



16S rRNA: data analysis

Dr. Natalia Zajac
Metagenomics, 03.2023





Microbes are everywhere

We live in a microbial world.

Coprinus comatus



Zea mays



Homo sapiens



Porphyra yezoensis

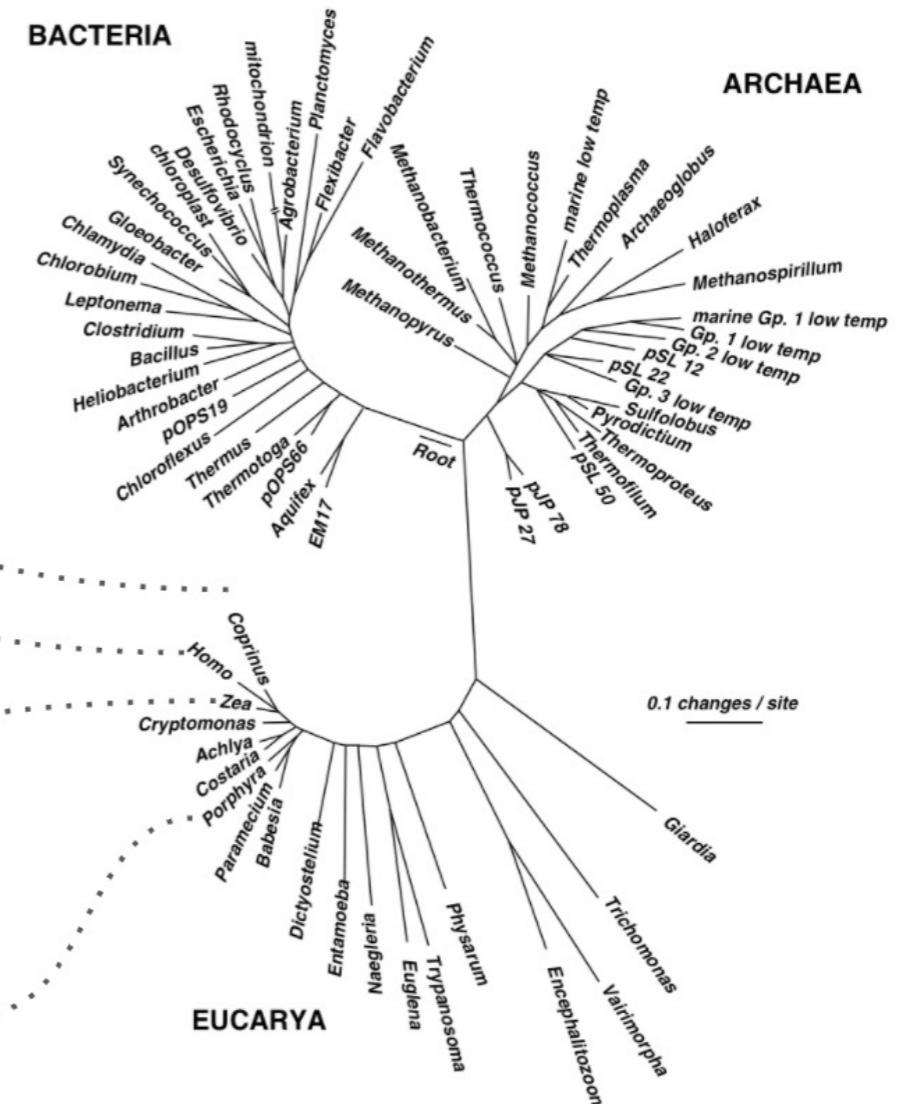


Image sources:

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.



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

Bertrand Routy,^{1,2,3} Emmanuelle Le Chatelier,⁴ Lisa Derosa,^{1,2,5} Connie P. M. Duong,^{1,2,5} Maryam Tidjani Alou,^{1,2,5} Romain Daillère,^{1,2,3} Aurélie Fluckiger,^{1,2,5} Meriem Messaoudene,^{1,2} Conrad Rauber,^{1,2,5} Maria P. Roberti,^{1,2,5} Marine Fidelle,^{1,2,5} Caroline Flament,^{1,2,5} Véronique 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,^{7,2,3} Bo Qu,^{1,2,5} Gladys Ferrere,^{1,2,5} Céline Clémenson,^{1,13} Laura Mezquita,^{1,14} Jordi Remon Masip,^{1,14} Charles Nalét,¹⁵ Solenn Brosseau,¹⁵ Courche Kaderbhai,¹⁶ Corentin Richard,¹⁶ Hira Rizvi,¹⁷ Florence Levenez,⁶ Nathalie Galleron,⁴ Benoit Quinquie,⁴ Nicolas Pons,⁴ Bernhard Ryffel,¹⁸ Véronique Minard-Colin,^{1,19} Patrick Gonin,^{1,20} Jean-Charles Soria,^{1,16} 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,^{1,2,14}

cancer immunotherapy (see the consumption is associated with poor profiled 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...



10
01
01
101010 01
101 10
010 01
010 01f. g. c. z.
01 1
10 0
+ 01 1

Outline

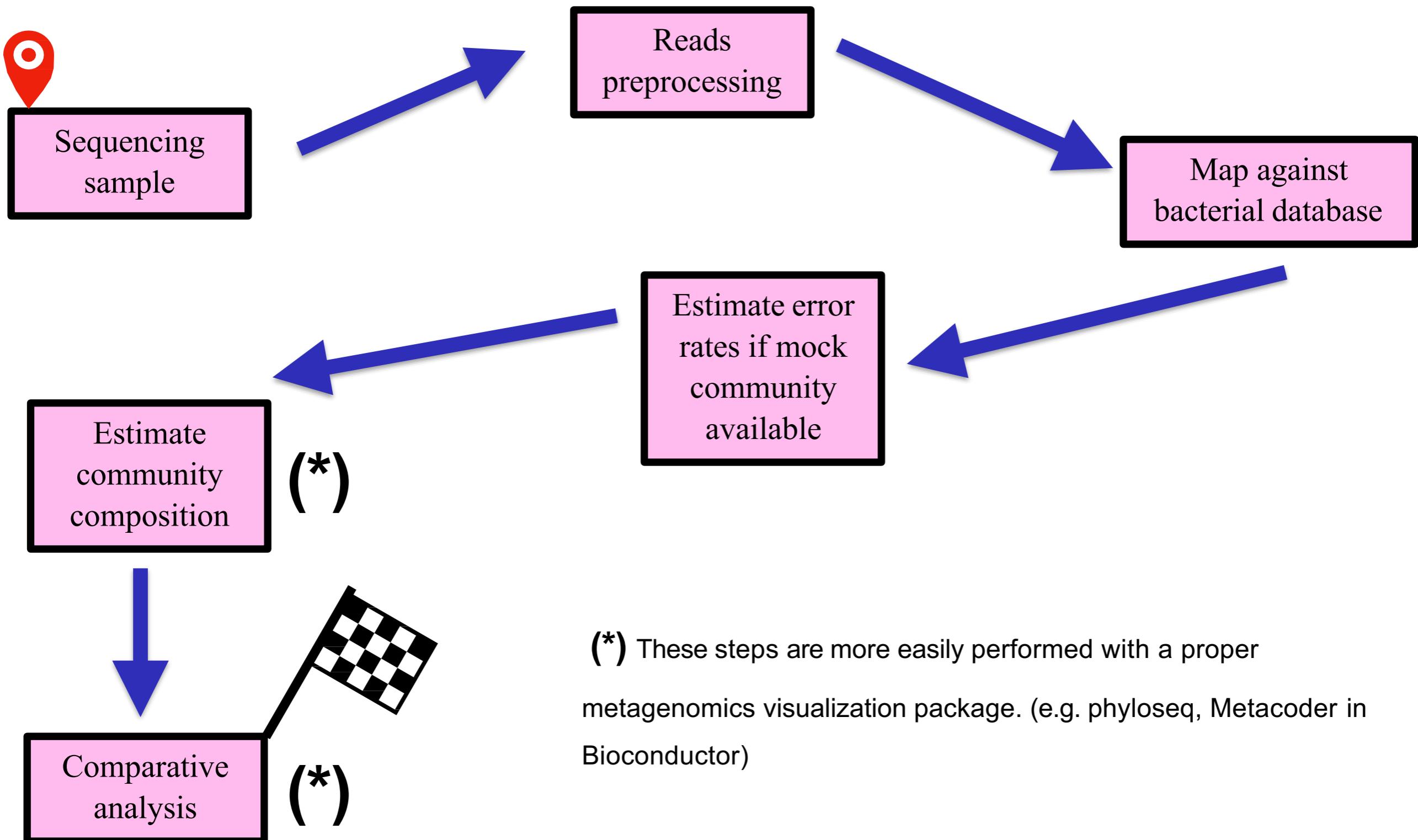
- Main goals
- Main challenges
- Approaches
- Databases
- File format

