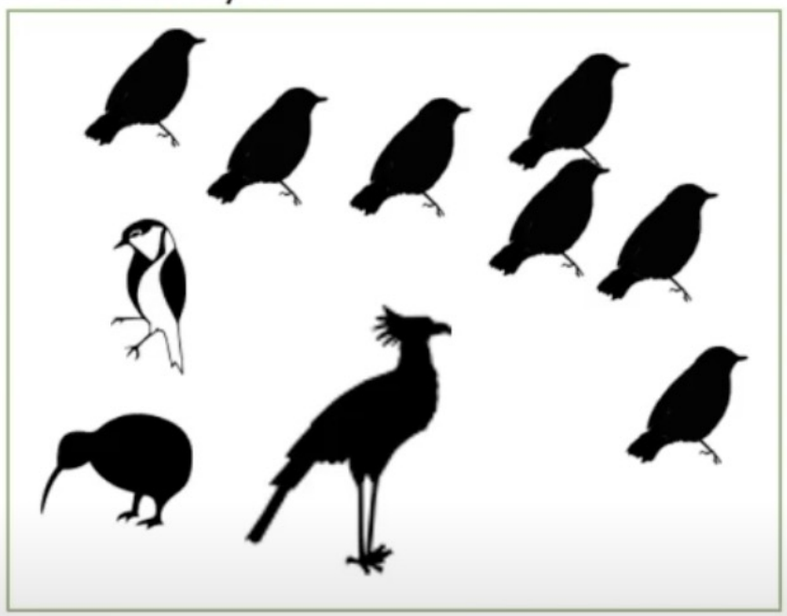
$$D = \frac{N(N-1)}{\sum n(n-1)}$$

N = Total number of organisms

n = Population of each individual species

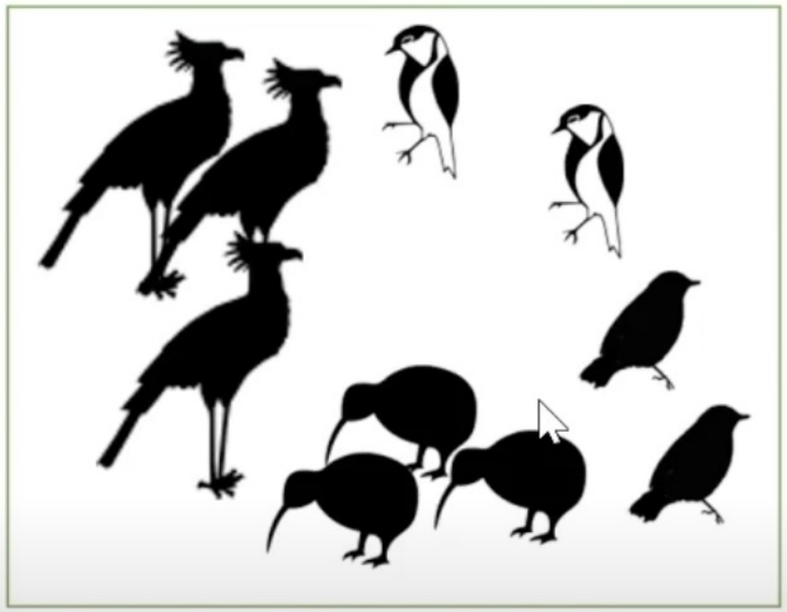
Simpson vs Shannon diversity index

Community A



Abundance = 10
Species Richness = 4
Diversity = ?

