Introductory Microbiome Bioinformatics

Lu Yang, Ph.D. and Jun Chen, Ph.D.

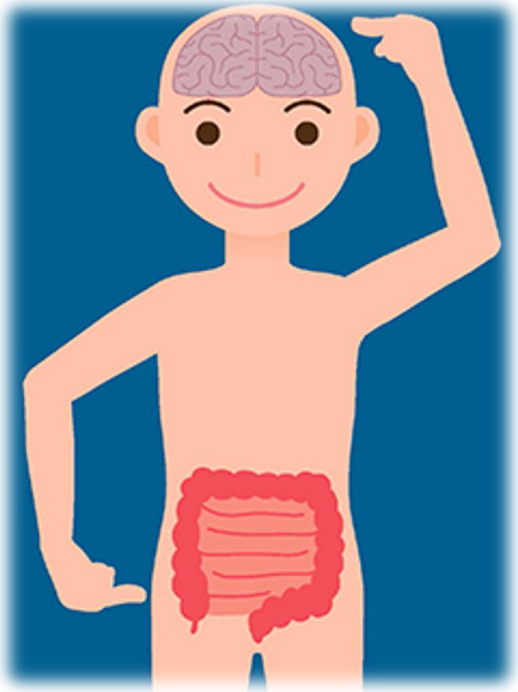
Department of Quantitative Health Sciences
Mayo Clinic

June 26, 2024

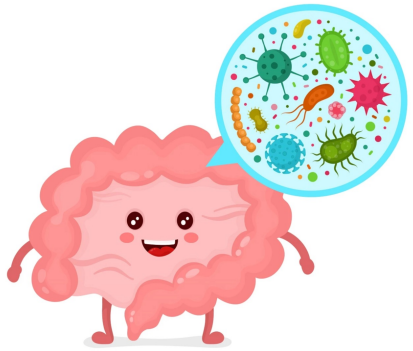
Outline

- ❑ Why We Study the Microbiome
- ❑ Sequencing Approaches to Study the Microbiome
  - ✓ 16S Amplicon Sequencing
  - ✓ Shotgun Metagenomic Sequencing
- ❑ Other Approaches

The Human Microbiome



- A healthy gut is like a forest, made up of various interacting species that compete for different ecological niches.
- The human body contains twice as many microbial cells as human cells.
- The genes in the human gut microbiota are estimated to be 100 times more than the genes in the human genome.
- The gut microbiota is made up of approximately 1000 different species.

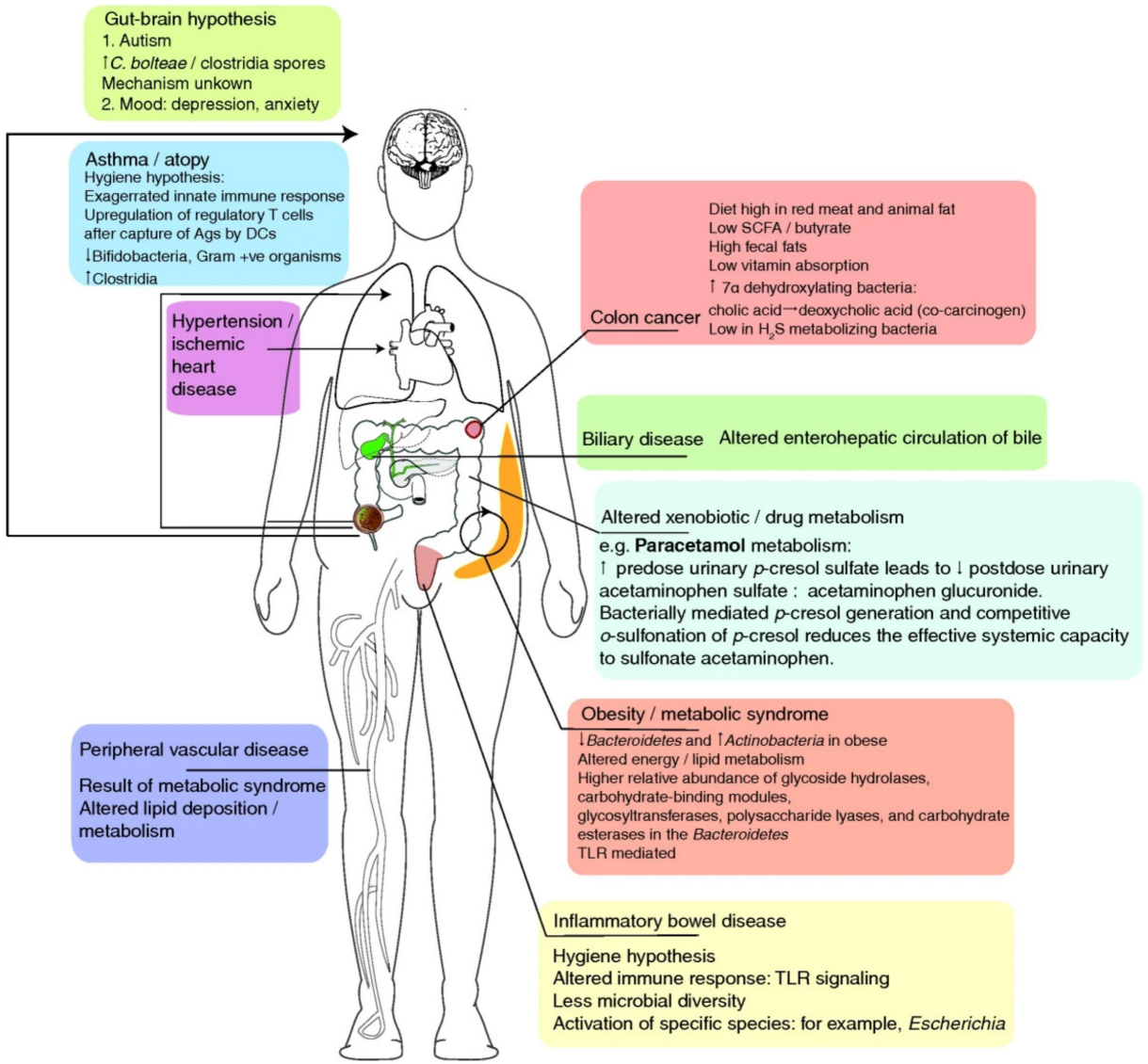


Bäckhed, F., et.al. (2005); Shreiner, A. B., et.al. (2015)