10
01
01
101010 01
101 10
010 01
010 01f. g. c. z.
01 1
10 0
01 1

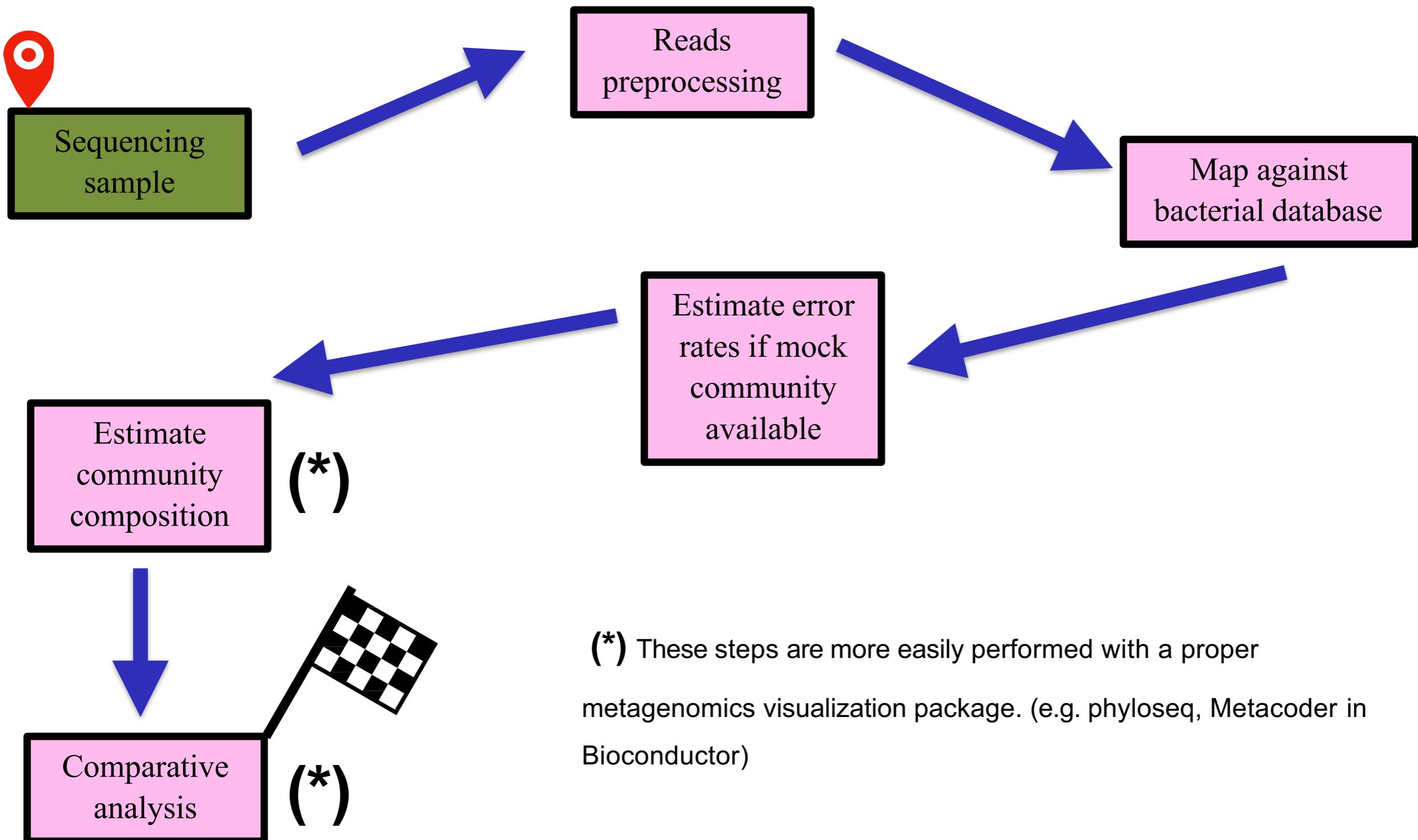
Ribosomal RNA

- 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

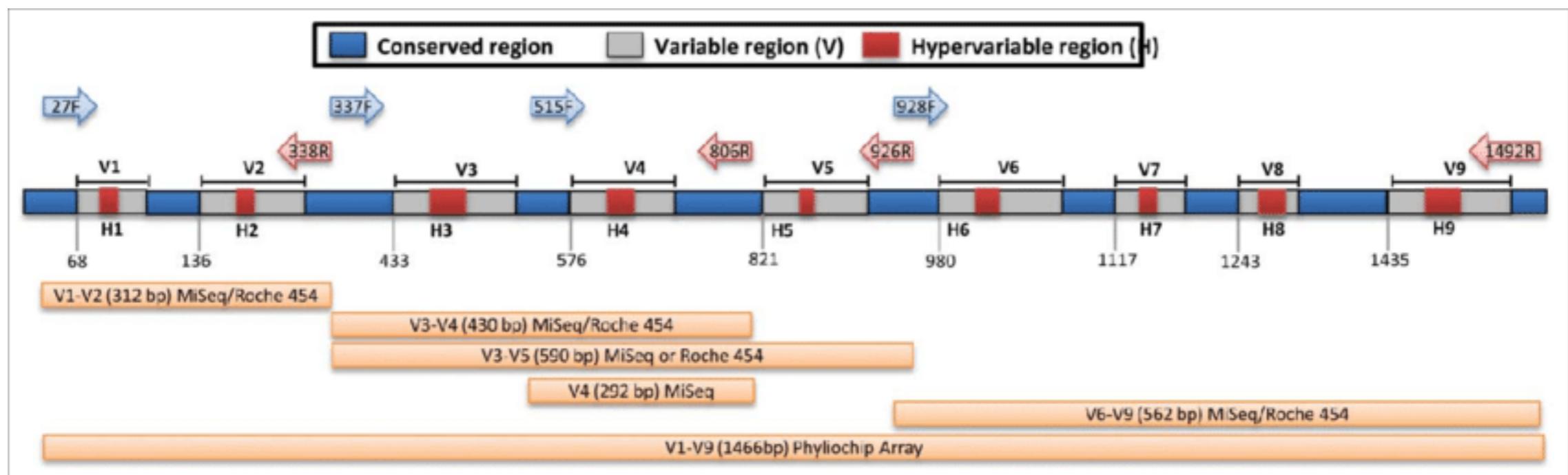


16S example workflow and associated challenges



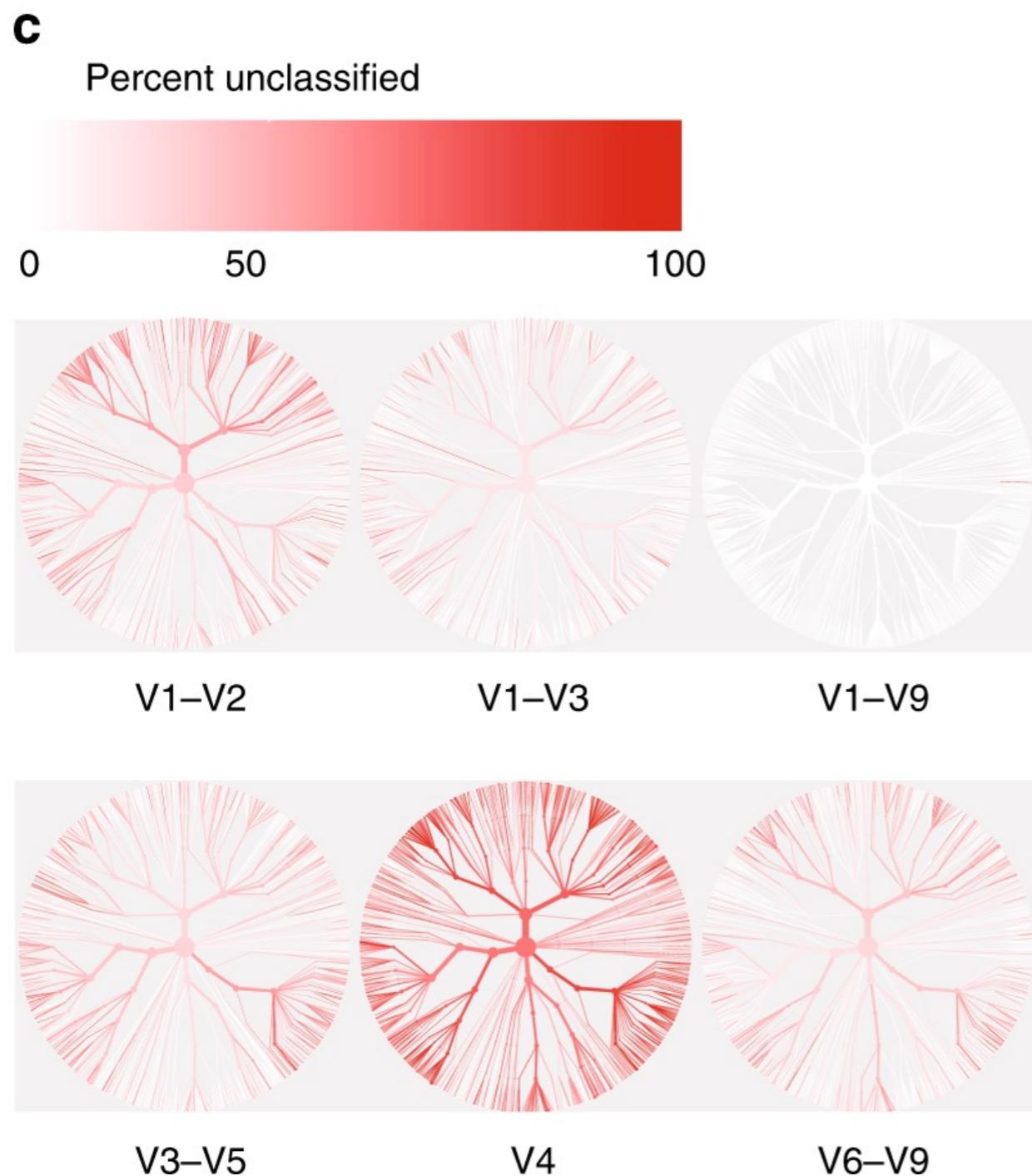


Sequencing sample



- 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?)

16S variable regions



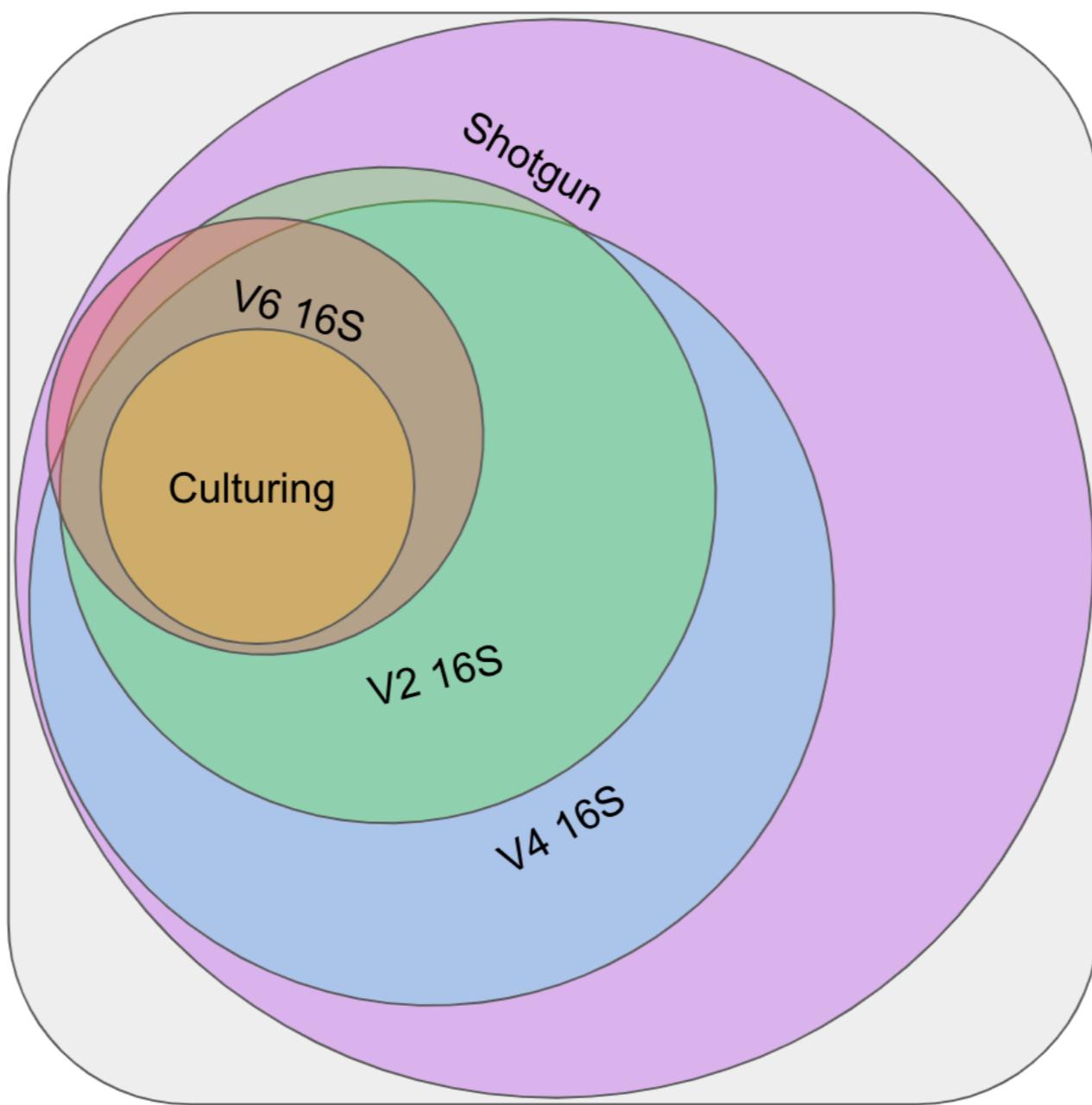
- V1–V2 region performed poorly at classifying sequences belonging to the phylum Proteobacteria
 - V3–V5 region performed poorly at classifying sequences belonging to the phylum Actinobacteria but great in identification of *Klebsiella*

Johnson et al. 2019

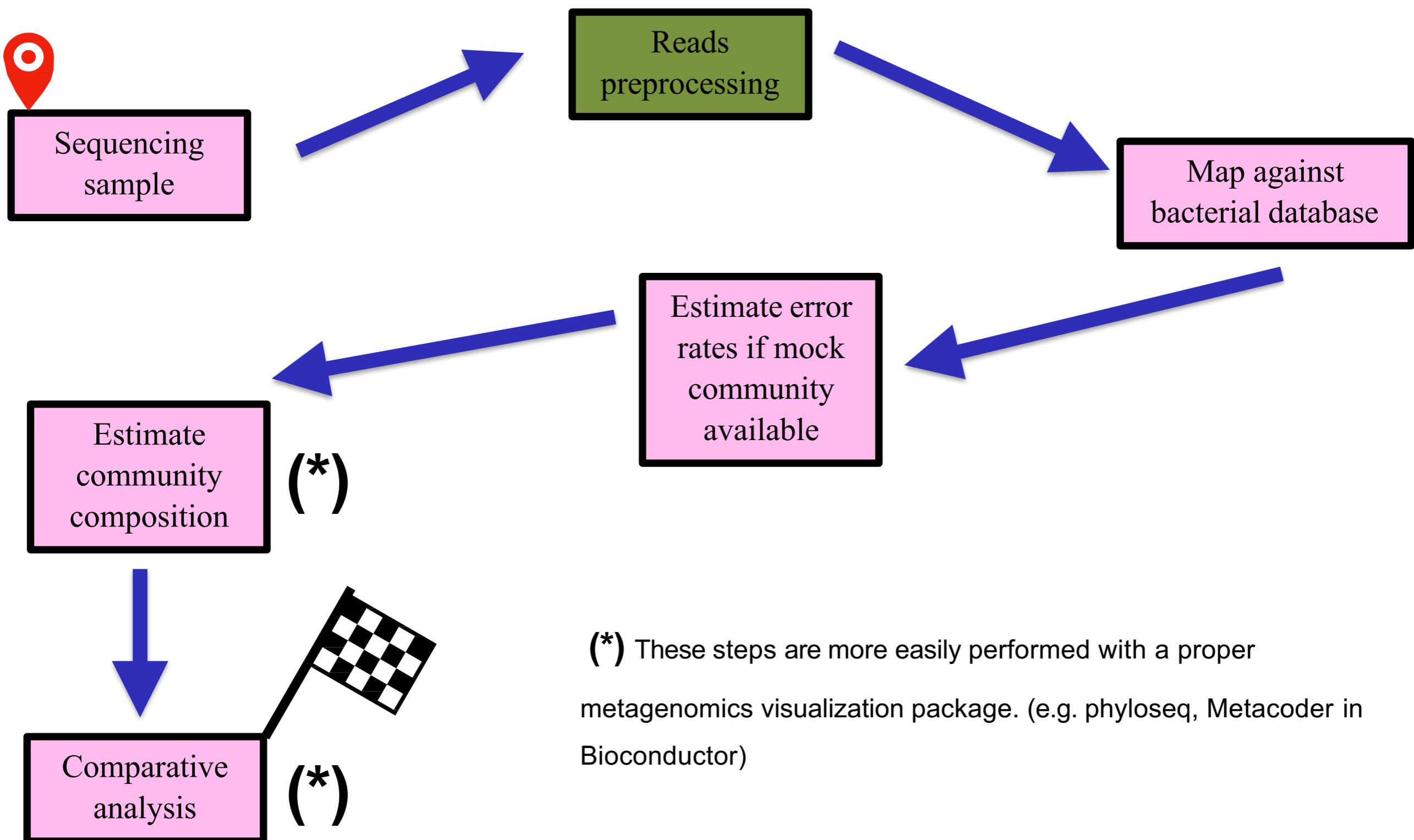
10
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
10110
01
101010
101
101
01001
10
01010
0101
10
01010
0101
10
01010
0101
10
01010
0101
10
01010
01010
01

Resolution

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



16S example workflow and associated challenges



10
01
01
101101 1
010 0
0101 10010 01
101 10
010 01

f. g. c. z.

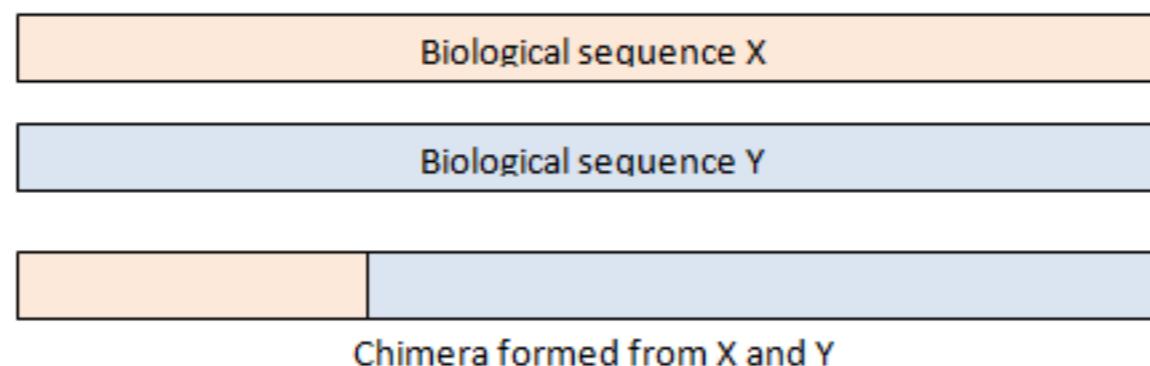
01 1
10 0
01 1
01 1

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?

As a general rule -> 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)
- (4) chimeras





Reads preprocessing

Step**Trim/Merge****Description**

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*

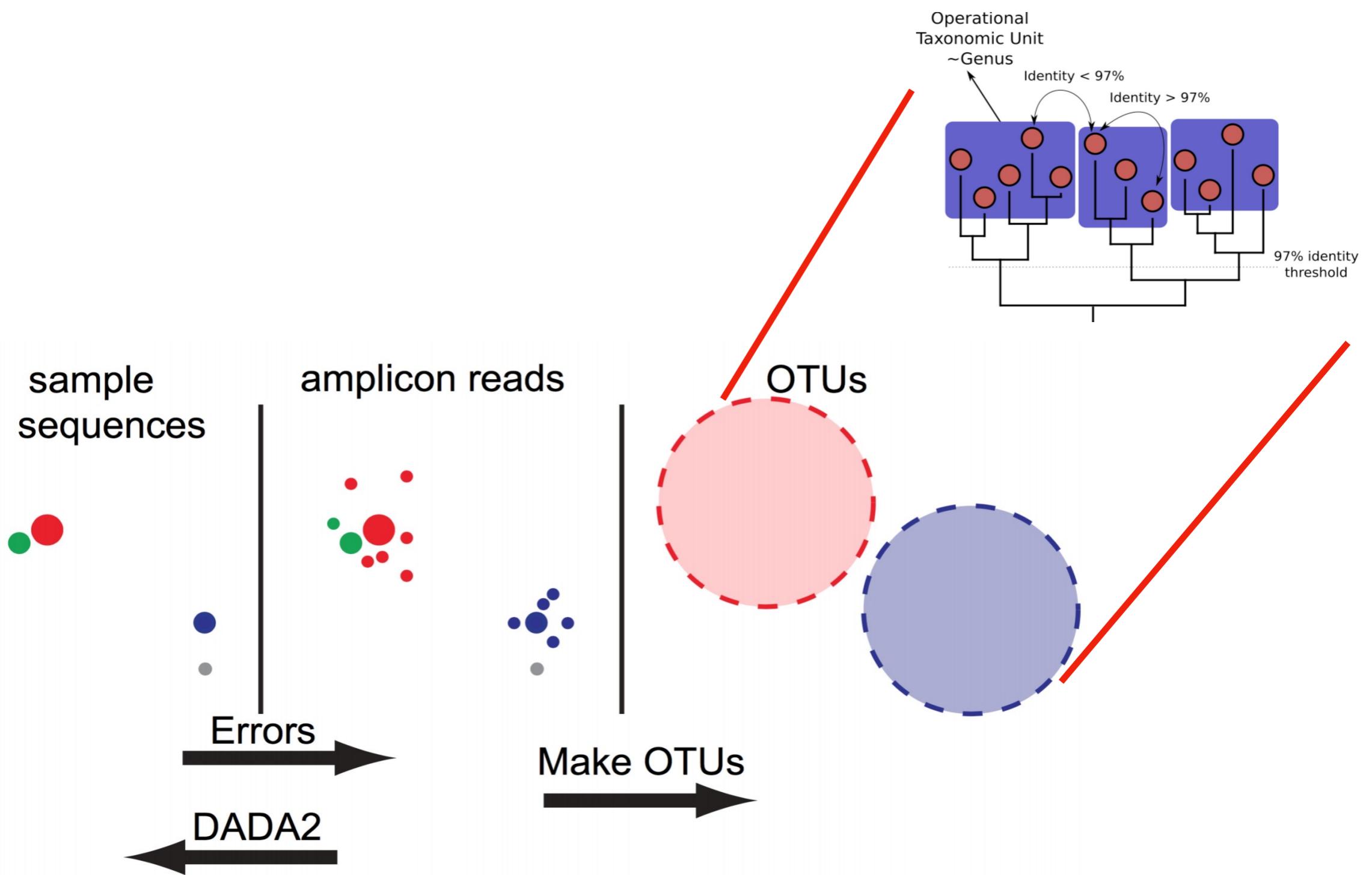
Discard all singletons, which are most likely chimeras, so set cluster size to >2