Community B



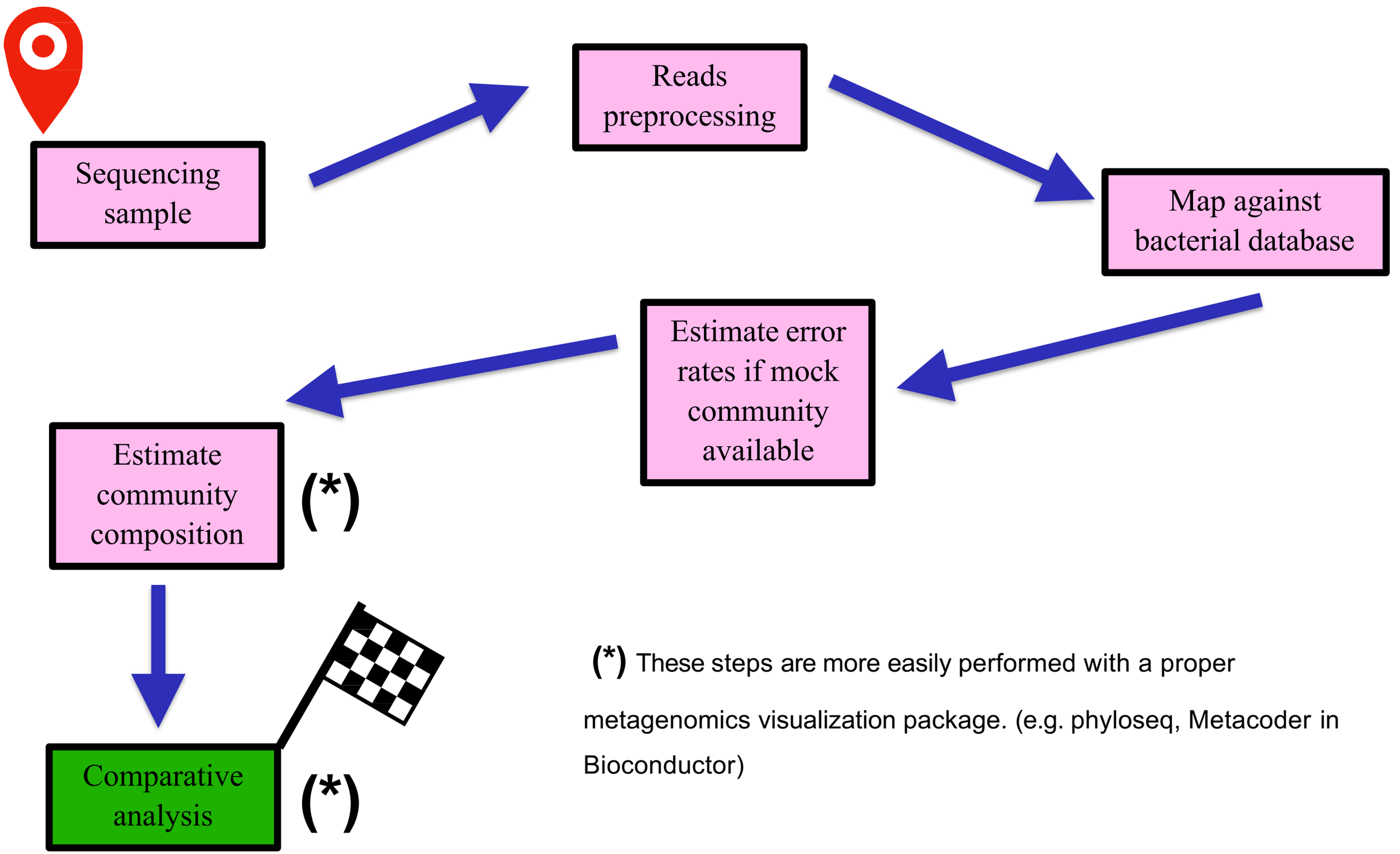
Abundance = 10
Species Richness = 4
Diversity =