Microbiome, Health & Disease

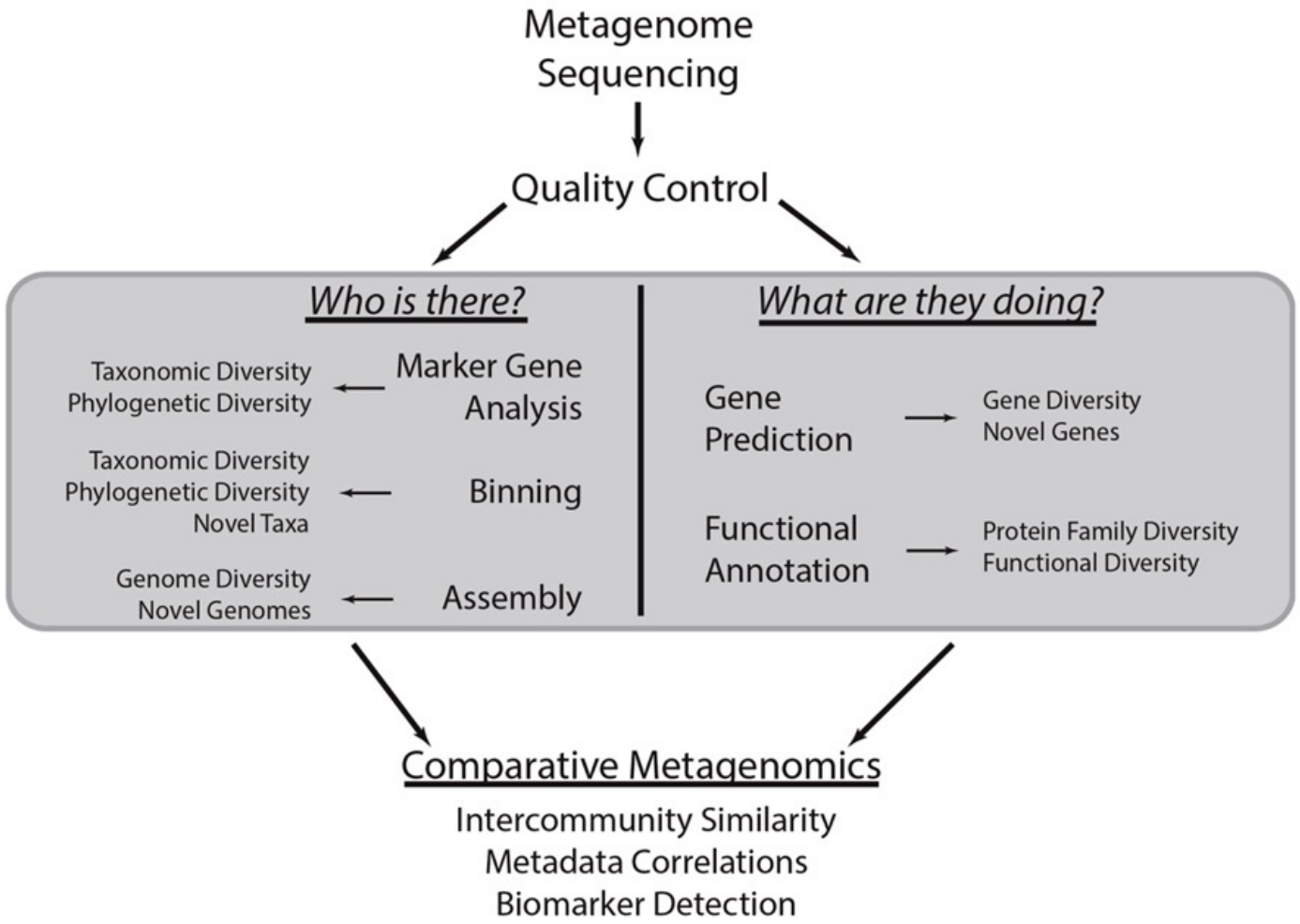
The human microbiome in health

- Digestive enzyme activity
- Synthesis of vitamins
- Interaction with immune system
- Protection from pathogens, etc.



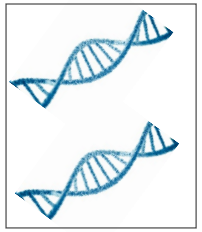
Kinross, J.M. Genome Medicine (2011)

Metagenomics: Sequencing Approaches to Study the Microbiome



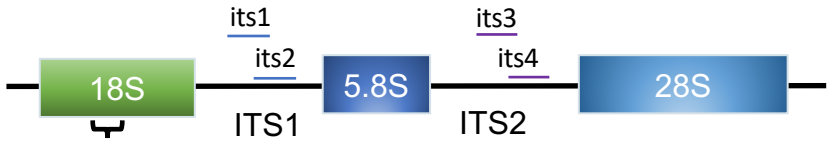
Sharpton, Thomas J. Frontiers in plant science (2014)

Sequencing Approaches



Targeted Approaches

16S: Targeting Archaea/Bacteria



18S: Targeting Eukaryotes
V4 has the most complete database info

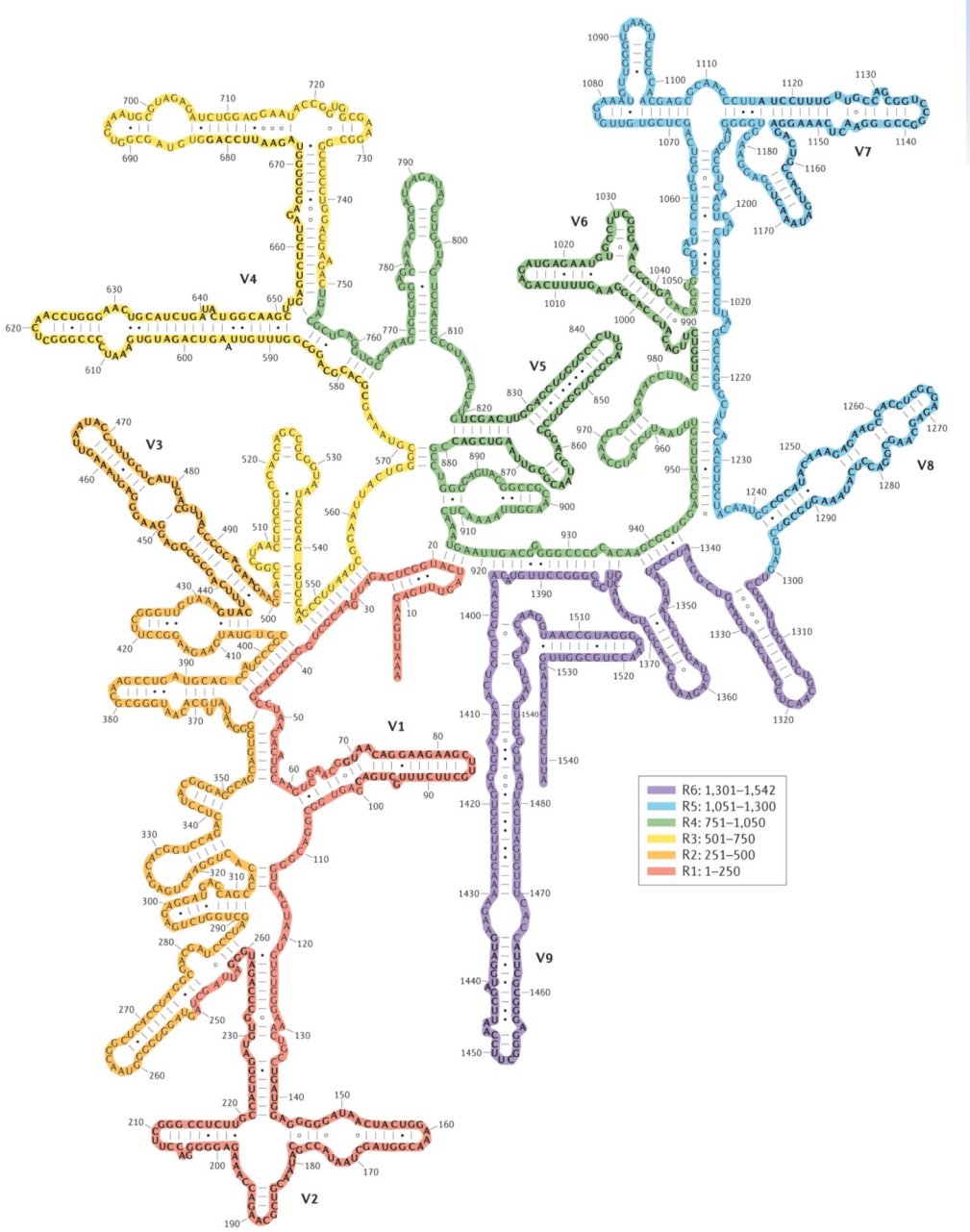
ITS: Targeting Fungi

Shotgun Approach

16S Amplicon Sequencing

Why 16S rRNA(Ribosomal RNA)?

- DNA → RNA → protein
- **Ribosomal RNA** (rRNA) is crucial for the survival of all living organisms.
- Genes that encode ribosomal RNA can serve as excellent **marker genes** since they are present in all living organisms.
- The **16S rRNA gene** is a DNA sequence corresponding to the rRNA in bacteria, found in the genomes of all bacterial species.
- The 16S rRNA is **highly conserved** throughout bacterial and other microbial evolution, earning it the designation of "the molecular fossil" of bacteria.

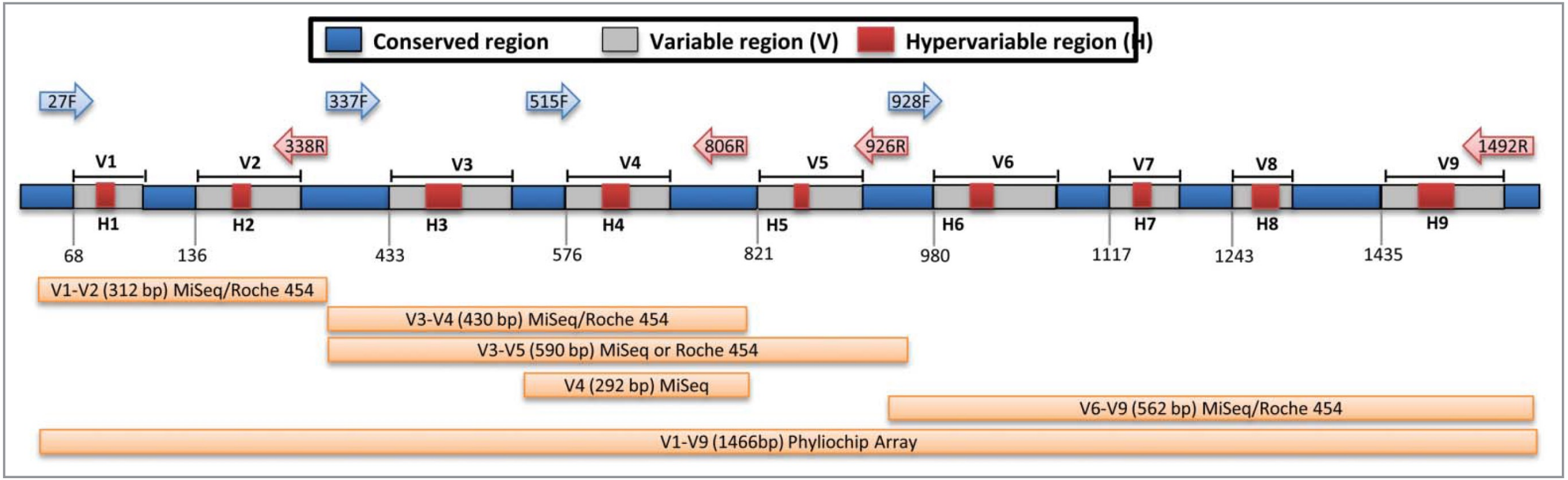


Yarza, P. et al. Nature Review Microbiology (2014)