Make OTU table*

Cluster sequences to the same OTU by 97% sequence identity, using the plus strand

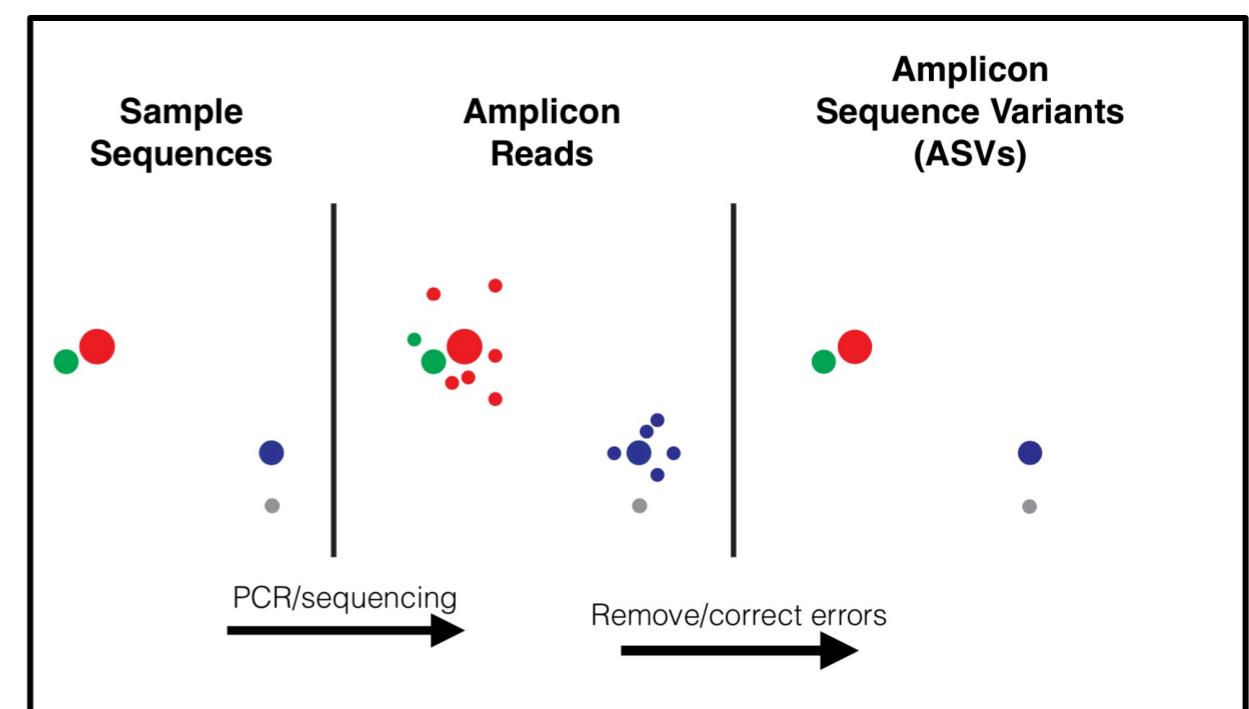
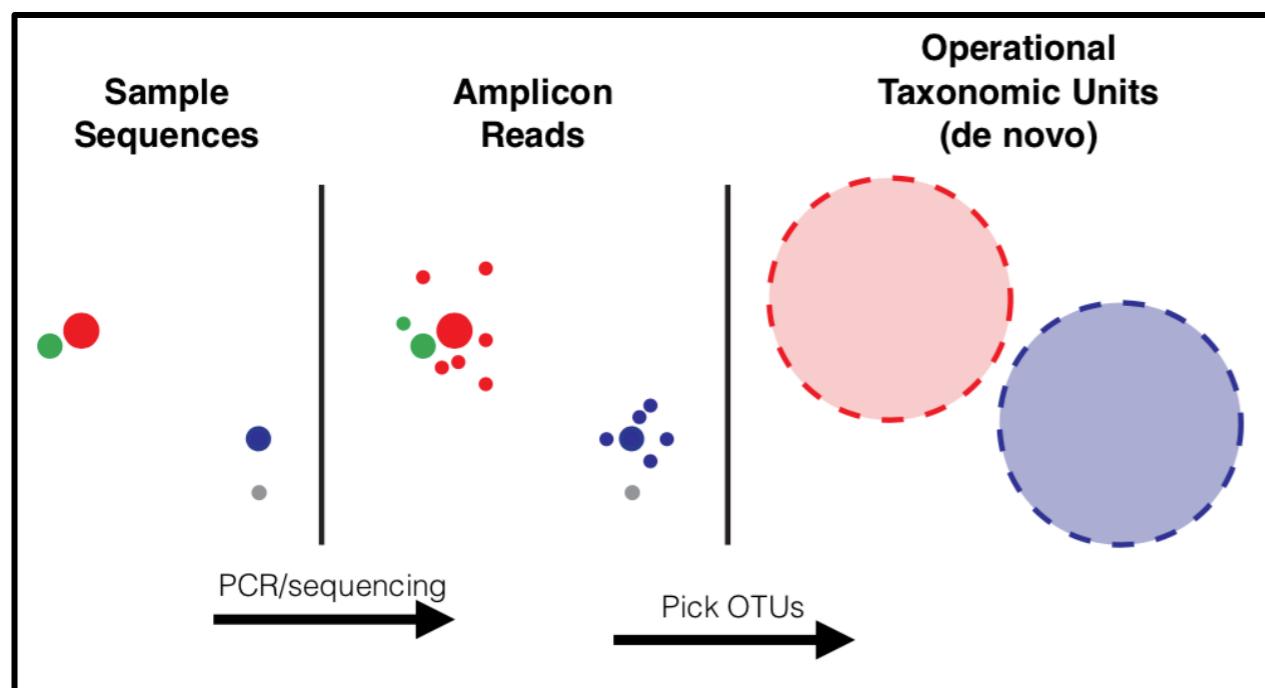
*optional

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





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



By Dr. Benjamin Callahan - Dr. Benjamin Callahan- powerpoint presentation, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=95743551>

10
01
01
101functional genomics center zurich
01 1
10 0
f. g. c. z.
01 1
01 1
10 0
+ 01 1

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

OTUs (e.g., Mothur)

Result of sequence clustering (typically at 97% identity)

Representative sequence obtained

Cannot distinguish very subtle strains

Needs fewer reads

Can get away with imperfect data

ASV (e.g., DADA2)

Only identical variants are clustered

Each cluster is a unique variant

Can distinguish complex communities

Requires more reads

Needs virtually error-free sequences

10
01
101101
1
010
0
0101
10010
01
101
10
010
01

f.g.c.z.

01
10
01
10
01

Clustering

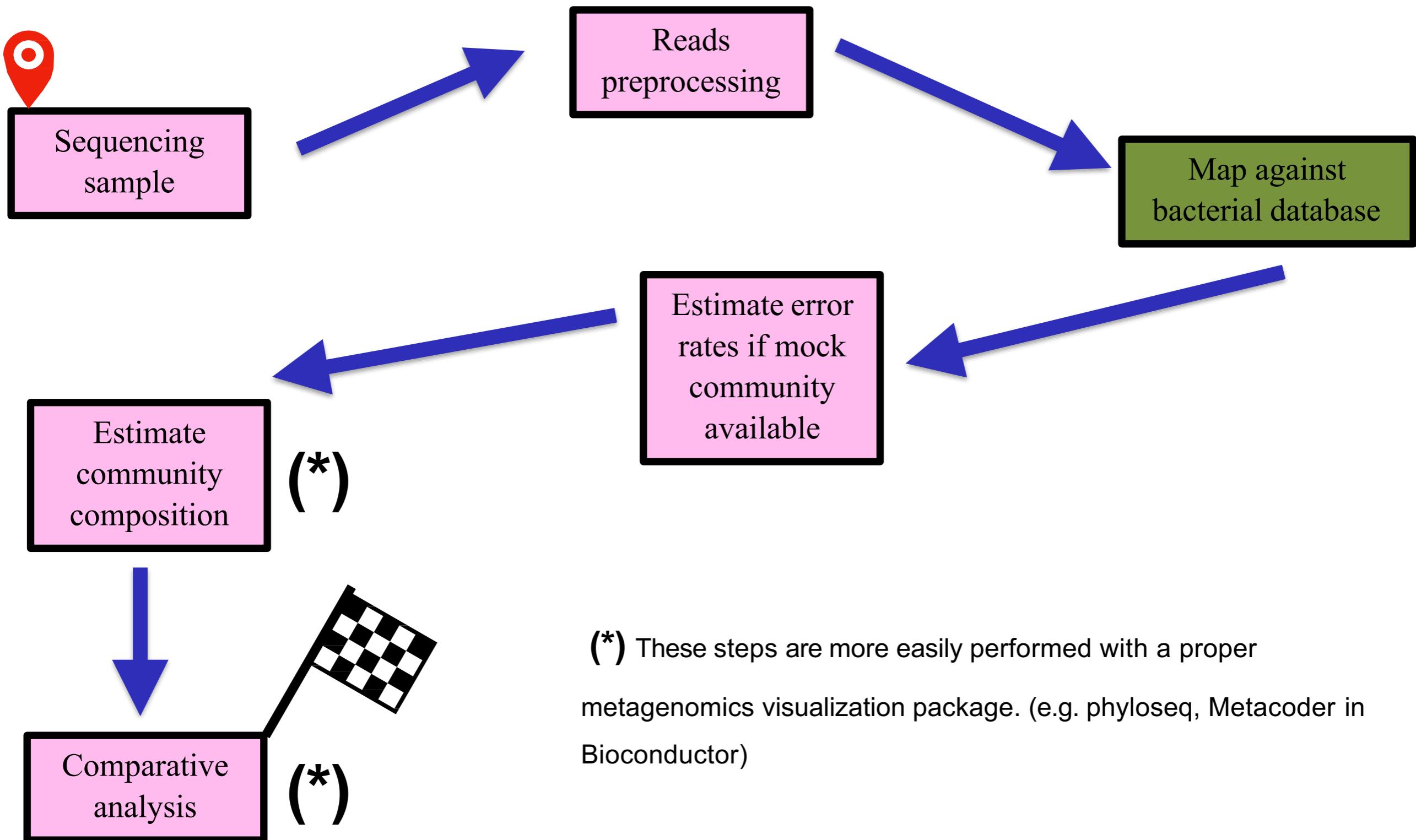
- Reference-free - no reference database and creates the clusters entirely from observed sequences
- Reference-based:
 - Closed-reference: uses a reference database of target gene sequences from known taxa and compares discovered sequences to them
 - Open-reference: open-reference clustering was developed, 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*

10
01
101010
101
10
01101
10
01
011

f. g. c. z.

0
1
0
0
1
0
0
1
1

16S example workflow and associated challenges



10
01
101101 1
010 0
0101 10010 01
101 10
010 01f. g. c. z.
01 1
10 0
01 1

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

10
01
010
0101

16S databases

[Genomics Inform.](#) 2018 Dec; 16(4): e24.

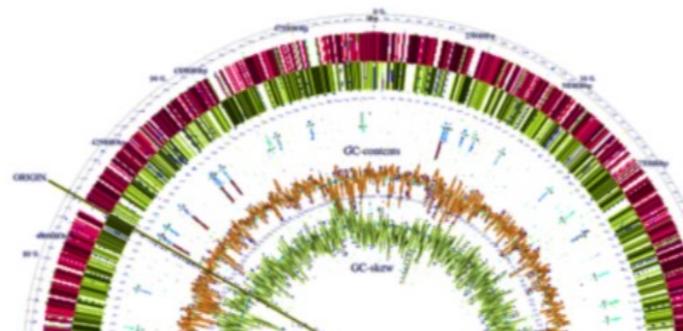
PMCID: PMC6440677

Published online 2018 Dec 28. doi: [10.5808/GI.2018.16.4.e24](https://doi.org/10.5808/GI.2018.16.4.e24)PMID: [30602085](#)

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

[Sang-Cheol Park¹](#) and [Sungho Won^{1,2,3,*}](#)

phylum *Latescibacteria* (0/556/0)
 phylum "Armatimonadetes" (0/1149/0)
 phylum "Verrucomicrobia" (0/10424/0)
 phylum "Acidobacteria" (0/15997/0)
 phylum Firmicutes (0/470524/0)
 phylum Cyanobacteria/Chloroplast (0/25864/0)
 phylum Marinimicrobia (0/997/0)
 phylum Aminicenantes (0/1546/0)
 phylum Omnitrophica (0/20/0)
 phylum Acetothermia (0/44/0)
 phylum Poribacteria (0/104/0)
 phylum Atribacteria (0/69/0)
 phylum Cloacimicrotes (0/179/0)
 phylum Candidatus Calescamantes (0/3/0)
 phylum candidate division WPS-1 (0/815/0)
 phylum candidate division WPS-2 (0/116/0)
 phylum Hydrogenedentes (0/460/0)
 phylum candidate division ZBS (0/76/0)
 phylum Ignavibacteriae (0/774/0)
 phylum Nitrospinae (0/537/0)
 ► Archaea Outgroup (0/1/0)
 ► unclassified_Bacteria (0/34557/0)
 domain Archaea (0/33971/0)
 phylum "Crenarchaeota" (0/1954/0)
 phylum "Euryarchaeota" (0/16984/0)
 phylum "Korarchaeota" (0/92/0)
 phylum "Nanoarchaeota" (0/139/0)



rrnDB: Stoddard et al.
NAR (2014)