Species	n	Pi	ln(Pi)	Pi x ln(Pi)
A				
B				
C				
D				
Total	10			

The formula: $H = -\sum(Pi) \times ln(Pi)$

where *Pi* = the proportion of individuals in each species

16S example workflow and associated challenges

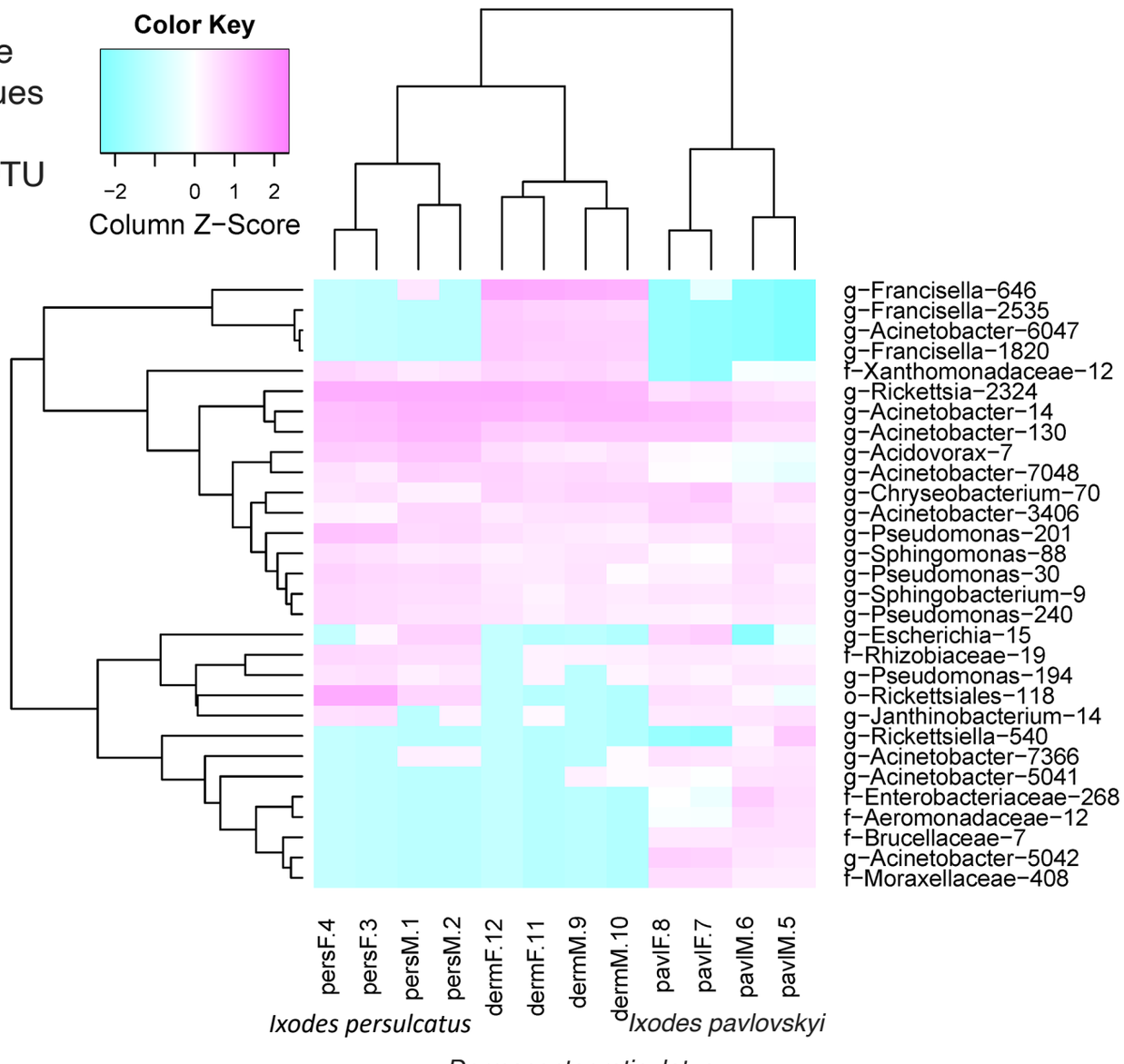


(*) These steps are more easily performed with a proper metagenomics visualization package. (e.g. phyloseq, Metacoder in Bioconductor)