Nature Reviews | Microbiology

16S Amplicon Sequencing

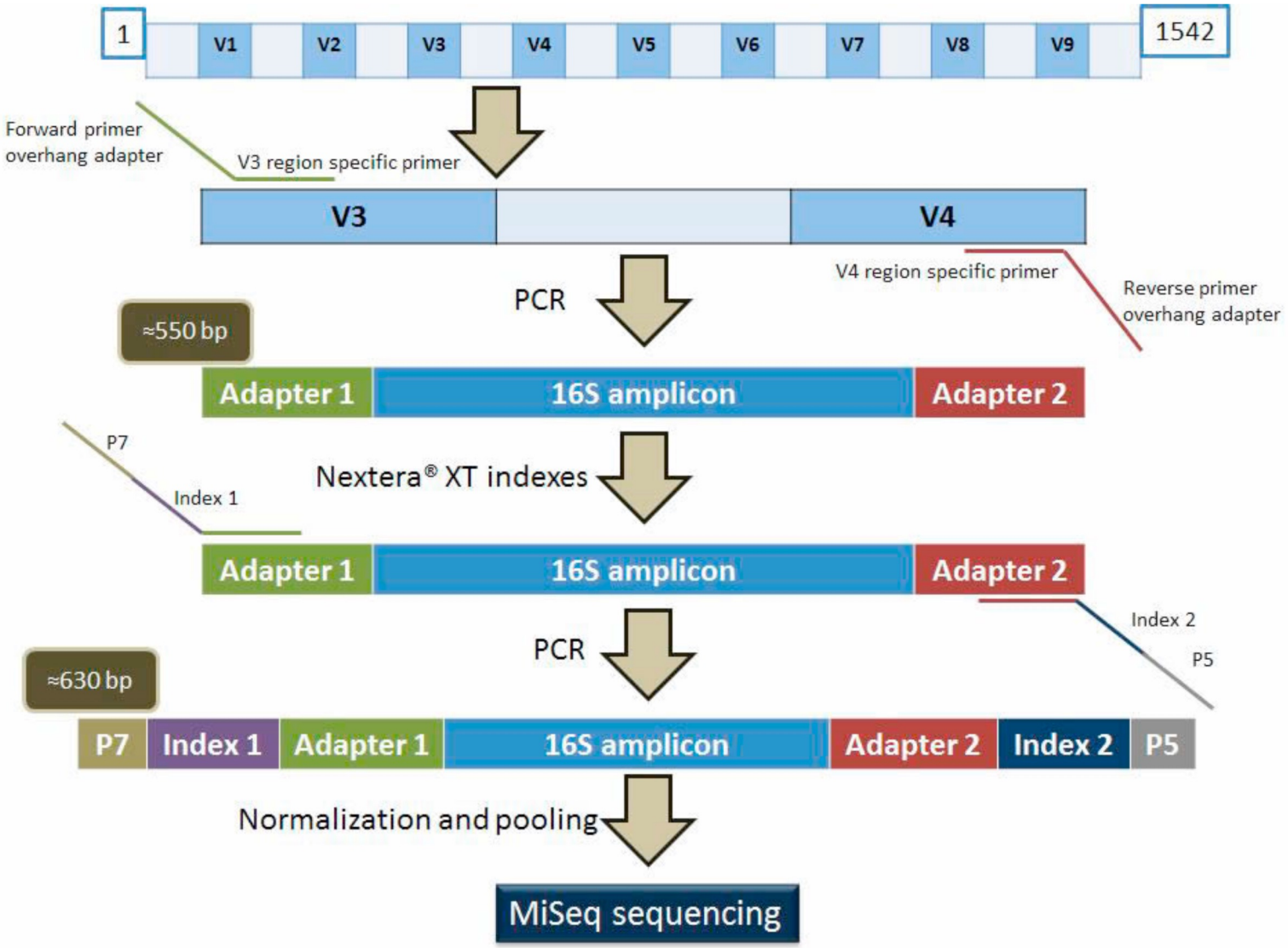
What?



Shahi, S. K., et. al, Gut Microbes (2017)

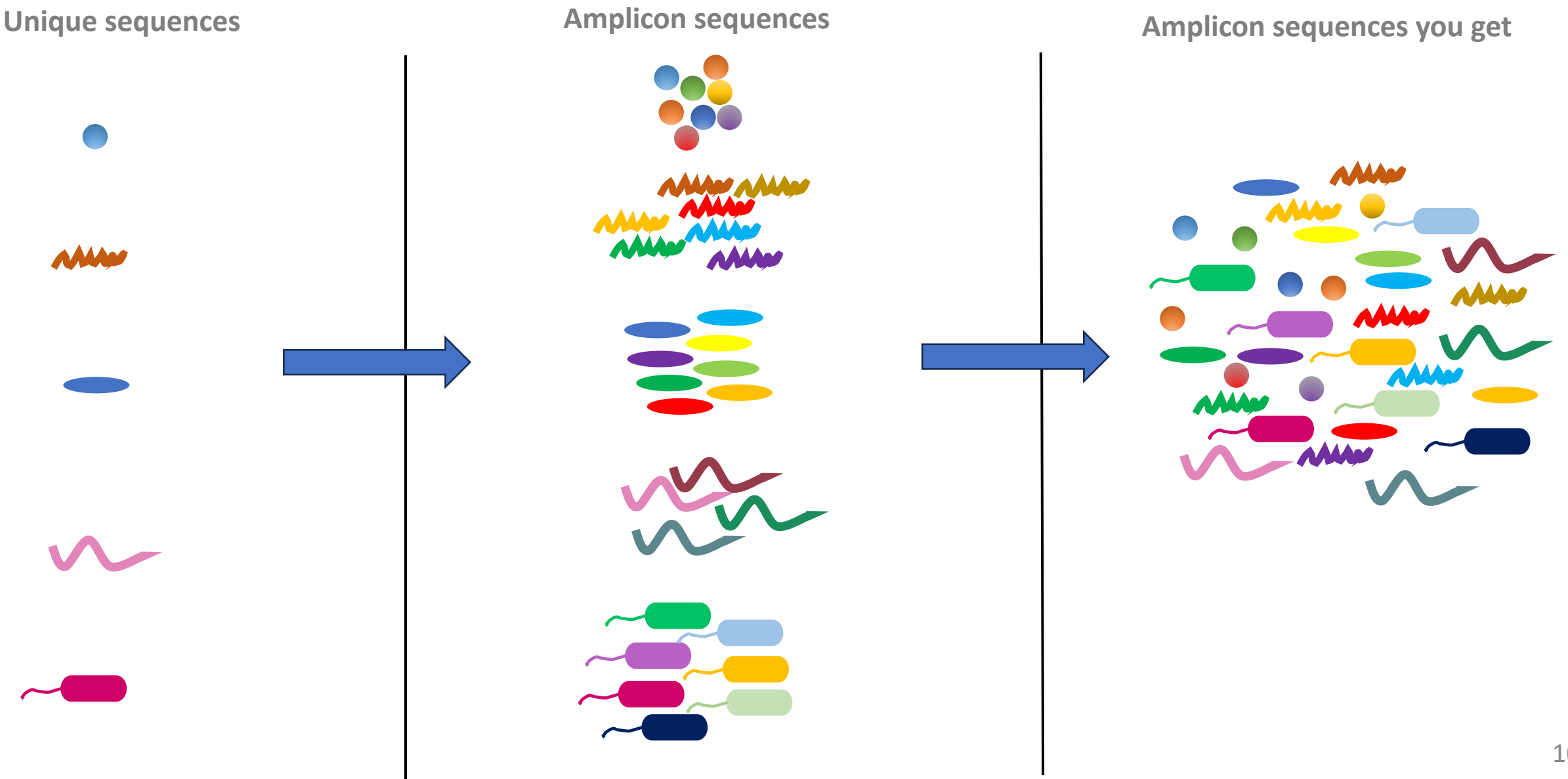
16S Amplicon Sequencing

How?