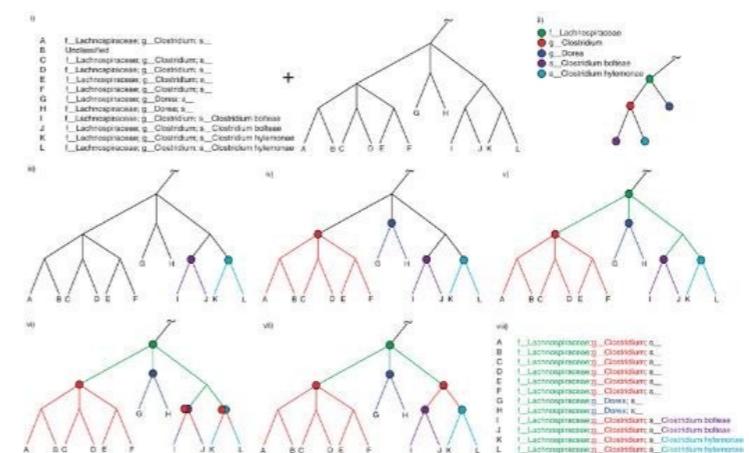
EZ BioCloud

EzBioCloud: Yoon
et al.
PubMed (2017)

RDP II: Cole et al.
NAR (2013)

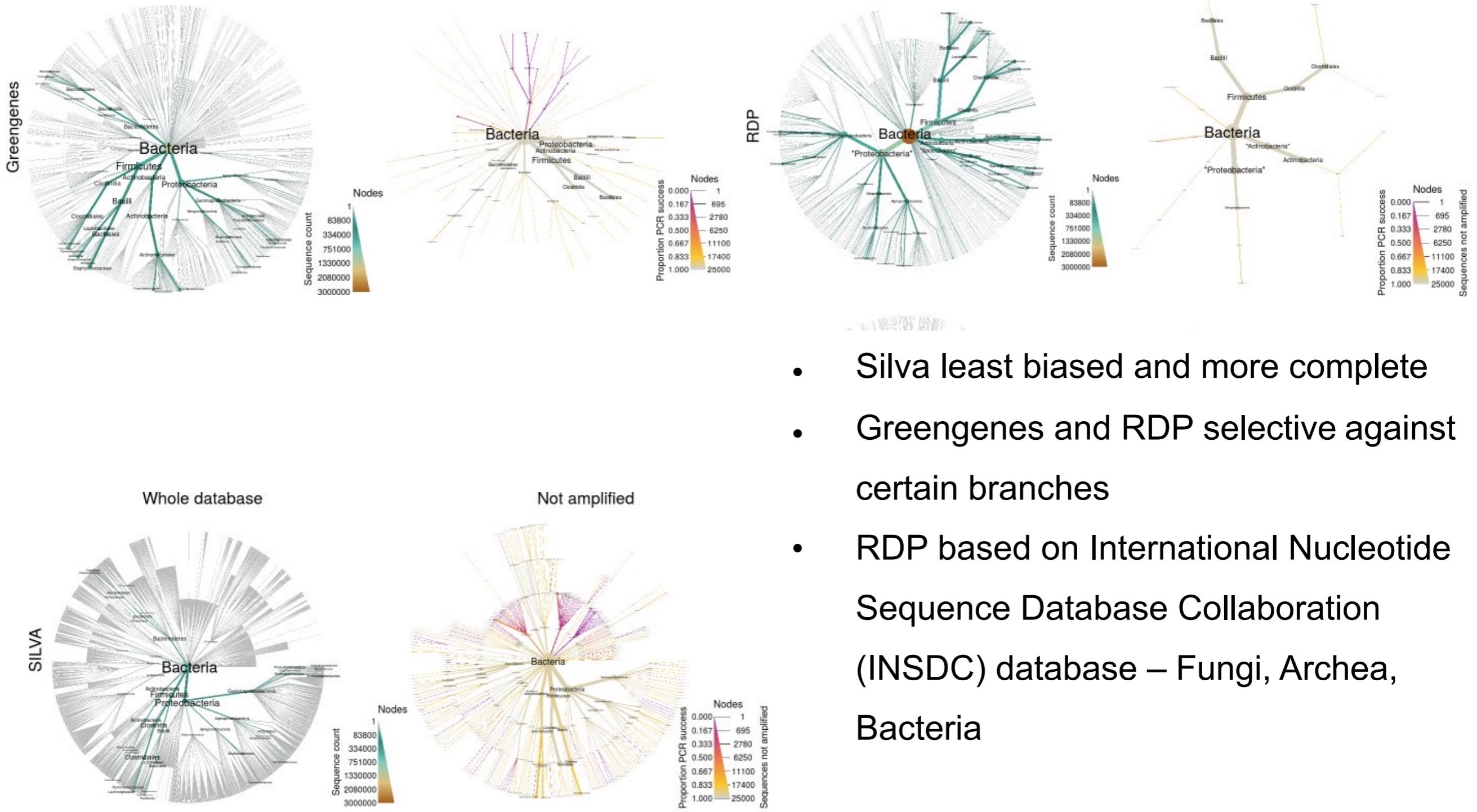


SILVA: Quast et al.
NAR (2013)



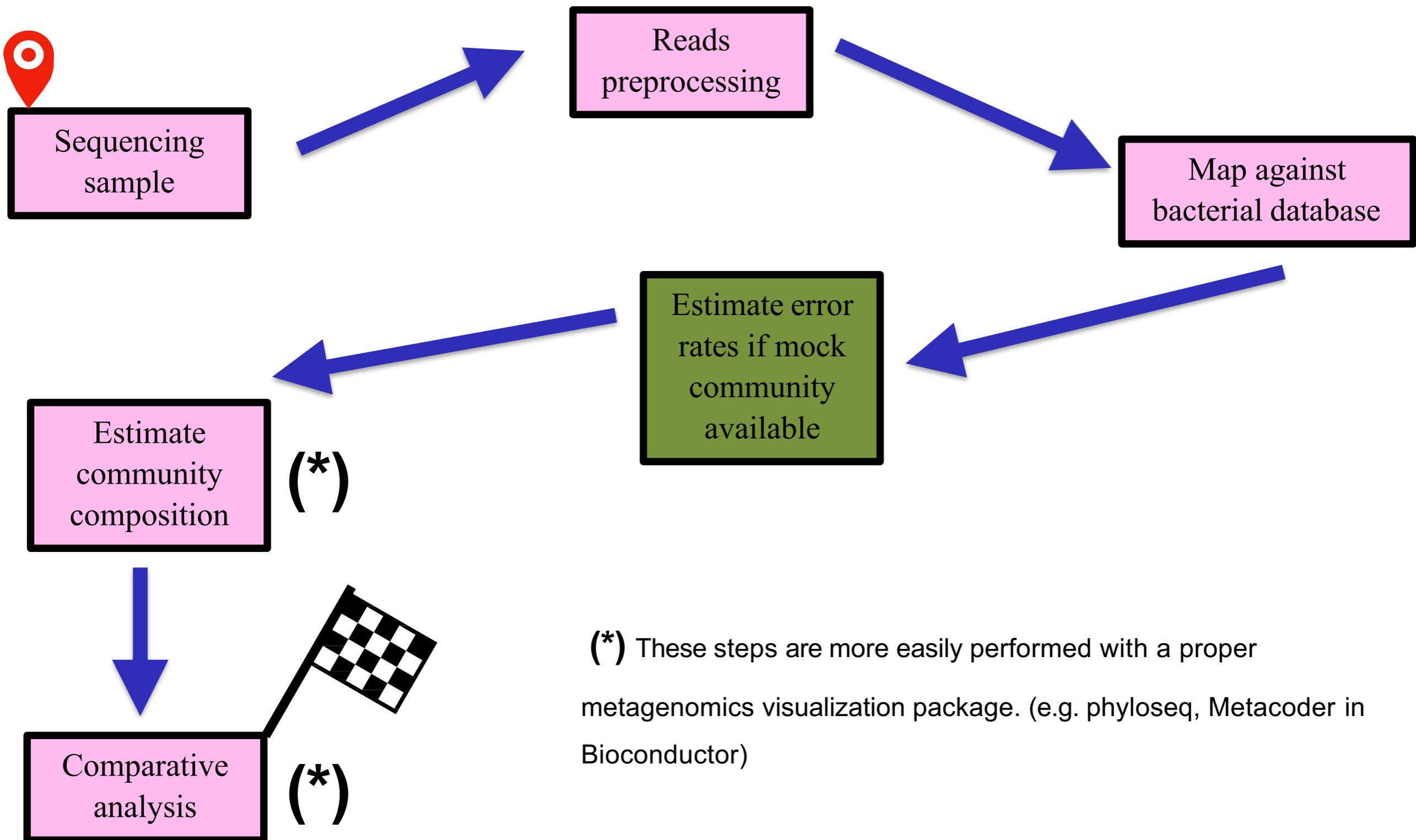
GreenGenes:
MacDonald et al. *ISME J* (2012)

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

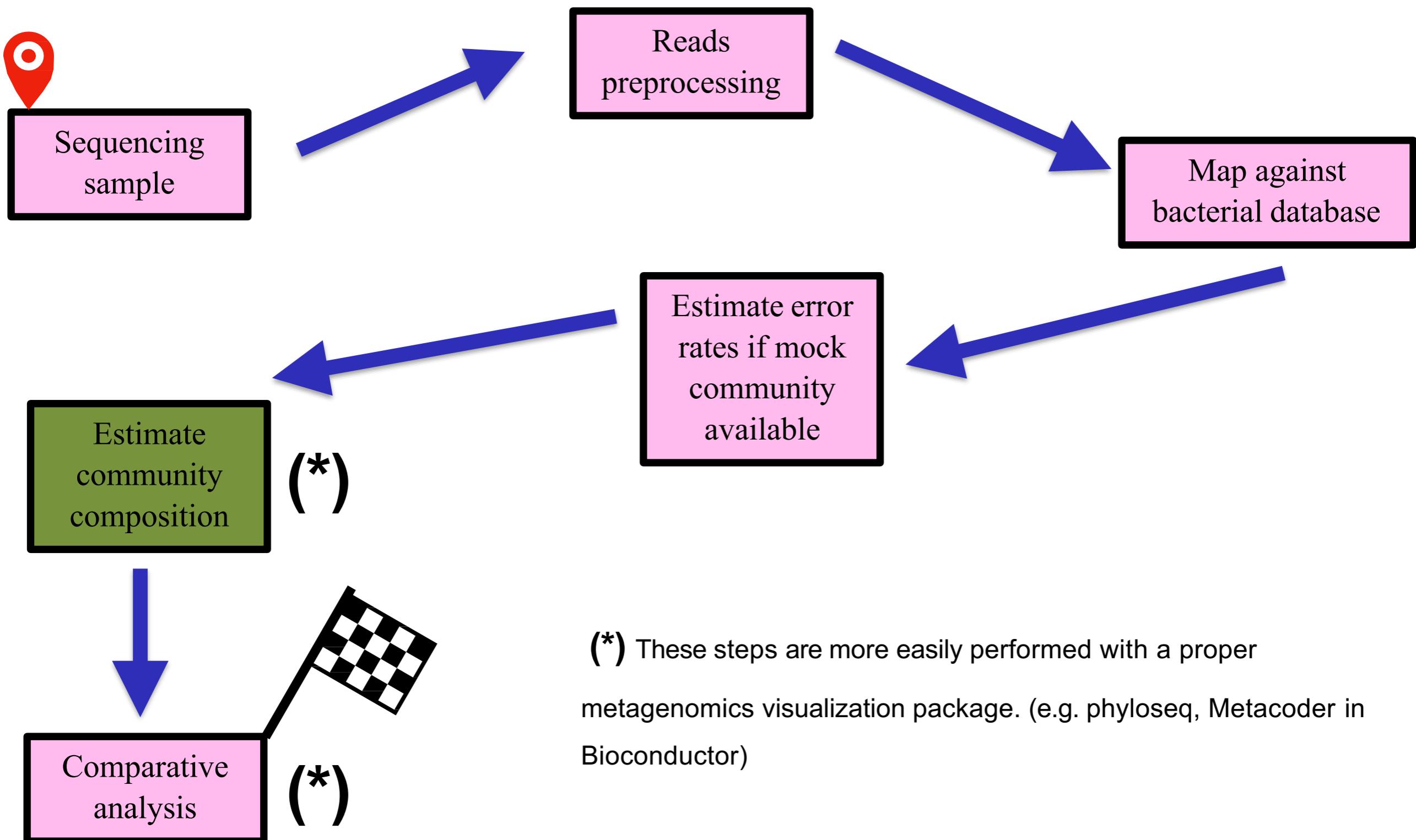


10
01
101101 1
010 0
0101 10010 01
101 10
010 01f. g. c. z.
01 1
10 0
01 1

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



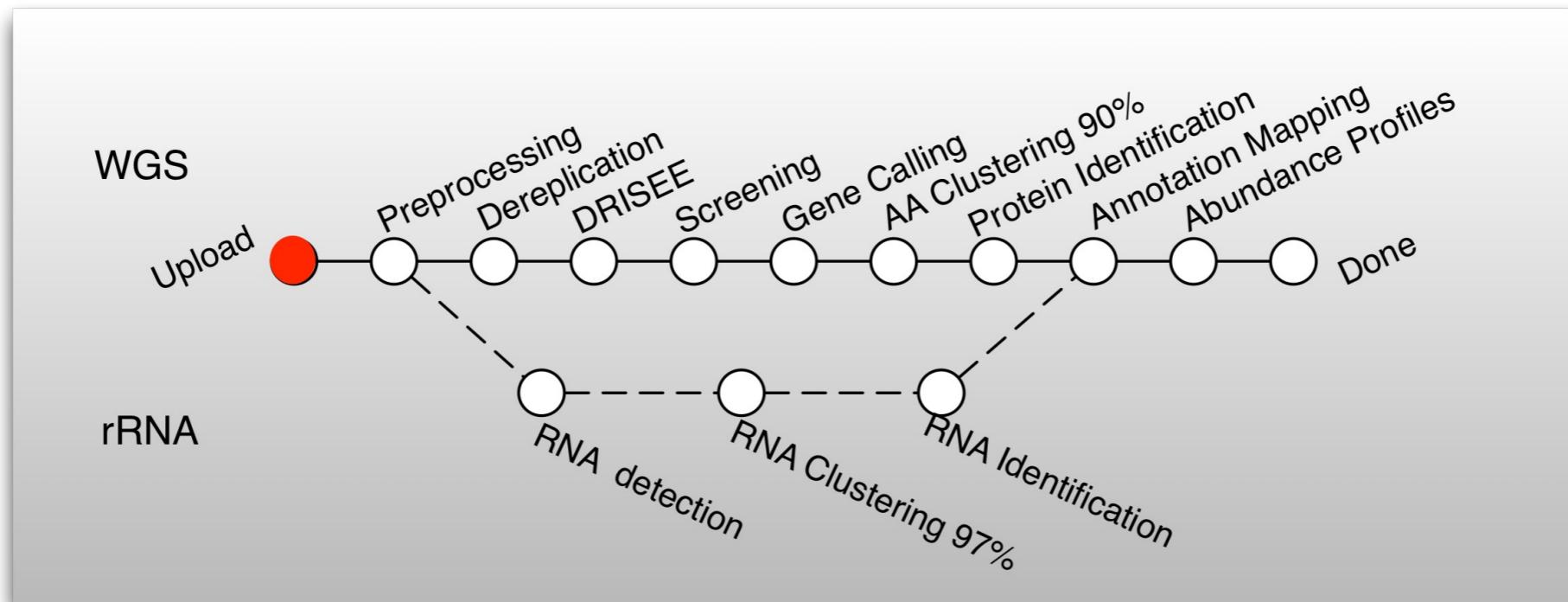
10
01
01
101functional genomics center zurich
010 01
101 10
010 01
010 01
f. g. c. z.
01 1
10 0
01 1
01 1

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?