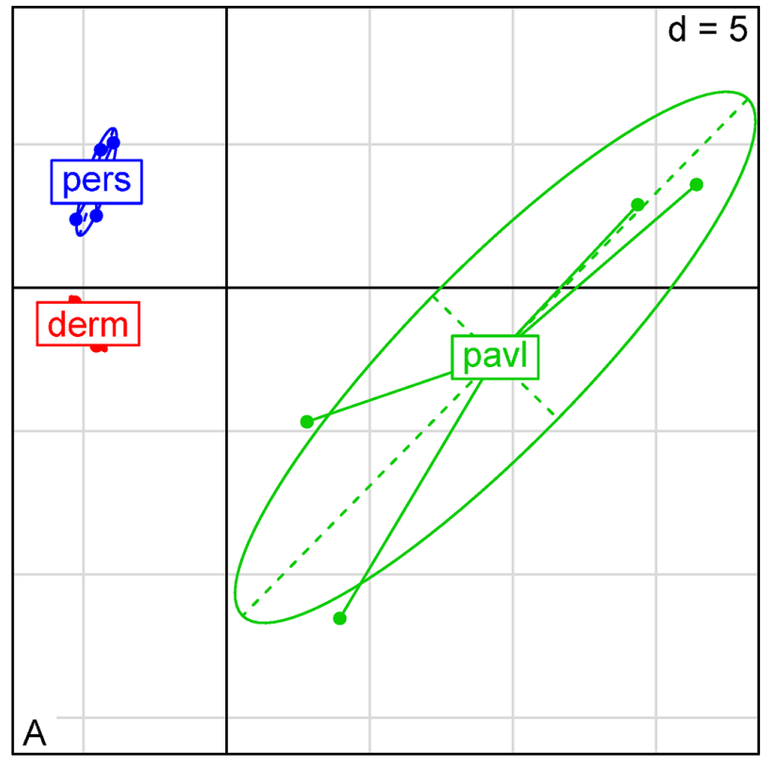
Comparative analysis: is there a host phenotype/environmental condition - community composition association?

Based on euclidean distance method for log-scaled in-sample taxon share values

(pink – higher OTU abundance)



Beta diversity – UniFRAC distance



Kurilshikov et al. 2015

Outline

- **File formats**
- **Processed data**
- **BIOM format**

Processed data: 16S

- The outcome of processed data consists of 3 information:
  - OTUs abundance matrix
  - OTU classification
  - Sample metadata

OTU abundance matrix can be dense or sparse

A dense representation of an OTU table:

OTU ID	PC.354	PC.355	PC.356
OTU0	0	0	4
OTU1	6	0	0
OTU2	1	0	7
OTU3	0	0	3

A sparse representation of an OTU table:

```
PC.354 OTU1 6
PC.354 OTU2 1
PC.356 OTU0 4
PC.356 OTU2 7
PC.356 OTU3 3
```

- Dense matrices contain zeros for OTUs not observed in certain samples
- Sparse do not; lot of space saved

Taxonomical classification

- Can be contained in the OTU matrix…

OTU ID	PC.354	PC.355	PC.356	Taxonomy
OTU0	0	0	4	Bacteria;Firmicutes;Bacilli;
OTU1	6	0	0	Bacteria;Firmicutes;Bacilli;
OTU2	1	0	7	Bacteria;"Proteobacteria";Gammaproteobacteria
OTU3	0	0	3	Bacteria;"Proteobacteria";Gammaproteobacteria

- Or be stored in a different file

OTU	Size	Taxonomy
Otu0001	2024	Bacteria;Firmicutes;Bacilli;Bacillales;Bacillaceae_1;Bacillus;
Otu0002	1552	Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Staphylococcus;
Otu0003	1207	Bacteria;Firmicutes;Bacilli;Bacillales;Listeriaceae;Listeria;
Otu0004	874	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus;
Otu0005	752	Bacteria;Firmicutes;Bacilli;Lactobacillales;Enterococcaceae;Enterococcus;