https://www.researchgate.net/profile/Emilio_Laserna-Mendieta/publication

16S Amplicon Sequencing



16S Amplicon Sequencing

Output(FASTQ)

Sequences

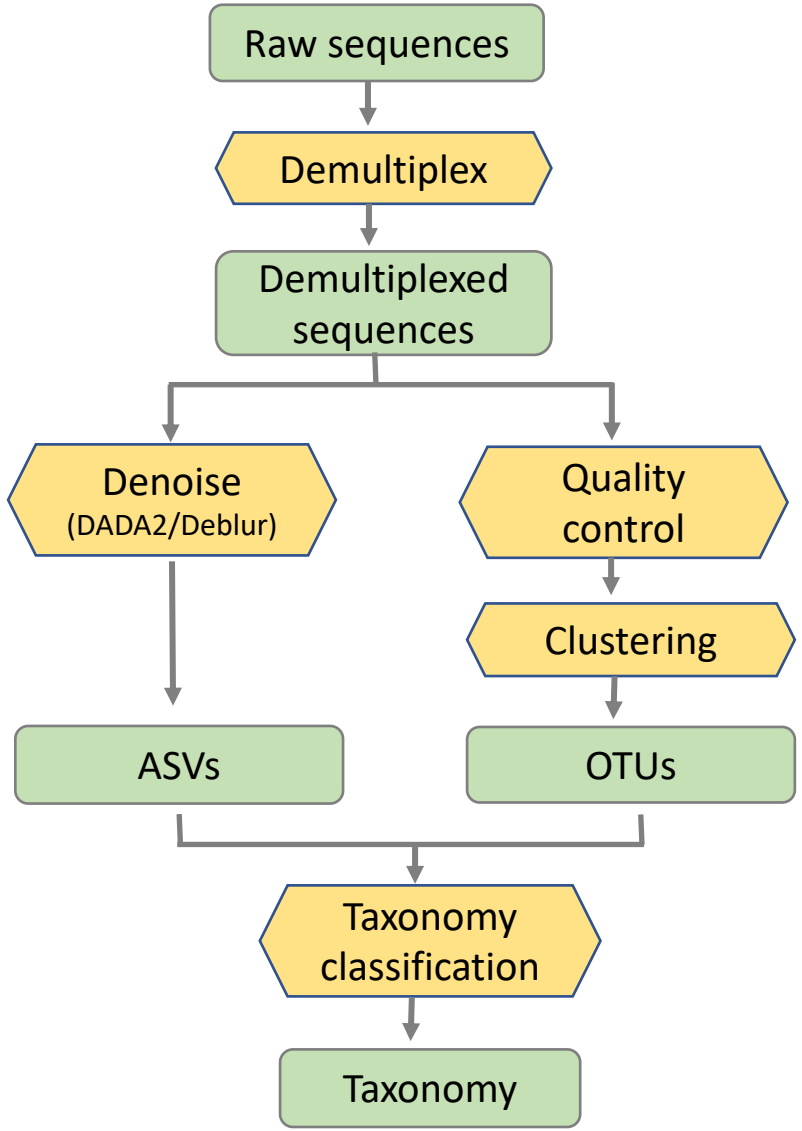
```
1 @HWI-EAS440_0386:1:23:17547:1423#0/1
2 TACGNAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAG
  TCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGATATCTTGAGTGCAG
  TTGAGGCAGGGGGGGATTGGTGTG
3 +
  IIIE)EEEEEEEEGFIIGIIIHIHHGIIIGIIHHHGIIHGHEGDGIFIGEHGIHHGHHGHHGGH
4 EEGHEGGEHEBBHBBEEDCEDDD>B?BE@@B>@@@@@CB@ABA@@?@@=>?08;3=;==8:5;@
  6?#####################
  @HWI-EAS440_0386:1:23:14818:1533#0/1
  CCCCNCAGCGGCAAAAATTAAAATTTTTACCGCTTCGGCGTTATAGCCTCACACTCAATCTTTT
  ATCACGAAGTCATGATTGAATCGCGAGTGGTCGGCAGATTGCGATAAACGGGCACATTAAATTT
  AAACTGATGATTCCACTGCAACAA
  +
  64<2$24;1)/:*B<?BBDDBBD<>BDD#################################
  ##########################################################
  ########################
```

@HWI-EAS440_0386	the unique instrument name
1	flowcell lane
23	tile number within the flowcell lane
17547	'x'-coordinate of the cluster within the tile
1423	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
1	the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

Symbol	ASCII Code	Q-Score
!	33	0
"	34	1
#	35	2
$	36	3
%	37	4
&	38	5
'	39	6
(	40	7
)	41	8
*	42	9
+	43	10
,	44	11
-	45	12
.	46	13
/	47	14
0	48	15
1	49	16
2	50	17
3	51	18
4	52	19
5	53	20

Symbol	ASCII Code	Q-Score
6	54	21
7	55	22
8	56	23
9	57	24
:	58	25
;	59	26
<	60	27
=	61	28
>	62	29
?	63	30
@	64	31
A	65	32
B	66	33
C	67	34
D	68	35
E	69	36
F	70	37
G	71	38
H	72	39
I	73	40

16S Amplicon Sequencing



qiime2

mothur

MG-RAST metagenomics analysis server

UPARSE OTU clustering

CLC Main Workbench

.....

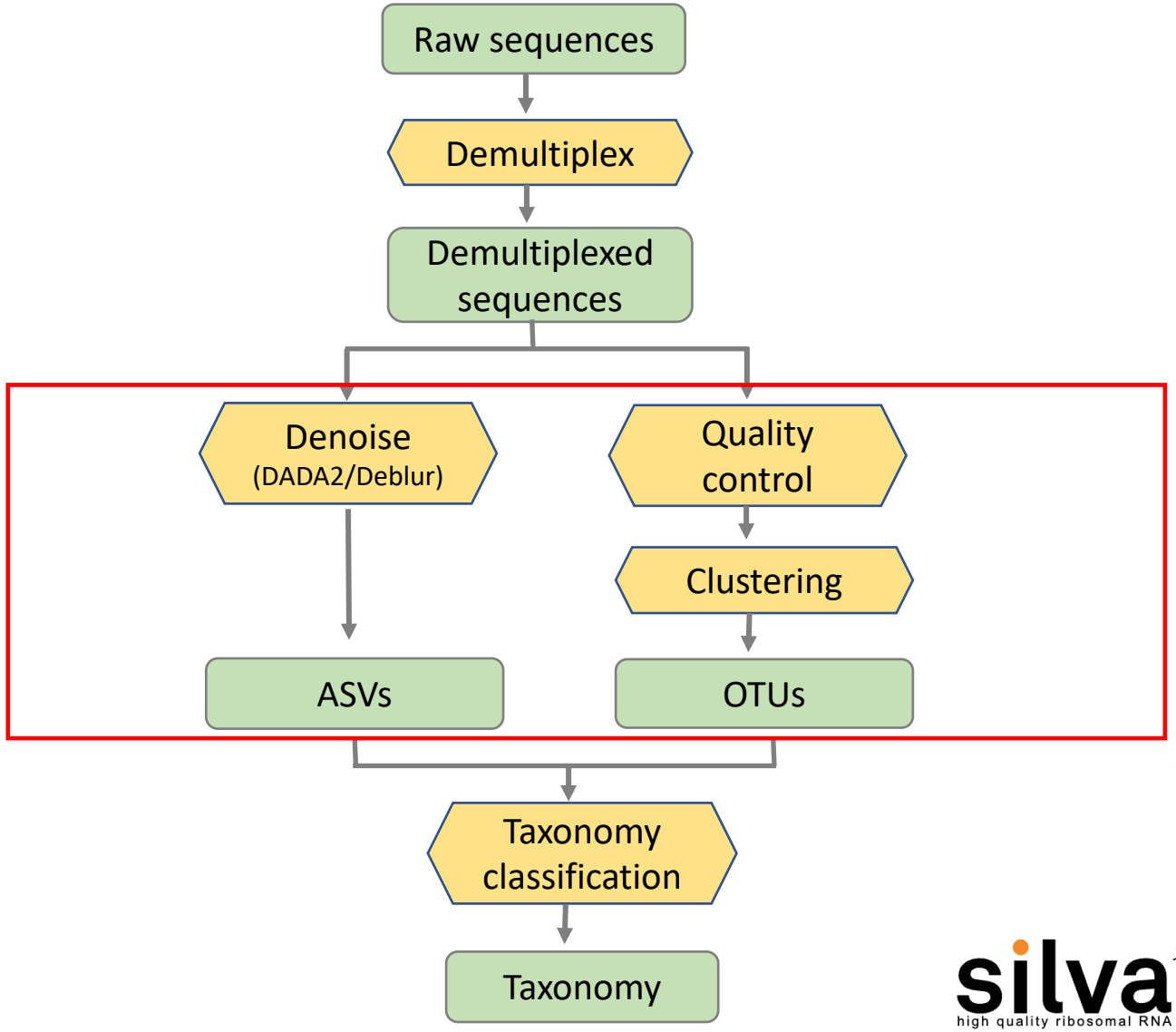
silva high quality ribosomal RNA databases

rdp

Welcome to GTDB
GENOME TAXONOMY DATABASE
596,859 genomes
Release 09-RS220 (24th April 2024)