10
01
01
101..
..
..
.....
...
...
...101 1
010 0
0101 10
...-
-
-functional genomics center zurich
010 01
101 10
010 01
f. g. c. z.
01 1
10 0
01 1
01 1

MG-RAST pipeline



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

rRNA identification

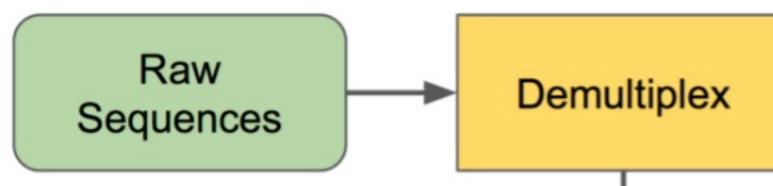
A BLAT similarity search for the longest cluster representative is performed against the M5rna database which integrates SILVA(Pruesse et al. 2007), Greengenes(DeSantis et al. 2006), and RDP(Cole et al. 2003).



Conceptual overview of QIIME 2

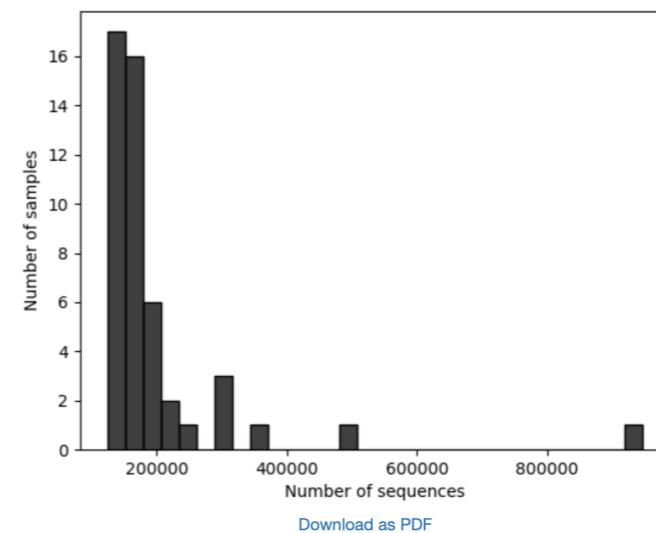
Now that we have read the glossary and key, let us examine a conceptual overview of the various possible workflows for examining amplicon sequence data:

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



1. Per sample read counts
2. Read frequency histograms

Forward Reads Frequency Histogram



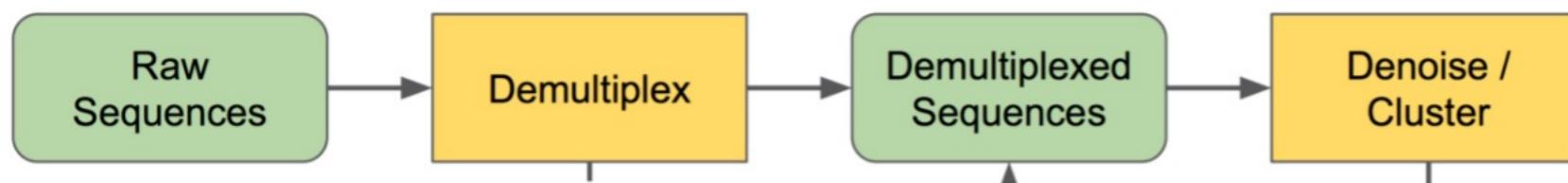
Requires installation but can be installed on a linux, windows or mac.



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

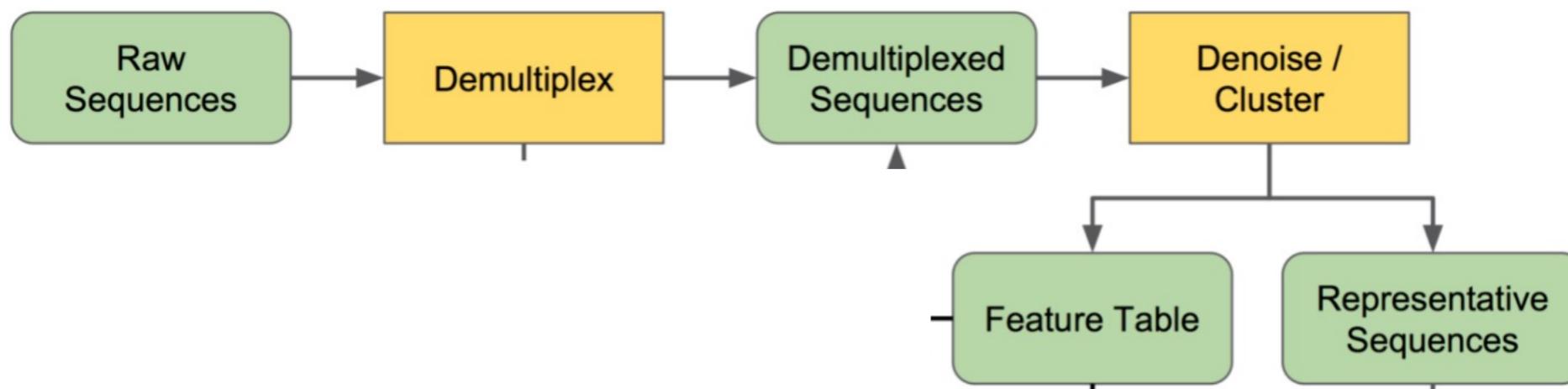
Requires installation but can be installed on a linux, windows or mac.



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

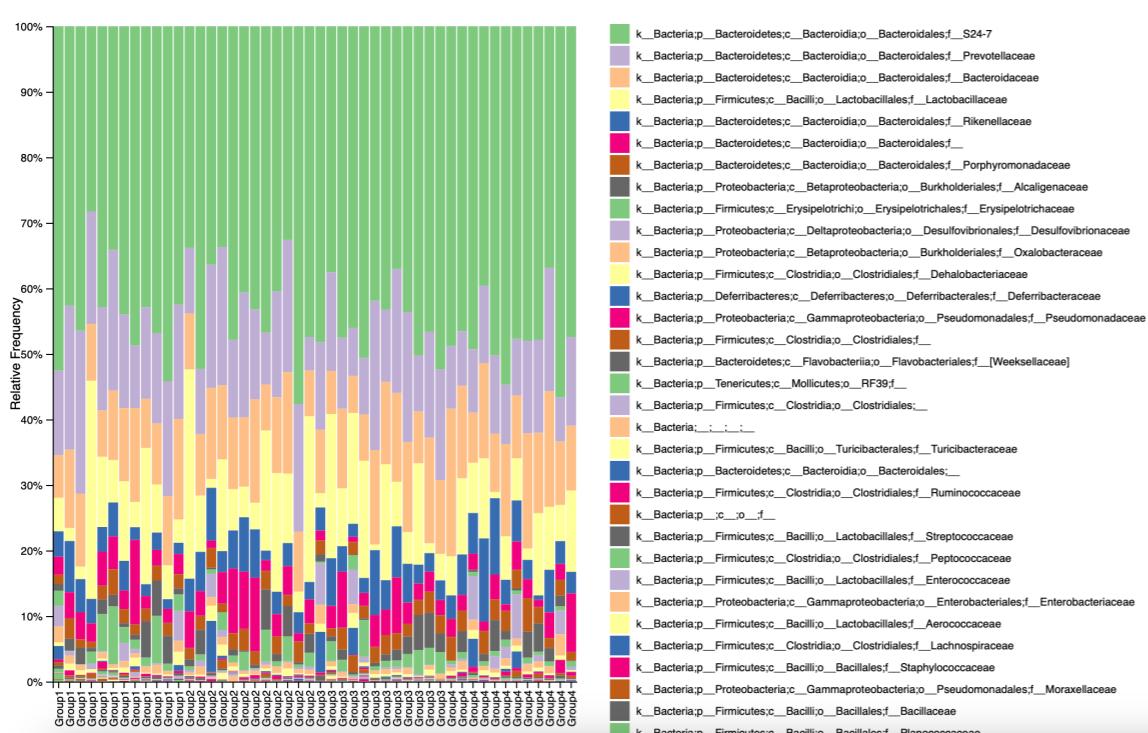
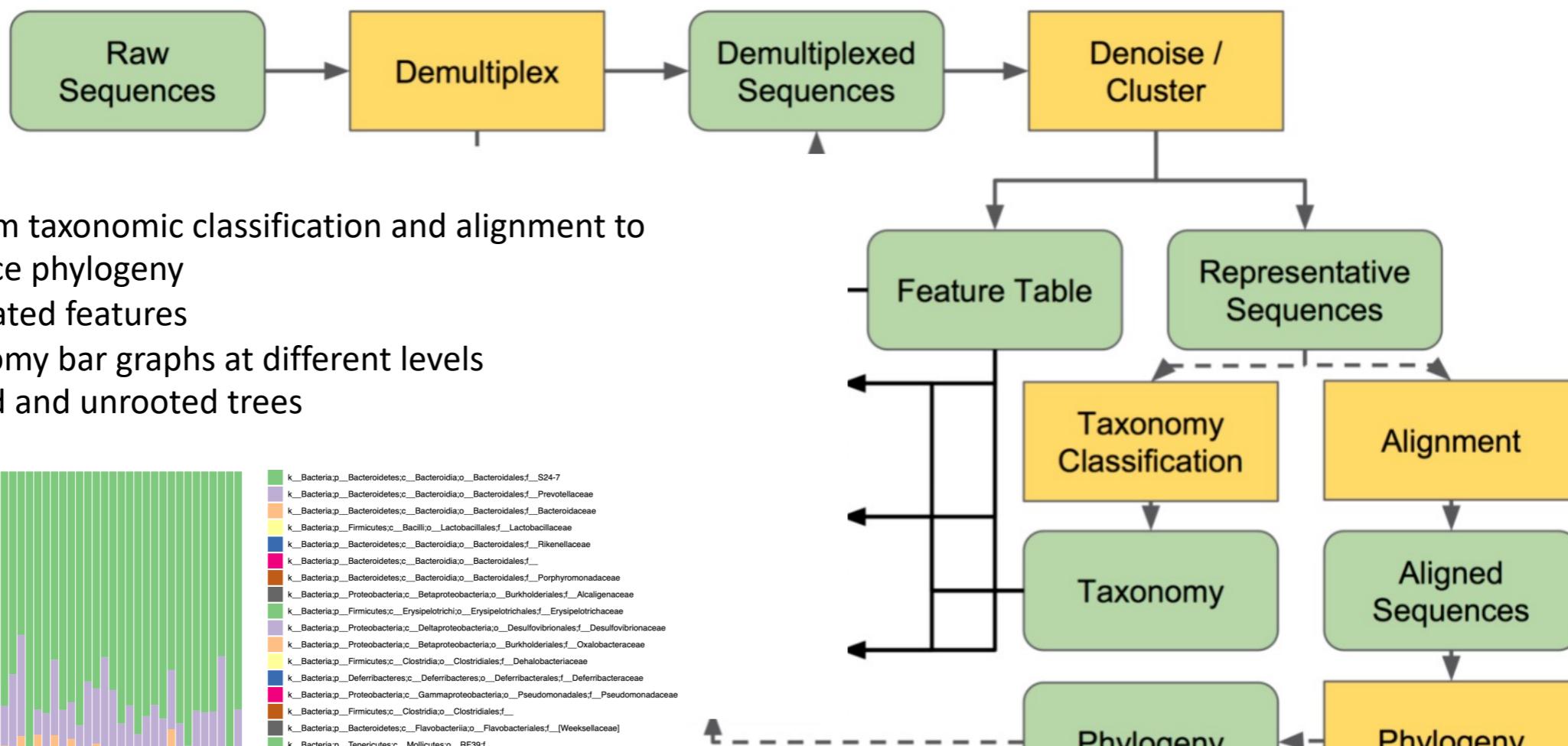
Requires installation but can be installed on a linux, windows or mac.

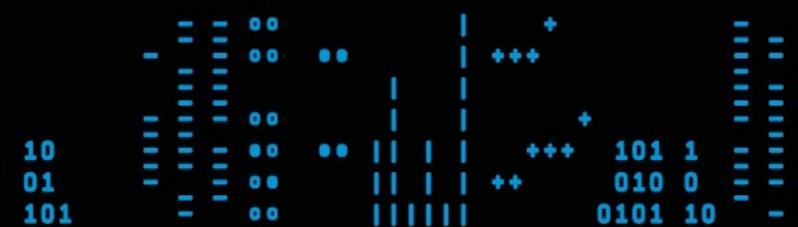
10
01
101...
...
...
101 1
010 0
0101 10...
...
...
010 01
101 10
010 01...
...
...
01 1
10 0
01 1functional genomics center zurich
f. g. c. z.0 1
1 0
0 1
1 0
+ 0 1