Sample information

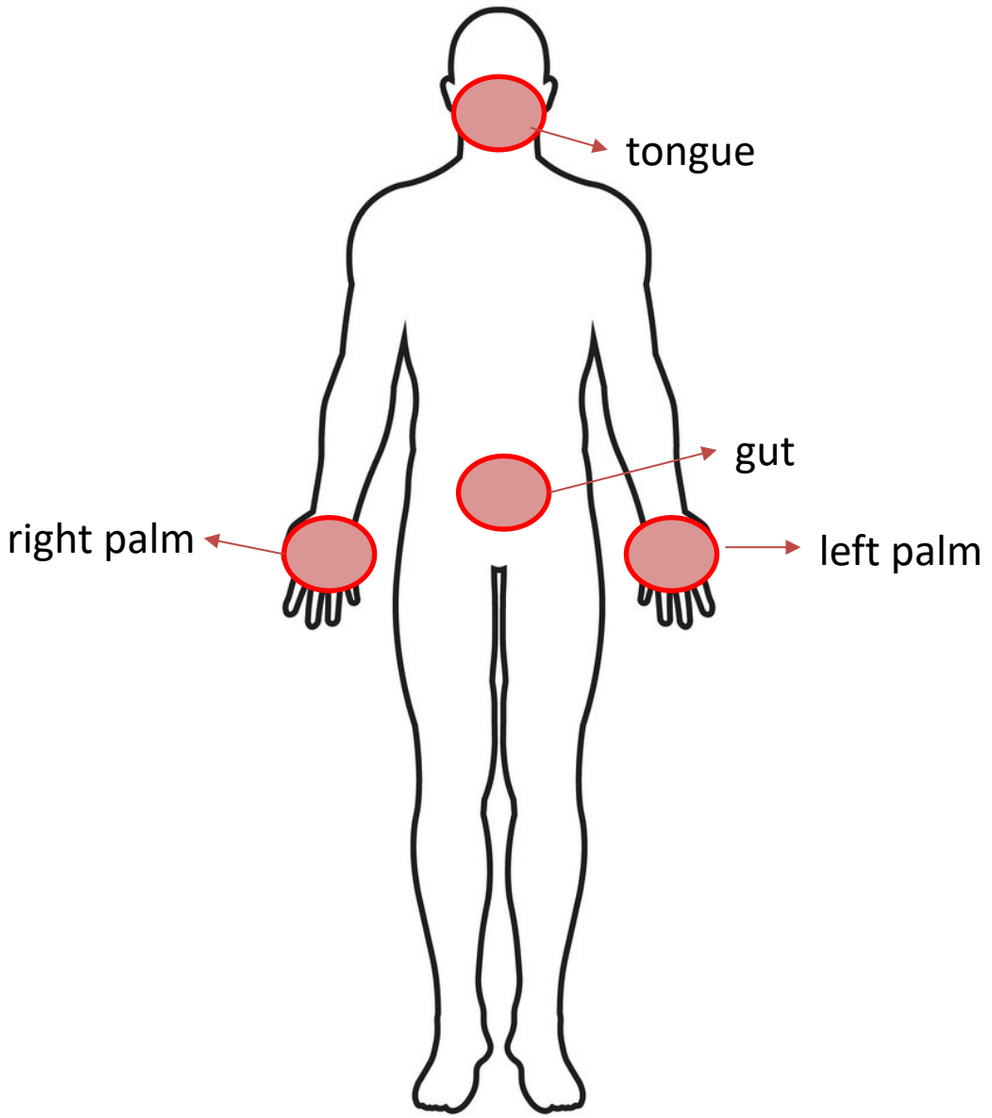
- Typical design matrix

Name	Technology [Factor]	Group [Factor]
post1	Illumina	post
post2	Illumina	post
post3	Illumina	post
pre1	Illumina	pre
pre2	Illumina	pre
pre3	Illumina	pre