GREENGENES

16S Amplicon Sequencing



qiime2

mothur

MG-RAST metagenomics analysis server

UPARSE OTU clustering

CLC Main Workbench

.....

silva high quality ribosomal RNA databases

rdp

Welcome to GTDB
GENOME TAXONOMY DATABASE
596,859 genomes
Release 09-RS220 (24th April 2024)

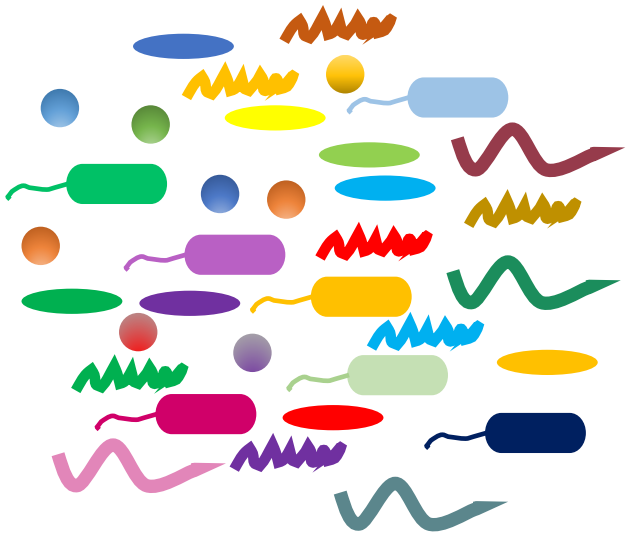
GREENGENES

16S Amplicon Sequencing

Why picking OTU/ASV?

- Sequencing is far from error-free. (i.e., base call errors)
- Tolerate/correct sequencing errors
- Reduce the dimensionality of the data

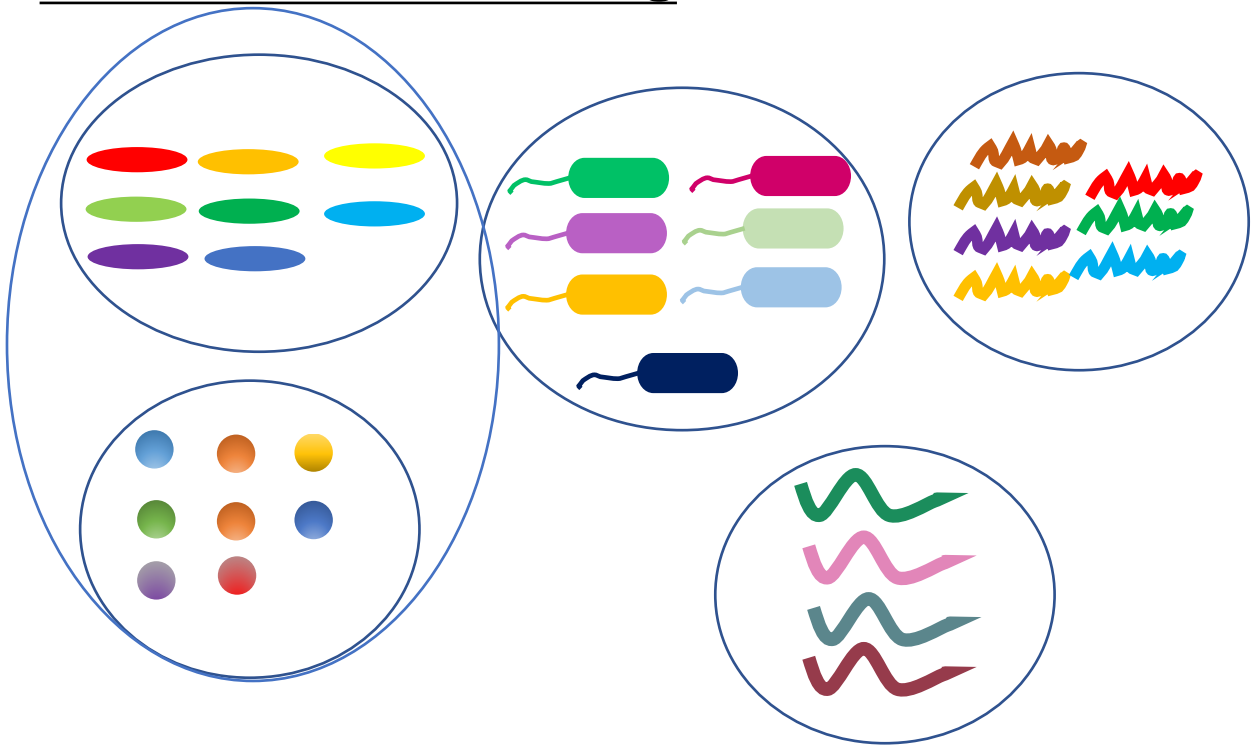
Sequences to be grouped/clustered



Number of sequence variants $\stackrel{?}{=}$ number of species

16S Amplicon Sequencing

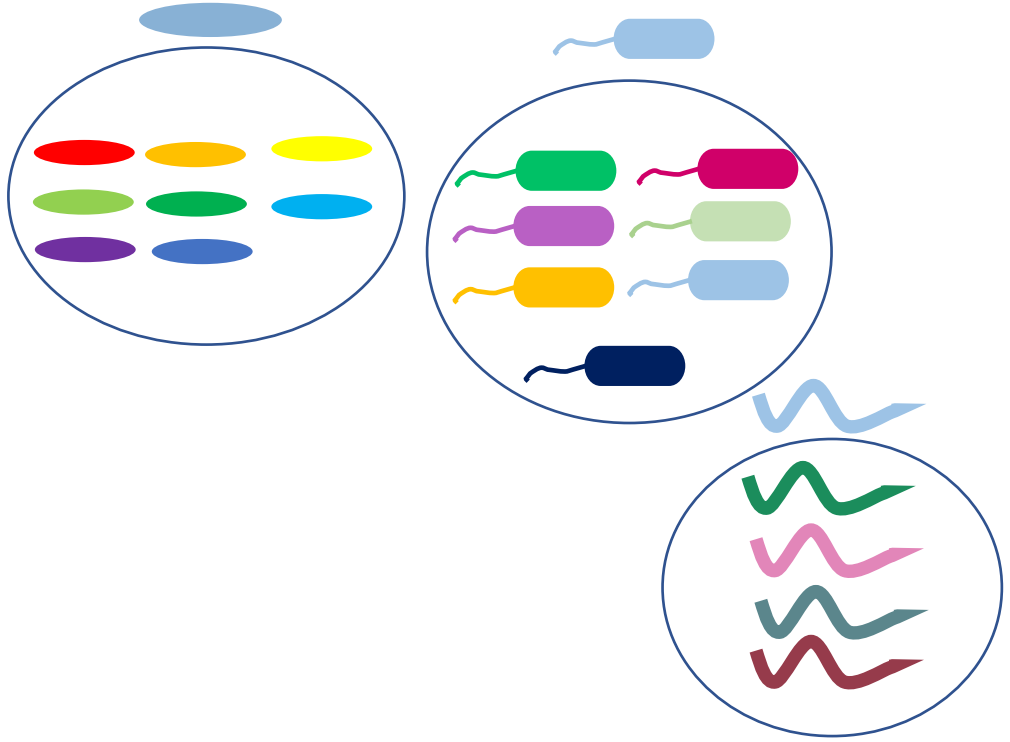
De Novo OTU Picking



- Reads are clustered against each other without using an external reference sequence collection. This process must be repeated when new sequences are added.
- No reads are lost.
- Suitable for situations where a reference sequence collection is unavailable for clustering.
- Processing large datasets can be slow.