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/>



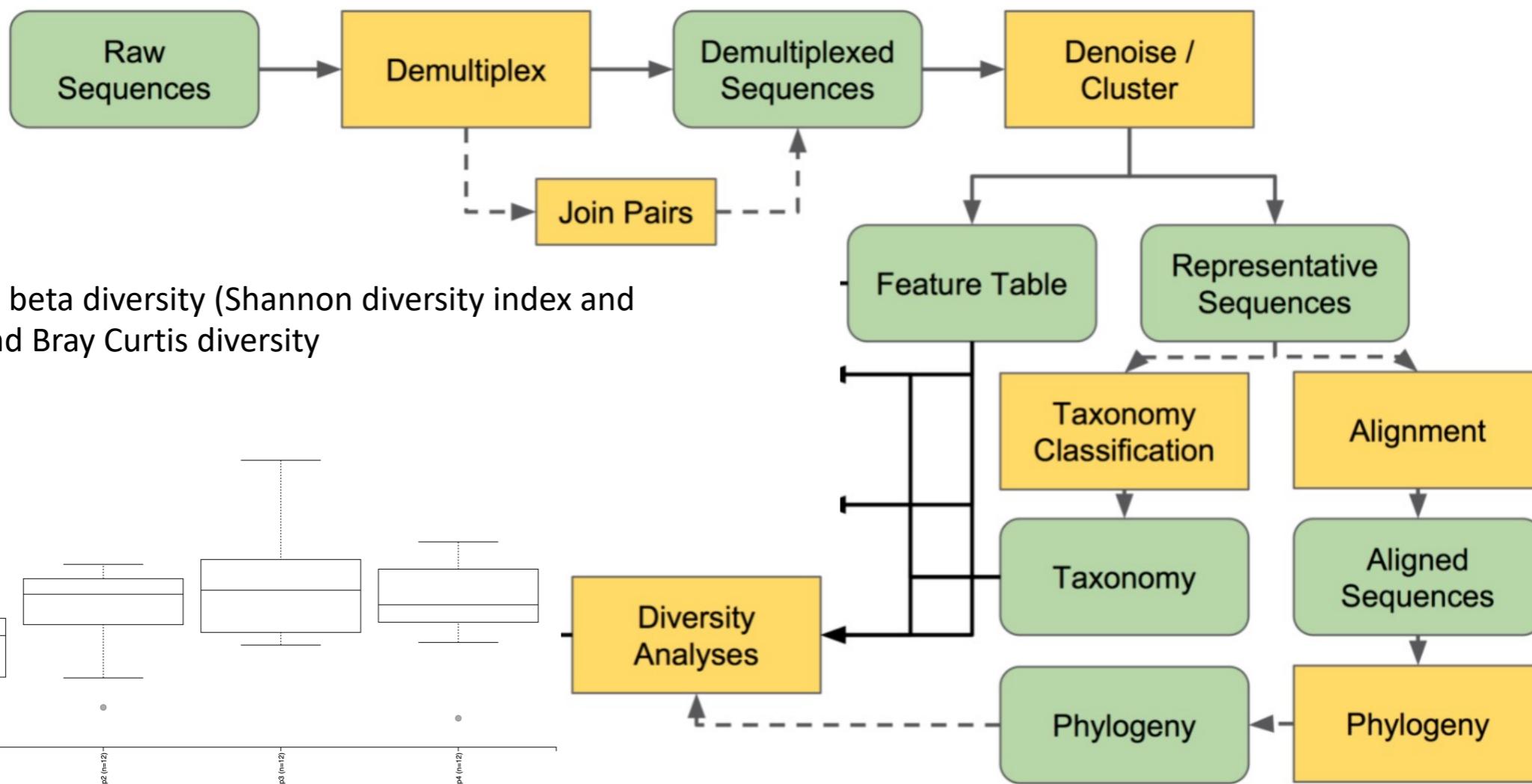


f. g. c. z.
01 1
10 0
01 1
10 0
01 01
01 1
10 10
01 01
01 1

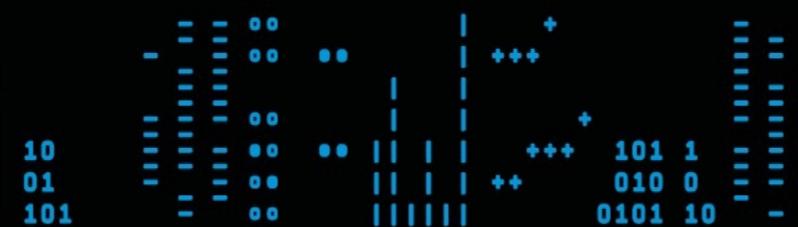
Conceptual overview of QIIME 2

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/>



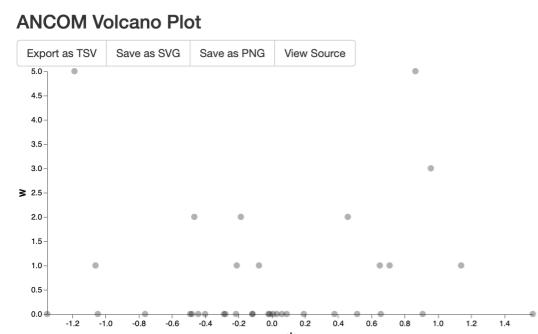
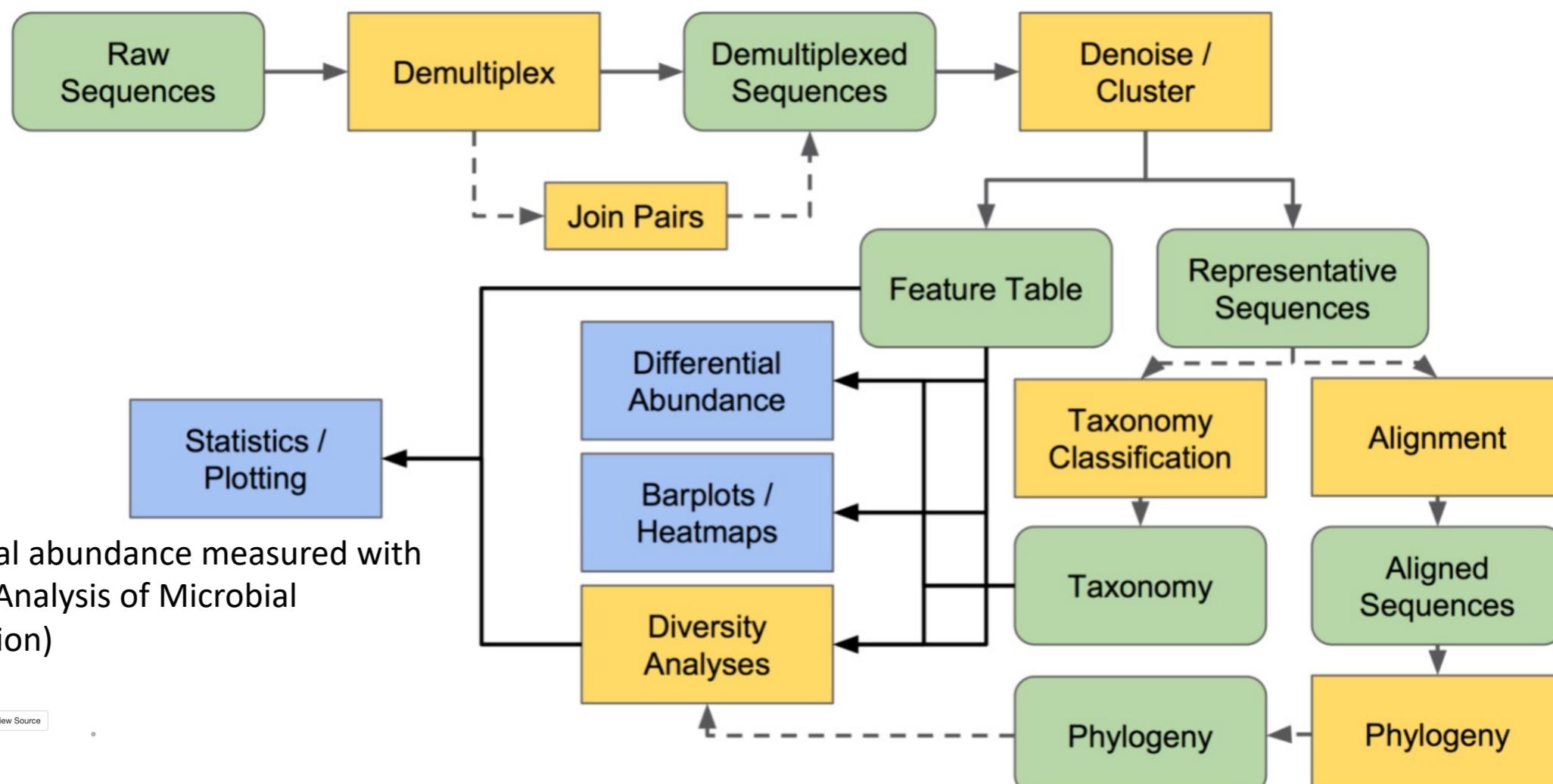
1. Alpha and beta diversity (Shannon diversity index and Jaccard and Bray Curtis diversity)

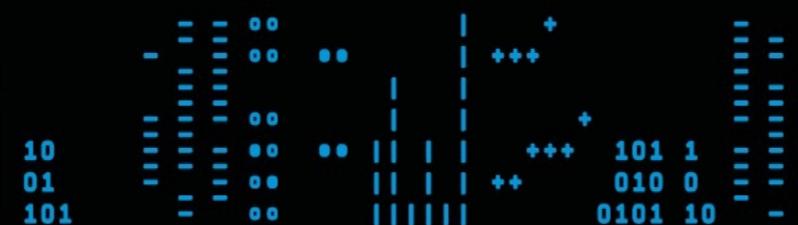


Conceptual overview of QIIME 2

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/>

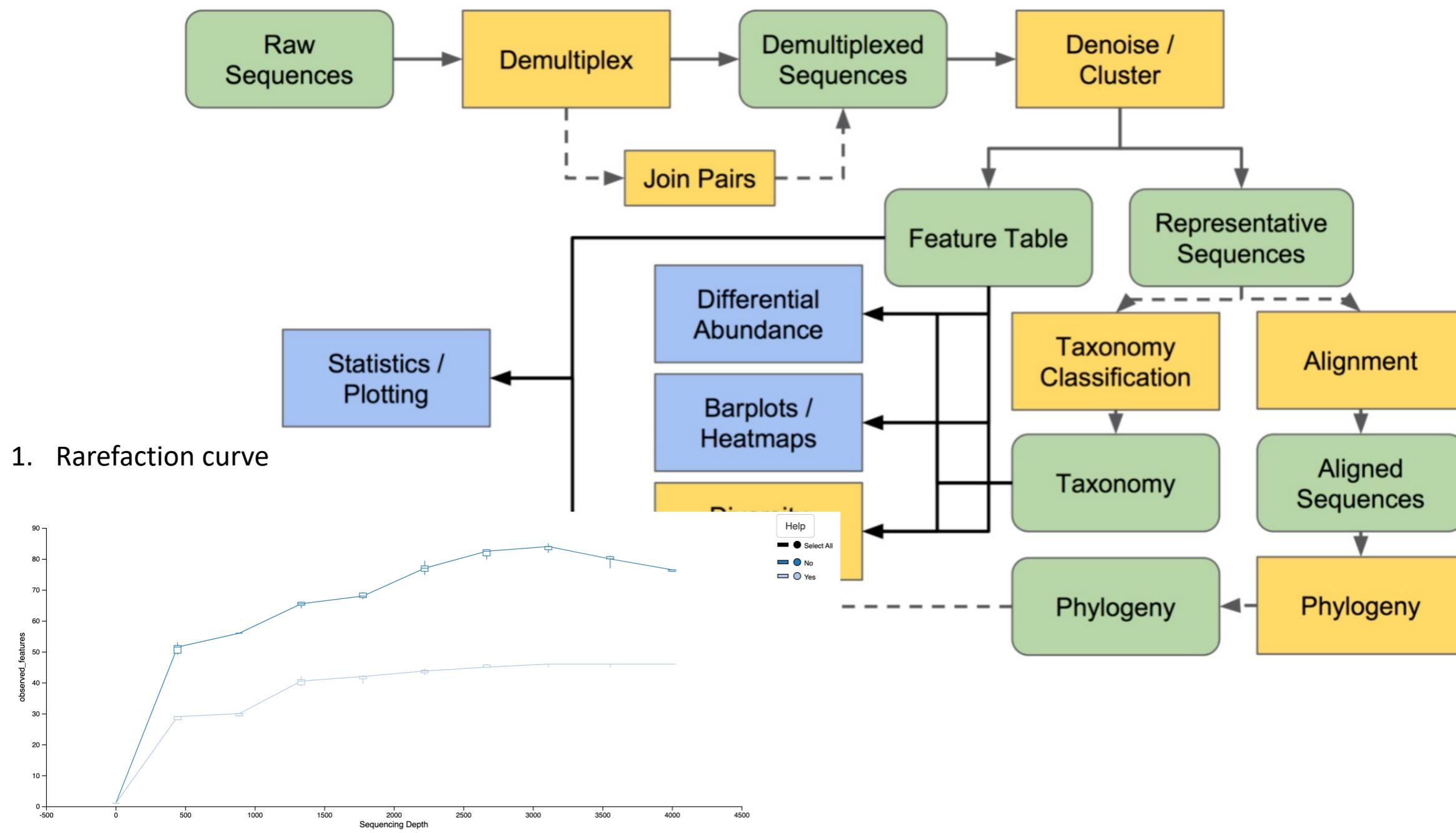




Conceptual overview of QIIME 2

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/>





A quick look at some other tools

USEARCH/UPARSE

Just a binary, super easy

Main clustering algorithm
also an implementation of
vsearch

High memory, multi-core
server

Scalable (simple binary)

[https://www.drive5.com/
usearch/manual/
uparse_pipeline.html](https://www.drive5.com/usearch/manual/uparse_pipeline.html)

Mothur

Single program with minimal
dependencies (read easy to
install and setup)

Reimplementation of tried-
and-tested algorithms (e.g.,
mapping, clustering)

High memory, multi-core
server

Not so much scalable (slows
down with sample size)

DADA2 within QIIME2

Divisive Amplicon Denoising Algorithm

Full R-based; easy to
proceed with downstream
analysis

Fast

Low memory, potentially
parallelizable

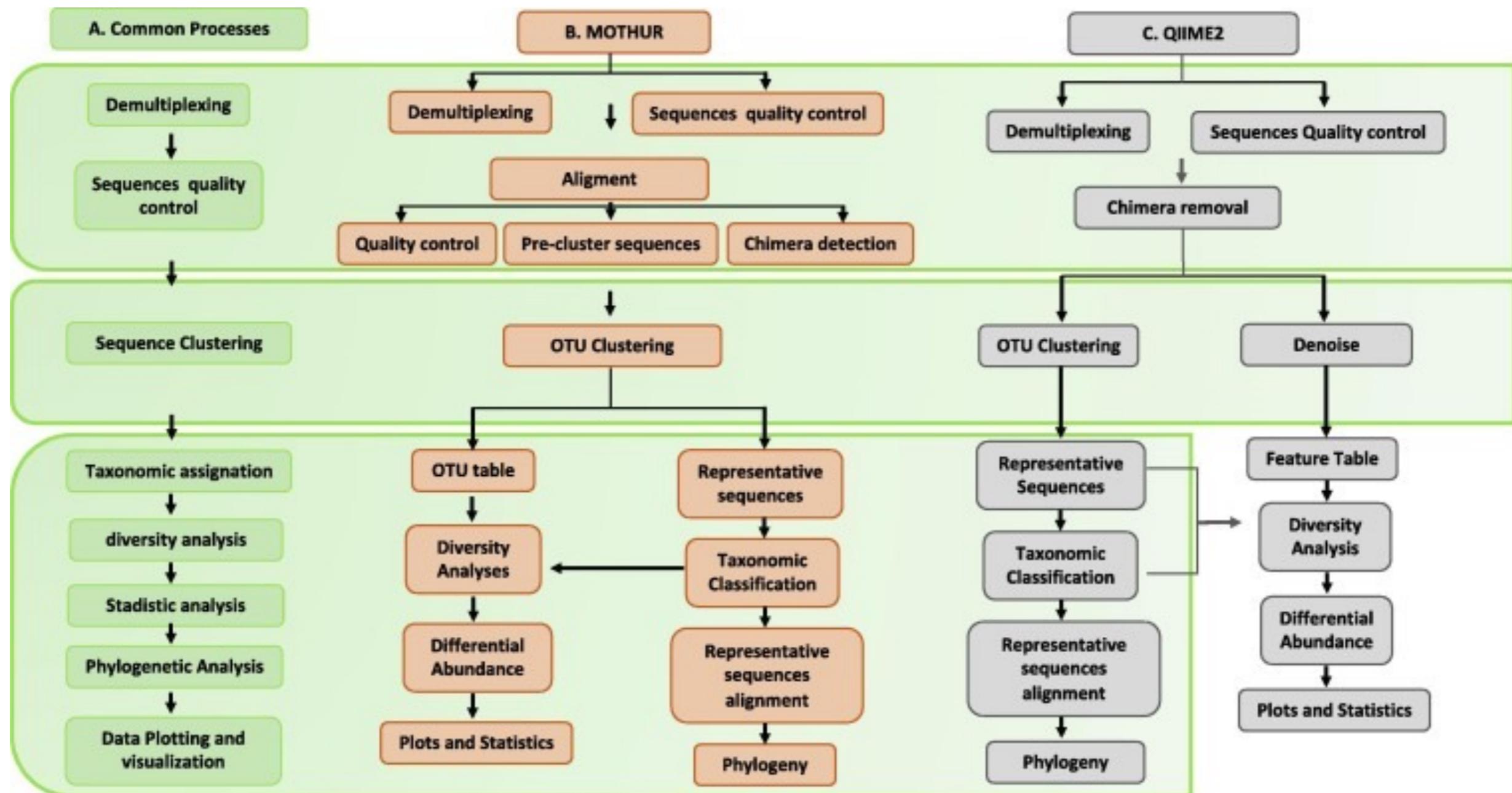
Heavily affected by errors;
requires more reads