Let's do a practical

- Earth Microbiome Project
- sequenced hypervariable region 4 **(V4) 16S rRNA**
- analysis of human microbiome samples
- Original study is from Copraso et al. 2011
- Analysed with Quantitative Insights Into Microbial Ecology (QIIME2)

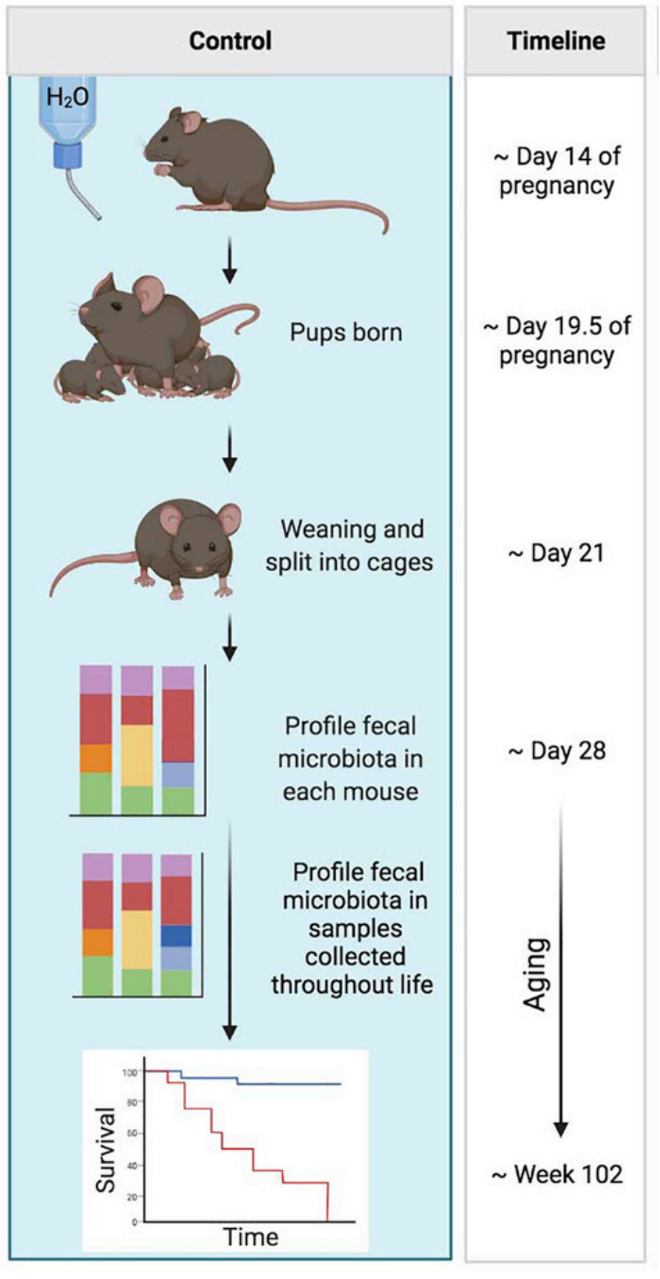
2 individuals: with/without antibiotics
396 time points (here 5 times points)



- https://pubmed.ncbi.nlm.nih.gov/21624126/

Let's do a practical

- Schloss mice
- sequenced hypervariable region 4 **(V4) 16S rRNA**
- analysis of mice gut microbiome post-weaning
- Original study is from Schloss et al. 2012
- Analysed with Mothur



- https://www.tandfonline.com/doi/full/10.4161/gmic.21008

Lets do some work!

Go to https://github.com/zajacn/metagenomics_course_FGCZ

Scroll to **QIIME2 analysis on SUSHI**