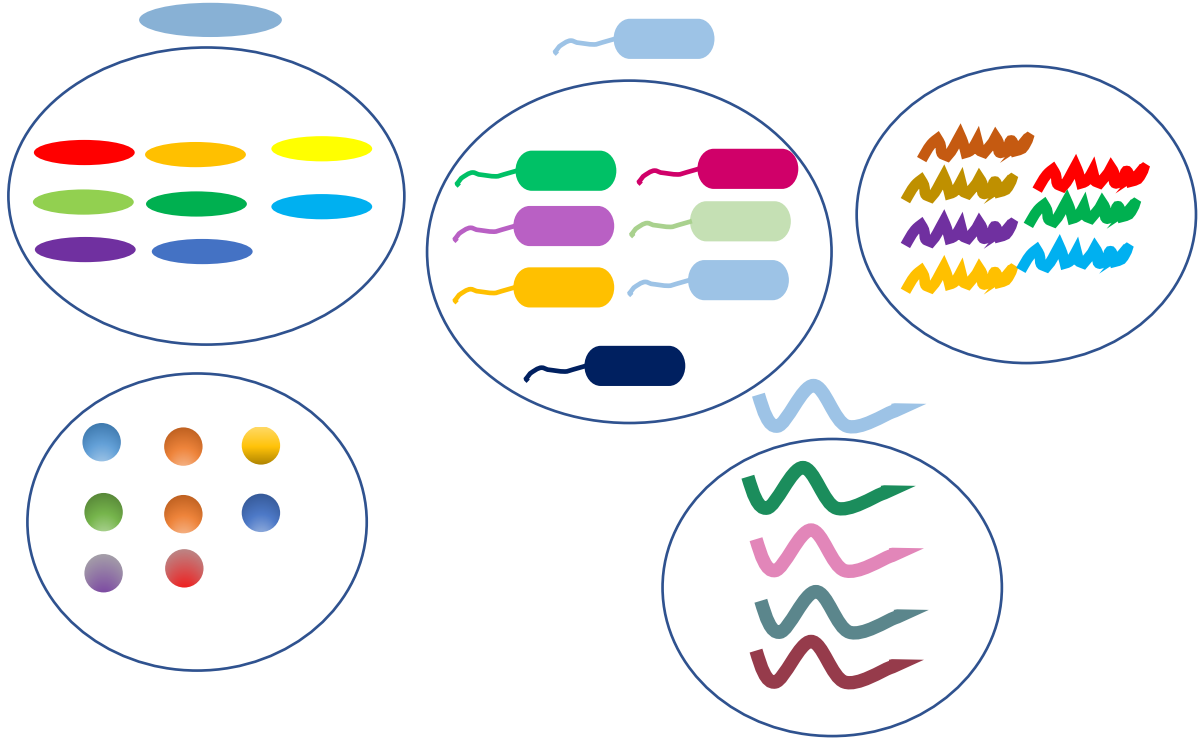
16S Amplicon Sequencing

Closed-reference OTU Picking



- Reads are clustered against a reference sequence collection, and any reads that do not match a sequence in the reference collection are excluded from further analysis.
- This method offers speed benefits, especially for analyzing large datasets.
- It is not suitable for detecting novel diversity.
- Not applicable if a reference sequence collection is unavailable for comparison.

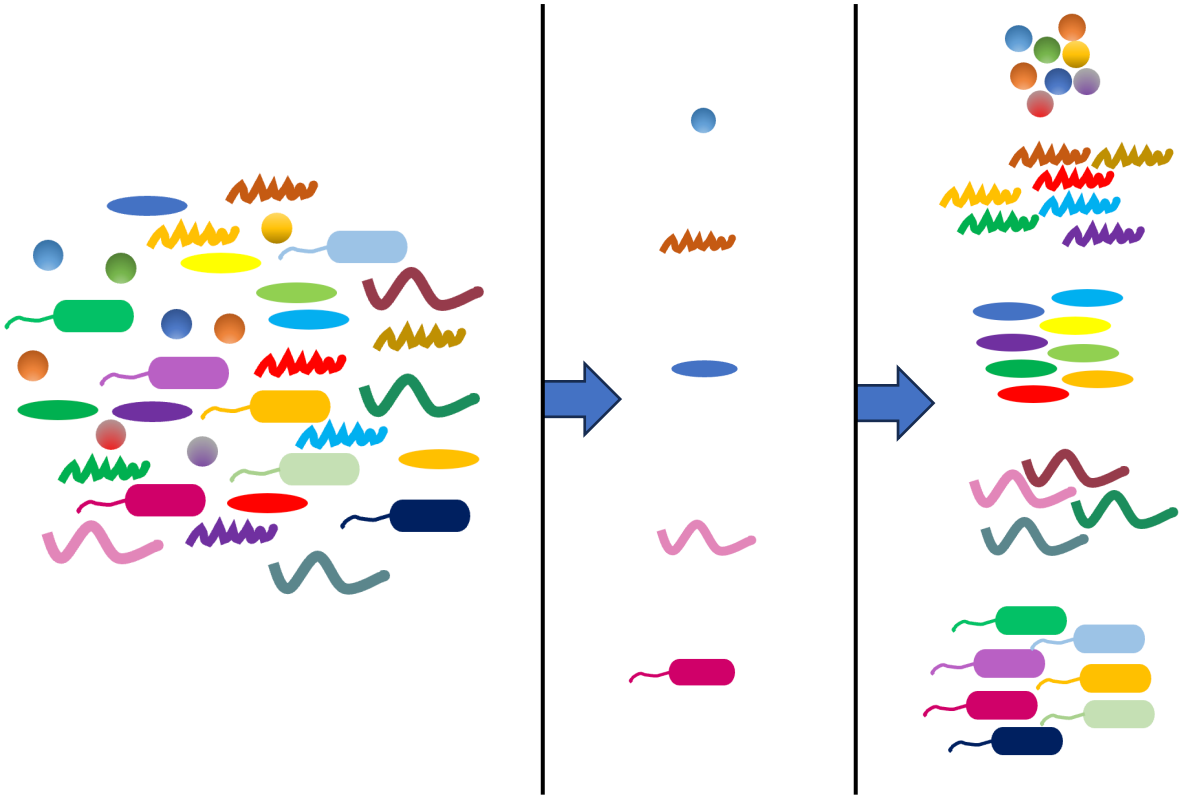
Open-reference OTU Picking



- Reads are clustered against a reference sequence collection, and any reads that do not match the reference collection are clustered de novo.
- All reads are included in the clustering process.
- This method is faster if a large portion of sequences match the reference database; otherwise, it remains slow, similar to de novo clustering.

16S Amplicon Sequencing

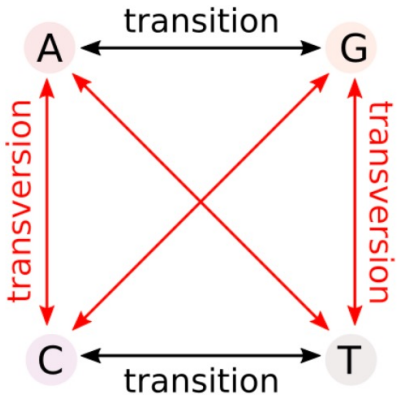
Error model based ASV inference(DADA2/DEBLUR)



- Statistically infer whether each sequence is an artifact or a real (error-free) sequence.
- Real amplicon sequence variants, not representative consensus sequences.
- Not reference-based.
- Facilitate comparison across studies.

Error model based ASV picking (DADA2)

Partition:	Parwise alignments - center *vs* unique reads	Abundance	Consensus Q score (at 25 bp)
Center:	ATGCATGCTACGTGCAGCTAGCTC**A**CATGCTAG...	10k reads	Q30
Unique2:	ATGCATGCTACGTGCAGCTAGCTC**G**CATGCTAG...	150 reads	Q3
Unique3:	ATGCATGCTACGTGCAGCTAGCTC**T**CATGCTAG...	50 reads	Q2
Unique4:	ATGCATGCTACGTGCAGCTAGCTC**C**CATGCTAG...	25 reads	Q1
...	...	...	...
Position:	1 5 10 15 20 25 30 ...	...	...



transition probabilities

transitions \ Q scores (16 … 41)	0	1	...	40
A2A	0.99	0.99	0.99	0.99
A2T	0.00	0.00	0.00	0.00
...	...	...	...	...
C2C	0.99	0.99	0.99	0.99