[https://www.mothur.org/wiki/
MiSeq_SOP](https://www.mothur.org/wiki/MiSeq_SOP)

[https://benjineb.github.io/
dada2/tutorial.html](https://benjineb.github.io/dada2/tutorial.html)

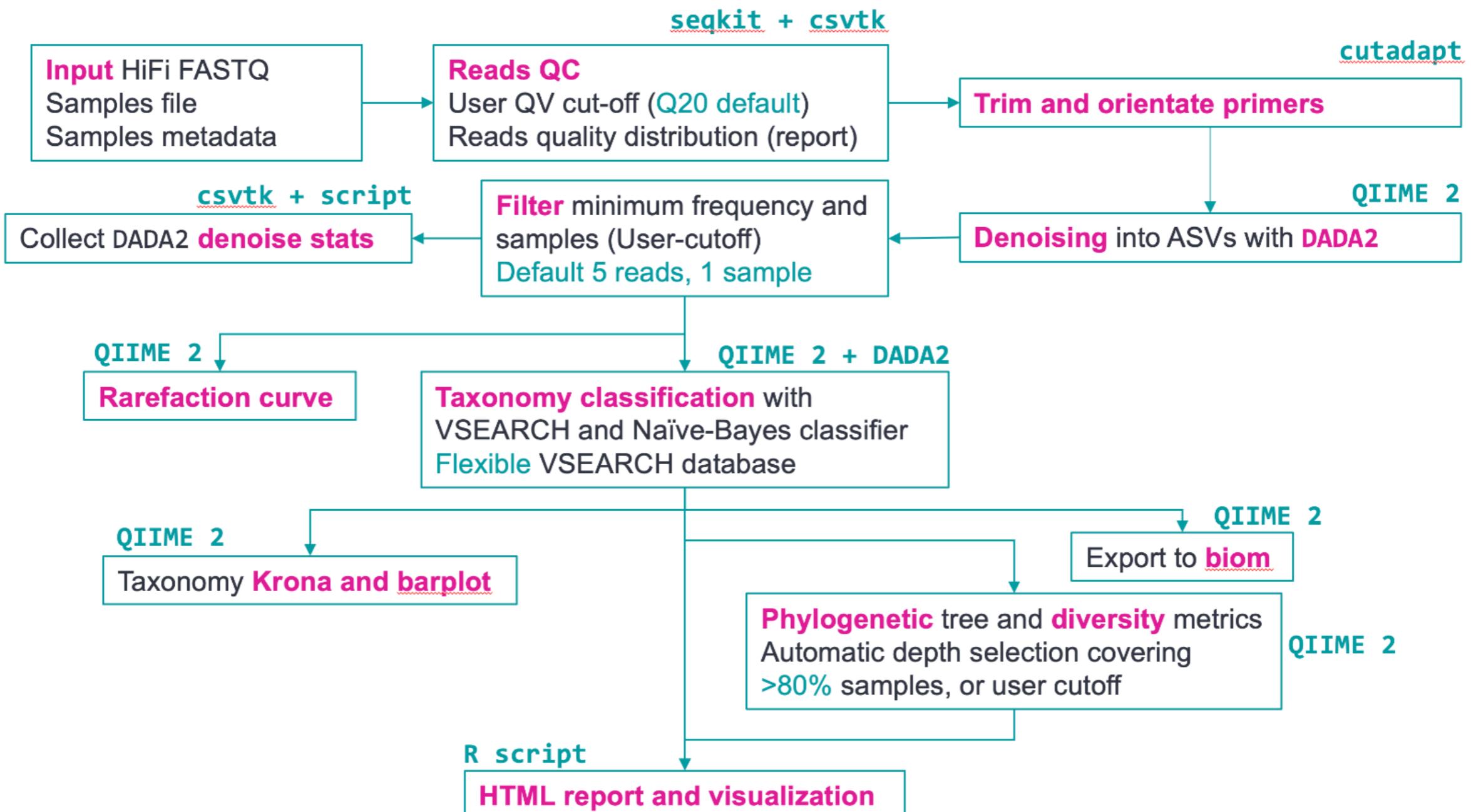
10
01
101101 1
010 0
0101 10...
0101 10...
010 01...
101 10...
010 01...
01 1...
10 0...
01 1...
01 1...
01 1...
01 1...
01 1...
01 1...
01 1...
01 1...
01 1...
01 1...
01 1...
01 1...
01 1...
01 1...
01 1...
01 1...
01 10
1
0

A quick look at some other tools



10
01
101...
101 1
010 0
0101 10
...
010 01
101 10
010 01
...
01 1
10 0
01 1

16S long read sequencing



10
01
101

Example of statistical analysis – diversity estimates

Alpha Diversity

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

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

where P_i = 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

10
01
101010
101
10
010
101
10
011

f. g. c. z.

01
10
011

Example of statistical analysis – diversity estimates

Alpha Diversity

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

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

where P_i = 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

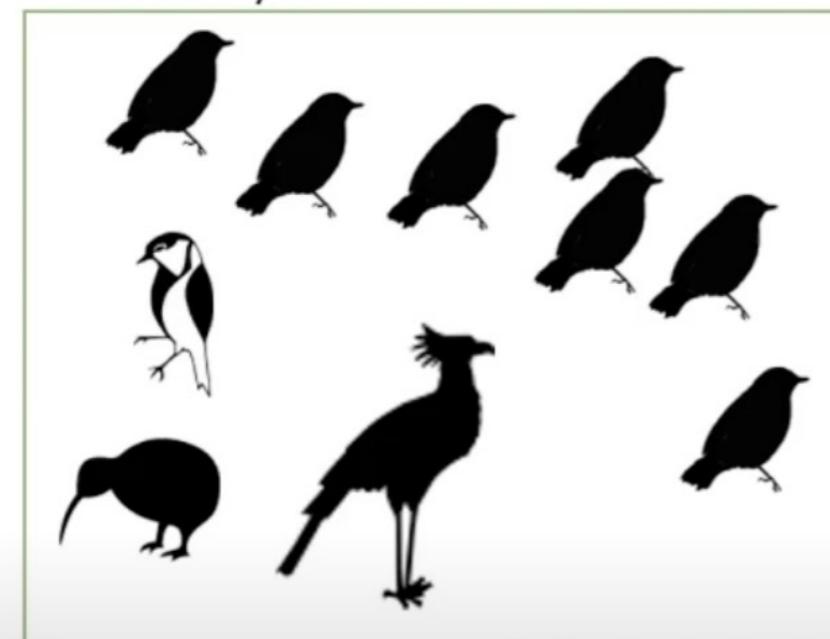
- based on abundance/read count data
 - differences in microbial abundances between two samples (e.g., at species level)
- values are from 0 to 1

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

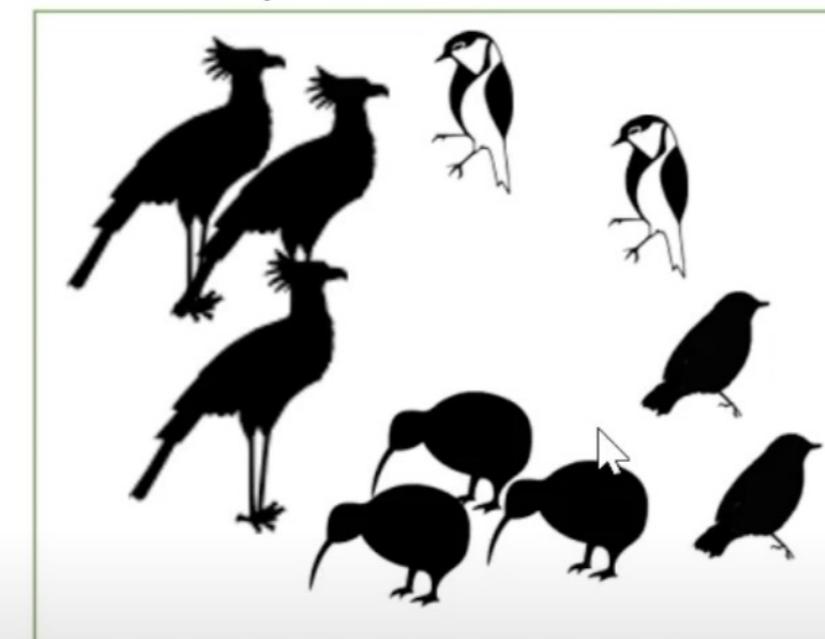
Shannon vs Simpson's diversity Index

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		

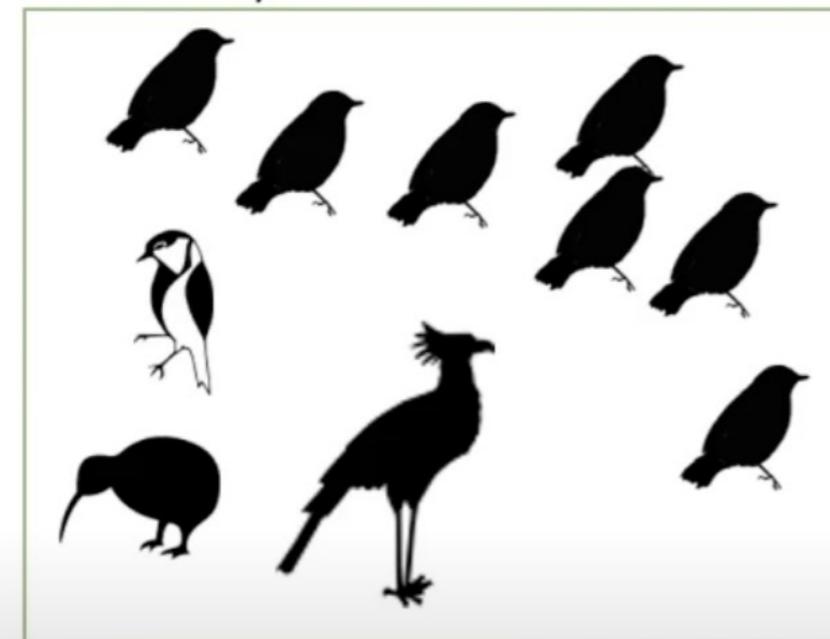
$$D = \frac{N(N-1)}{\sum n(n-1)}$$

N = Total number of organisms

n = Population of each individual species

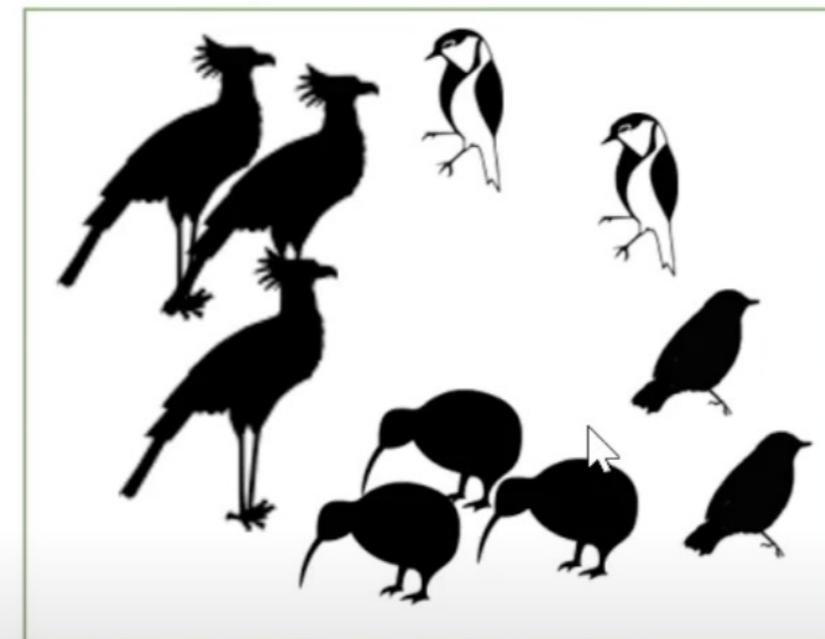
Shannon vs Simpson's diversity Index

Community A



Abundance = 10
Species Richness = 4
Diversity = ?

Community B



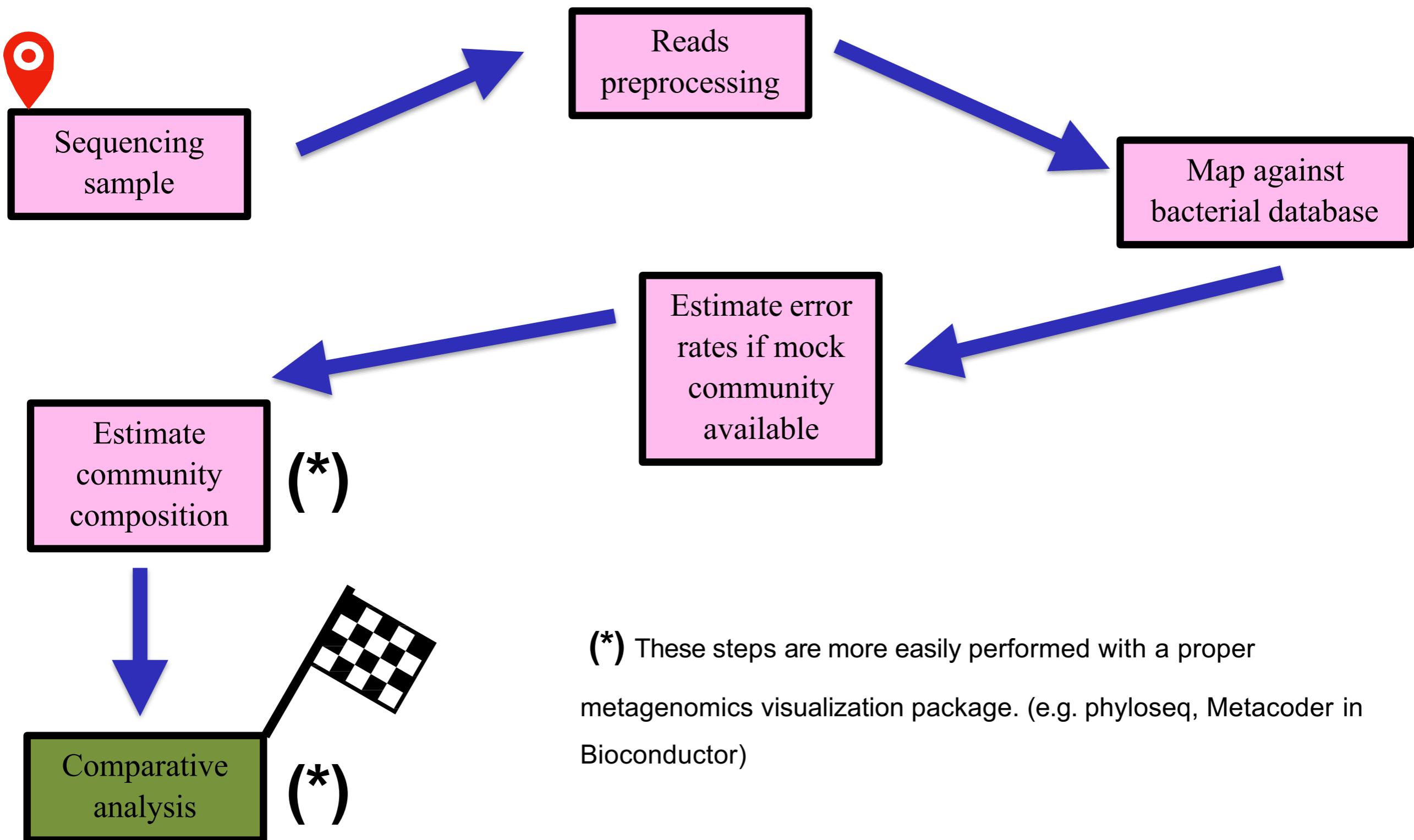
Abundance = 10
Species Richness = 4
Diversity =