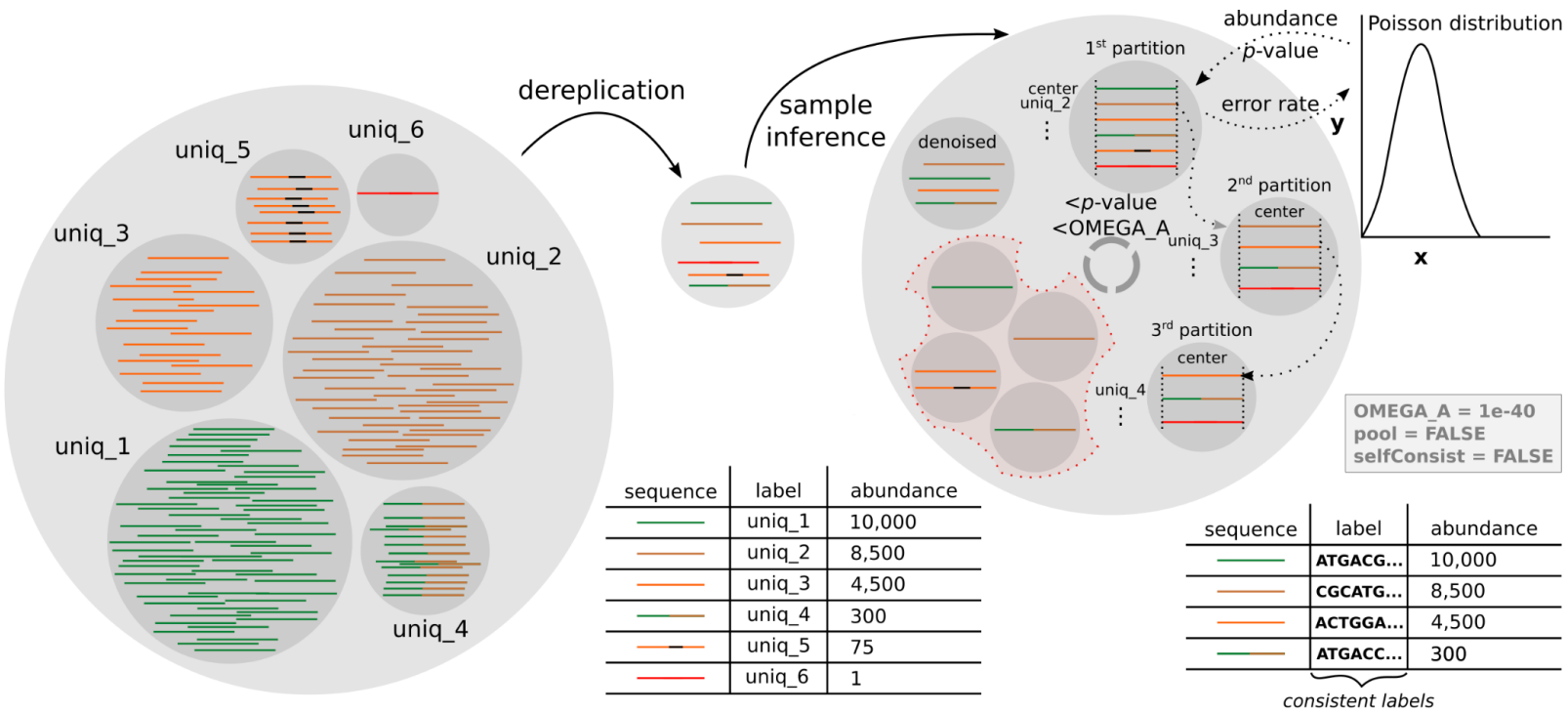
nbases = 1e+08

$\frac{1\text{x}10^8 \text{ bases}}{250 \text{ bp long}}$ = $4\text{x}10^5$ reads

estimate error rates - *learnErrors()*

https://igcbioinformatics.github.io/biomeshinycourse/pages/dada2/Biodata.ptCrashCourses.html

Error model based ASV picking (DADA2)



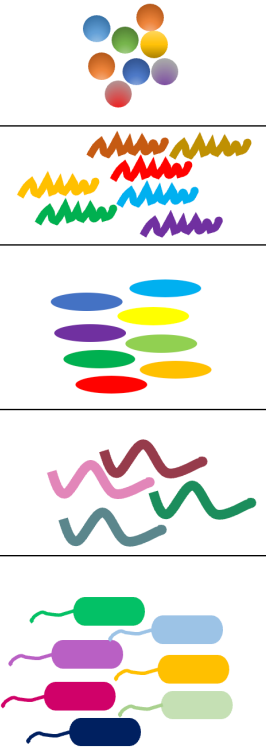
denoise unique sequences (dereplicate and exclude singletons) - *dada()*

https://igcbioinformatics.github.io/biomeshinycourse/pages/dada2/Biodata.ptCrashCourses.html

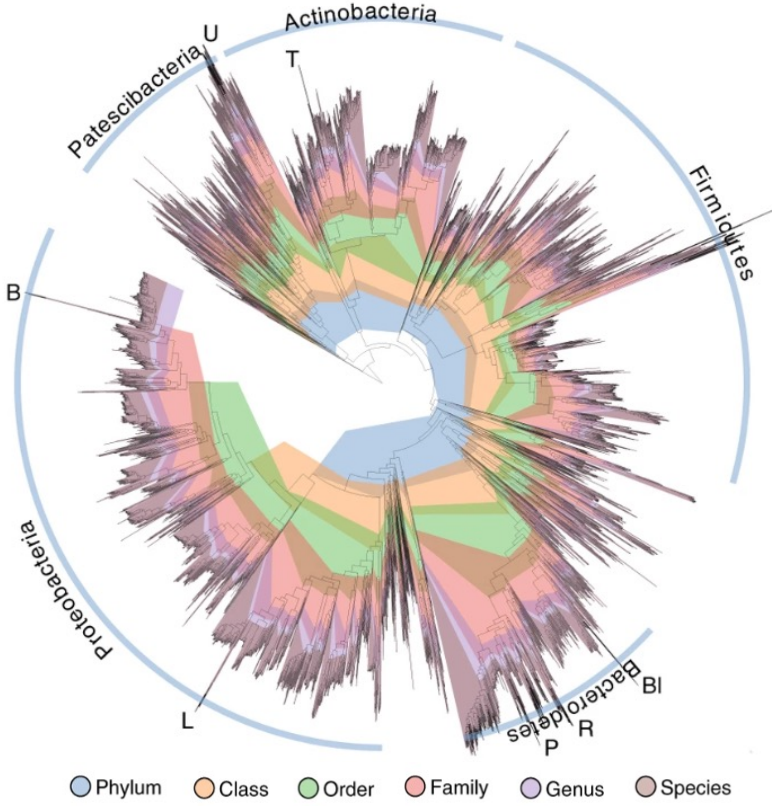
16S Amplicon Sequencing

Phylogenetics and taxonomy

ASV	Counts
ATGACG…	8000
CGCATG…	5000
ACTGGA…	2000
ATGACC…	4000
CTCACG…	600



?



Donovan H Parks. Nucleic Acids Research (2022)

silva
high quality ribosomal RNA databases

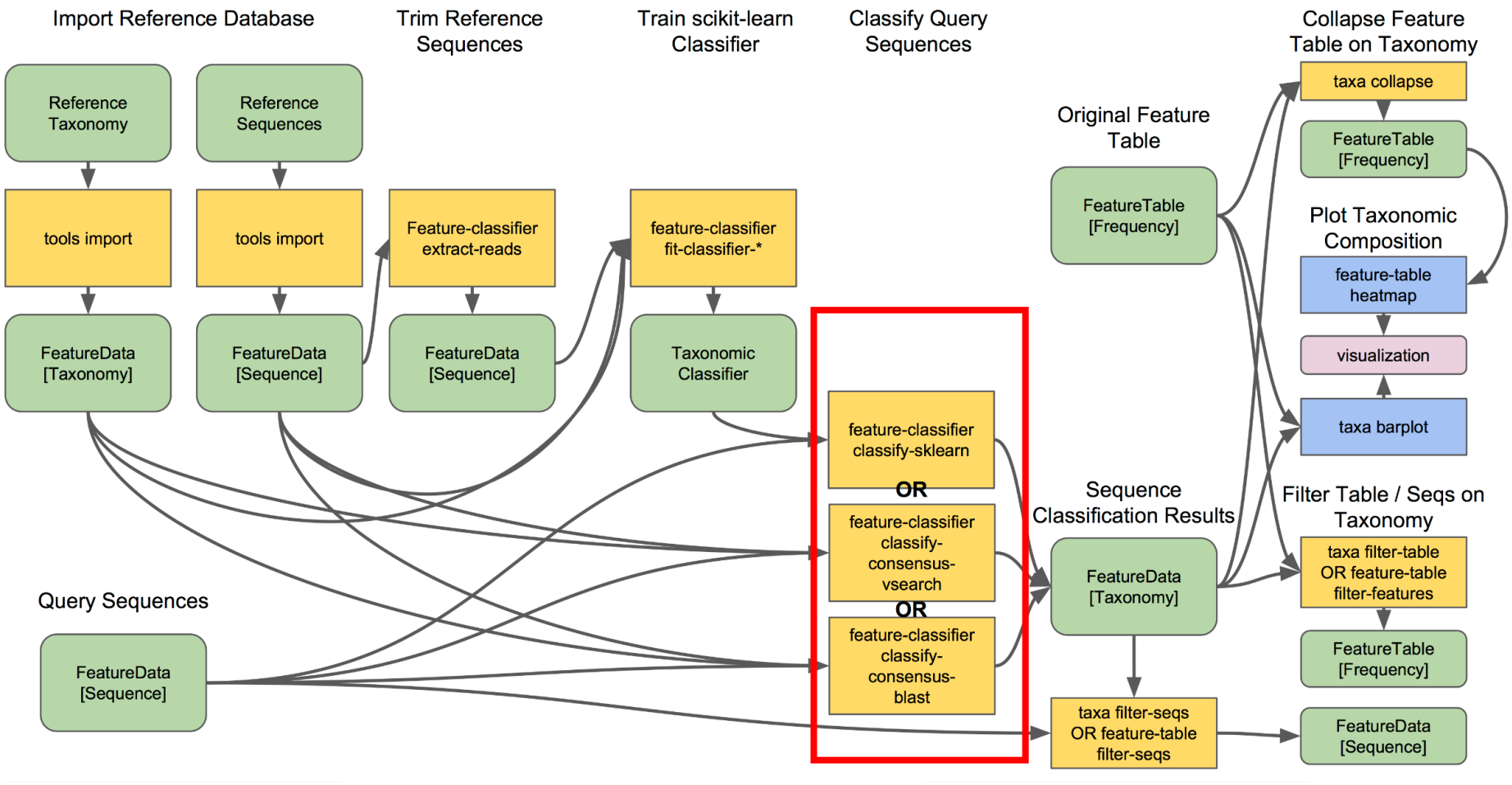
Welcome to GTDB
GENOME TAXONOMY DATABASE
596,859 genomes
Release 09-RS220 (24th April 2024)