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

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

where P_i = the proportion of individuals in each species

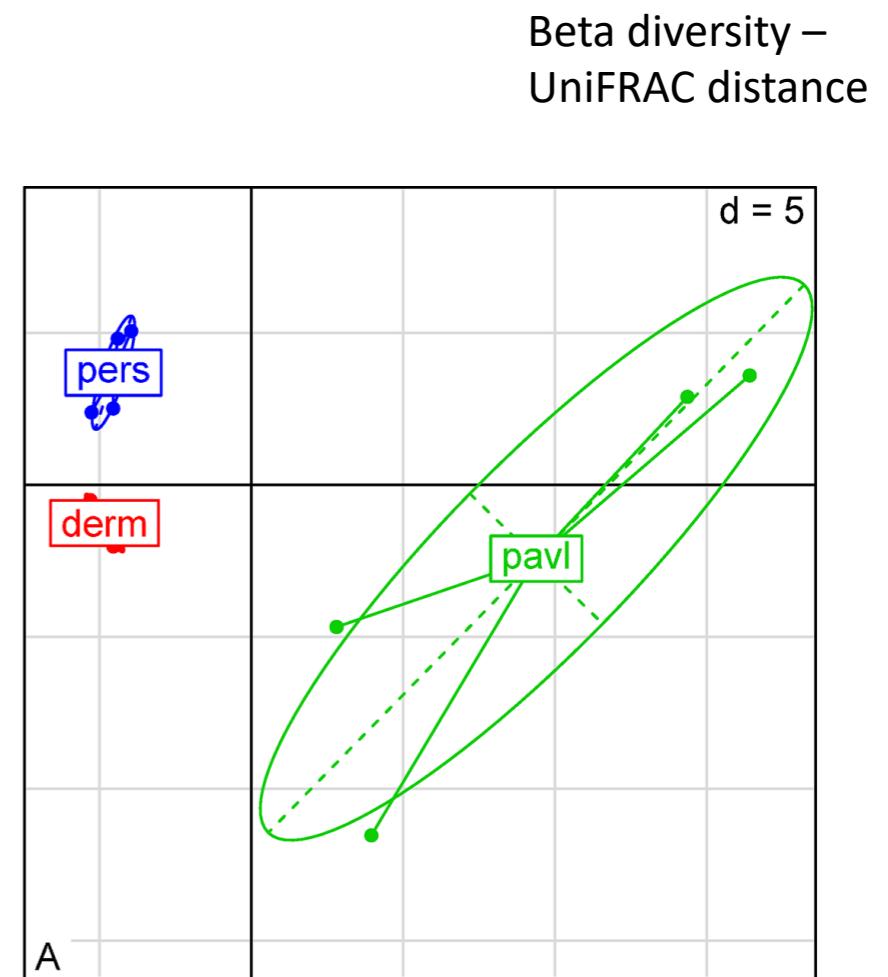
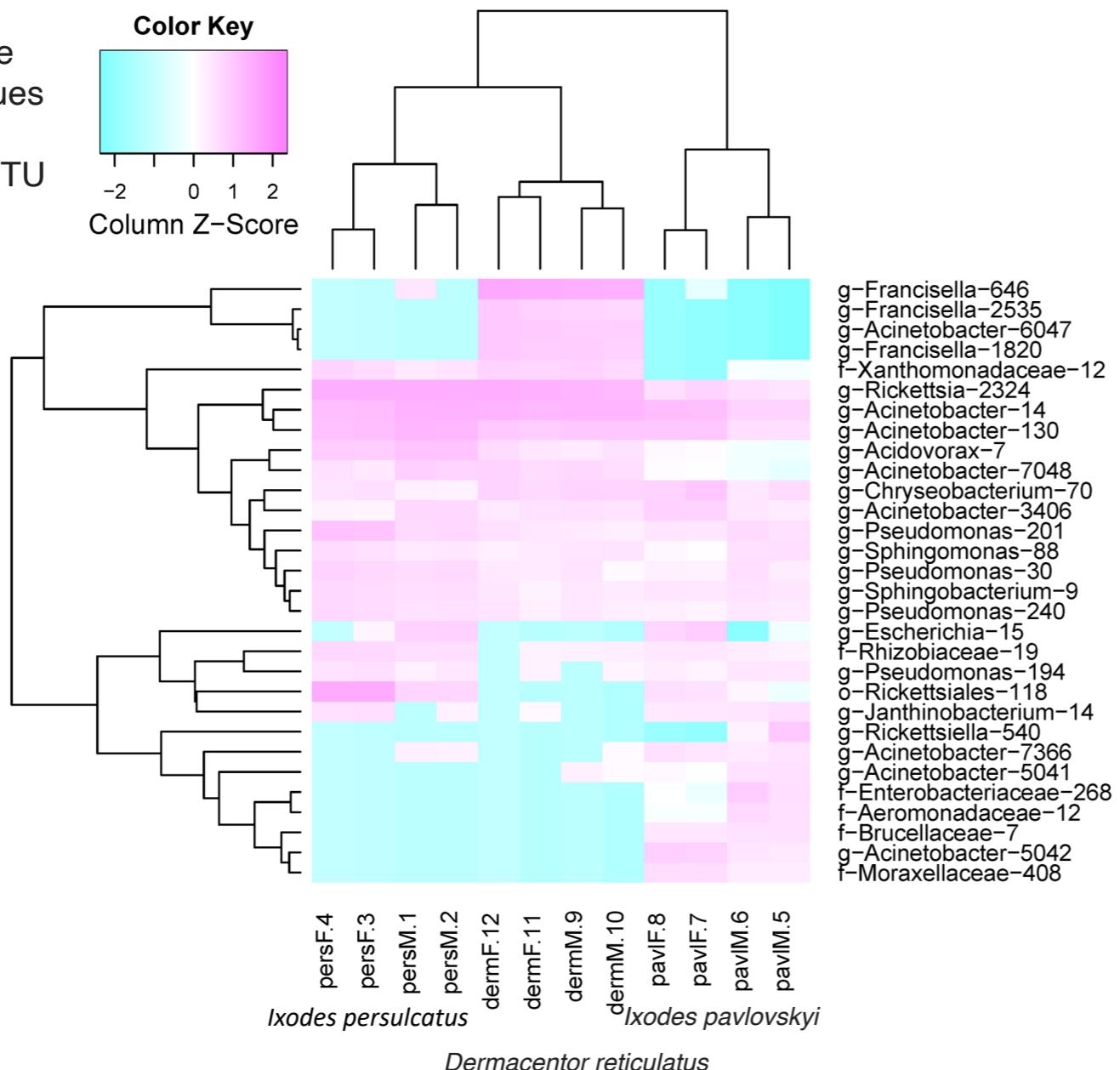
16S example workflow and associated challenges



10
01
10110
01
010
0101
10
010
01
011

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)



10
01
01
101..
..
..
..+++
++
010 1
010 0
0101 10
-
...
| + + + +..
..
..
..010 01
101 10
010 01
-
..functional genomics center zurich
f. g. c. z.
01 1
10 0
+ 01 1

Technical details

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

10
01
01
101101 1
010 0
0101 10010 01
101 10
010 01

f. g. c. z.

01 1
10 0
01 1

Processed 16S data

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

10
01
01
101functional genomics center zurich
010 01
101 10
010 01
f. g. c. z.
01 1
10 0
01 1

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



Taxonomic 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;

10
01
01
101..
..
..
101 1
010 0
0101 10
|
010 01
101 10
010 01
+ 01 101 1
10 0
01 1
+ 01 1

Sample Information

- Typical design matrix

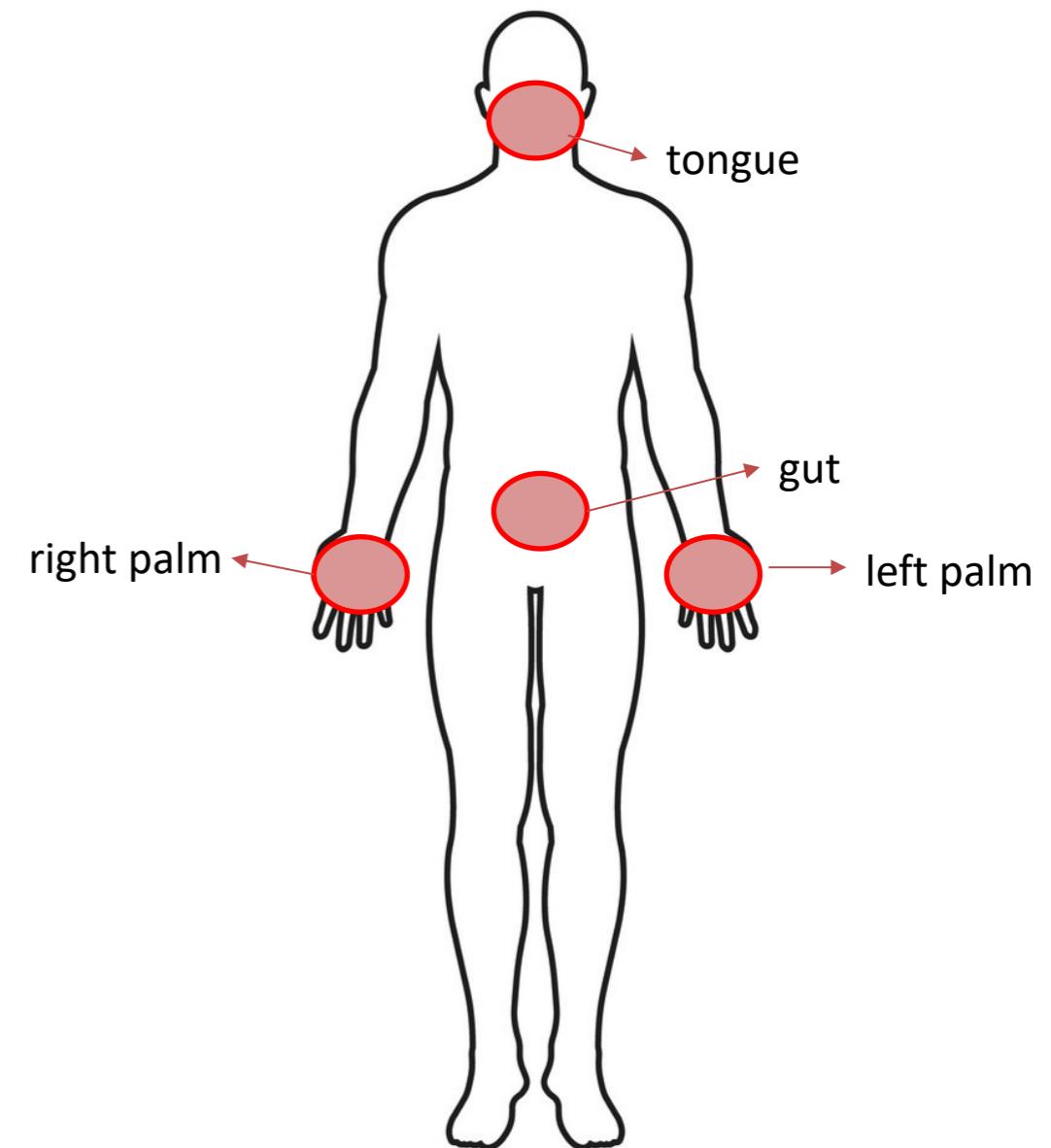
Name	Technology	[Factor]	Group	[Factor]
post1	Illumina		post	
post2	Illumina		post	
post3	Illumina		post	
pre1	Illumina		pre	
pre2	Illumina		pre	
pre3	Illumina		pre	

10
01
101..
..
.....
...
...
101 1
010 0
0101 10..
..
..
....
..
..
010 01
101 10
010 01..
..
..
010 01
010 01..
..
..
f g c z0 1
0 1
0 1
10 0
0 1 1

Practical

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

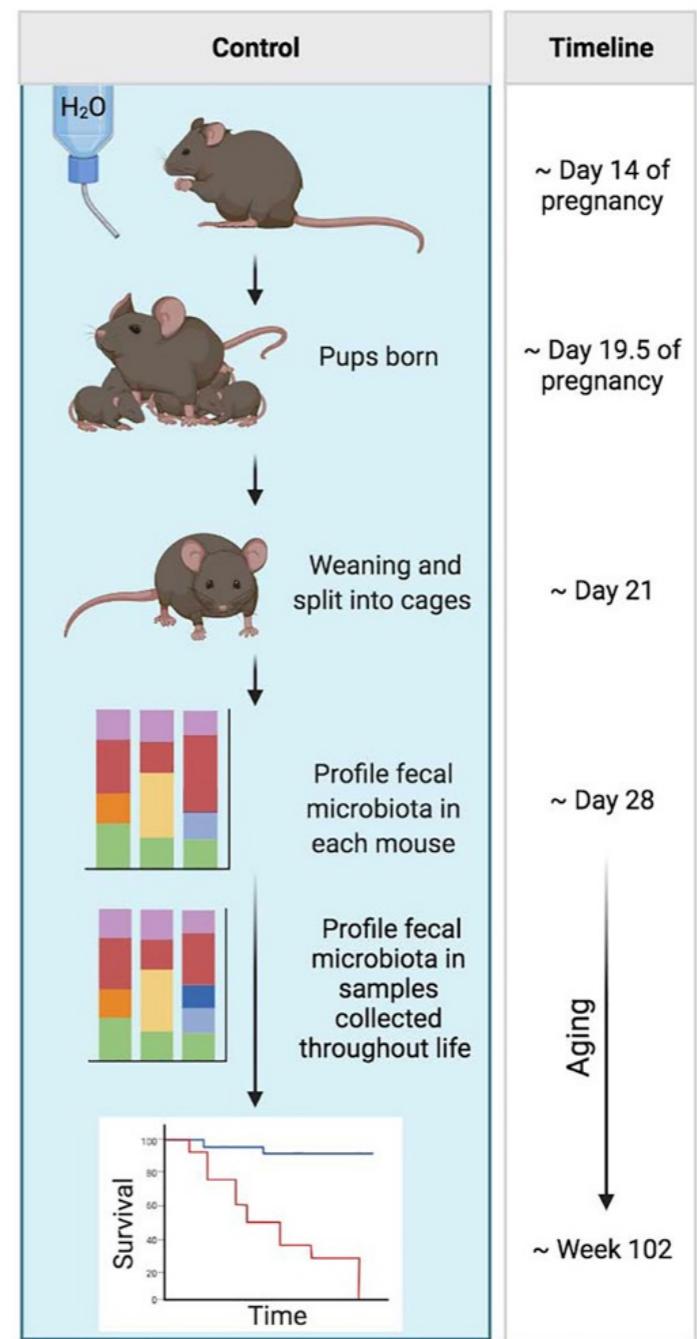
- Earth Microbiome Project
- sequenced hypervariable region 4 (**V4**) 16S rRNA
- analysis of human microbiome samples
- Original study is from Copraso et al. 2011
- Analysed with Quantitative Insights Into Microbial Ecology (QIIME2)





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>



10

01

101

101

1

0

010

0

10

0101

10

010

01

-

101

10

1

010

01

-

f

g

c

z

01

1

0

10

0

10

01

1

01

1

0

1

01

1

Lets do some work!

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

Scroll to **QIIME2 analysis on SUSHI**