GREENGENES

rdp

16S Amplicon Sequencing

Phylogenetics and taxonomy



https://docs.qiime2.org/2024.5/tutorials/overview

MAYO CLINIC

Bioinformatic products and metadata

Taxonomy Table

OTU.id	Kingdom	Phylum	Class	Order	Family	Genus	Species
368907	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	Propionibacteriumacnes
12724	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	Propionibacteriumgranulosum
93749	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	Propionibacteriumacidipropionici
378053	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Tessaracoccus	*NA*

OTU/ASV Table

OTU.id	sample1	sample2	sample3	sample4	sample5
36155	33783	654	9	19	10
47916	22598	2279	12	5	11
279297	17373	1540	7	7	4
573135	14062	14173	50	5	7
578268	13903	677	6	3	3
27669	12801	3702	252	1	1
557785	11082	627	3	5	3

Sample Metadata

SampleID	SampleType	Description
sample1	Soil	Calhoun South Carolina Pine soil, pH 4.9
sample2	Soil	Cedar Creek Minnesota, grassland, pH 6.1
sample3	Soil	Sevilleta new Mexico, desert scrub, pH 8.3
sample4	Feces	M3, Day 1, fecal swab, whole body study
sample5	Feces	M1, Day 1, fecal swab, whole body study

16S Amplicon Sequencing

Bioinformatic products and metadata

Taxonomy Table

OTU.id	Kingdom	Phylum	Class	Order	Family	Genus	Species
368907	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	Propionibacteriumacnes
12724	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	Propionibacteriumgranulosum
93749	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	Propionibacteriumacidipropionici
378053	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Tessaracoccus	*NA*

OTU/ASV Table

OTU.id	sample1	sample2	sample3	sample4	sample5
36155	33783	654	9	19	10
47916	22598	2279	12	5	11
279297	17373	1540	7	7	4
573135	14062	14173	50	5	7
578268	13903	677	6	3	3
27669	12801	3702	252	1	1
557785	11082	627	3	5	3

Sample Metadata

SampleID	SampleType	Description
sample1	Soil	Calhoun South Carolina Pine soil, pH 4.9
sample2	Soil	Cedar Creek Minnesota, grassland, pH 6.1
sample3	Soil	Sevilleta new Mexico, desert scrub, pH 8.3
sample4	Feces	M3, Day 1, fecal swab, whole body study
sample5	Feces	M1, Day 1, fecal swab, whole body study

16S Amplicon Sequencing

MAYO CLINIC

Cons of 16S

- Taxonomy resolution: does not really get to the species/strain level.
- No functional profiling (only prediction with PICRUSt, Tax4Fun).
- Amplicon bias and primer bias: if the primer does not match, sequencing will fail.
- No cross-domain analysis, only bacteria.

Shotgun Metagenomic Sequencing

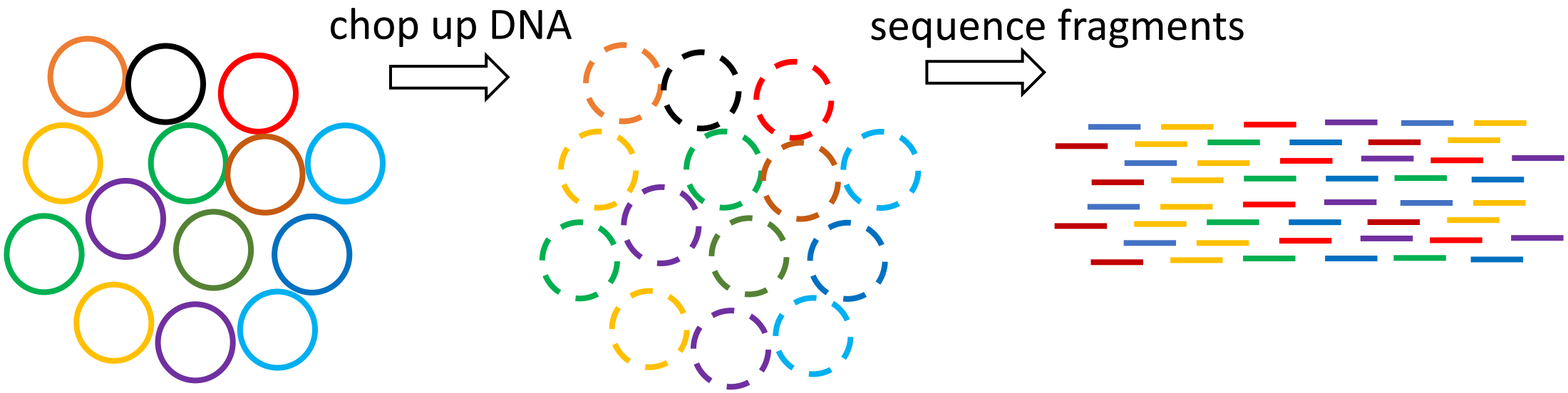
MAYO CLINIC

Why?

- Comprehensively samples all genes in all organisms present in given complex samples.
- Identifies rare or novel organisms.
- High taxonomy resolution.
- Cross-domain, including viruses and fungi.
- No amplification bias.
- Functional profiling.
- Can study antibiotic-resistant genes.
- More expensive.
- Computation-intensive.
- …

Shotgun Metagenomic Sequencing

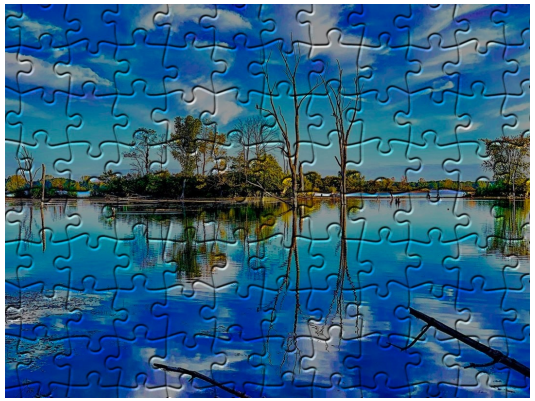
What?

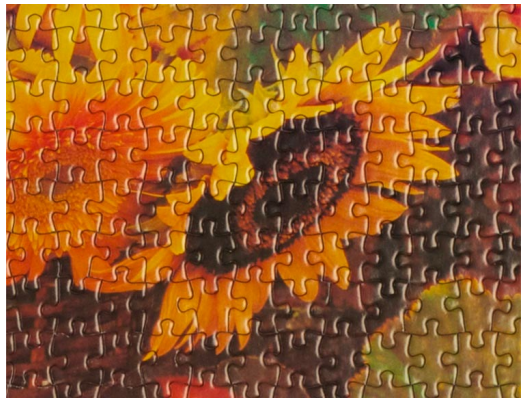


- Microbial DNA is fragmented and sequenced directly.
- Infer the taxonomic and functional content based on the sequence fragments.

Shotgun Metagenomic Sequencing

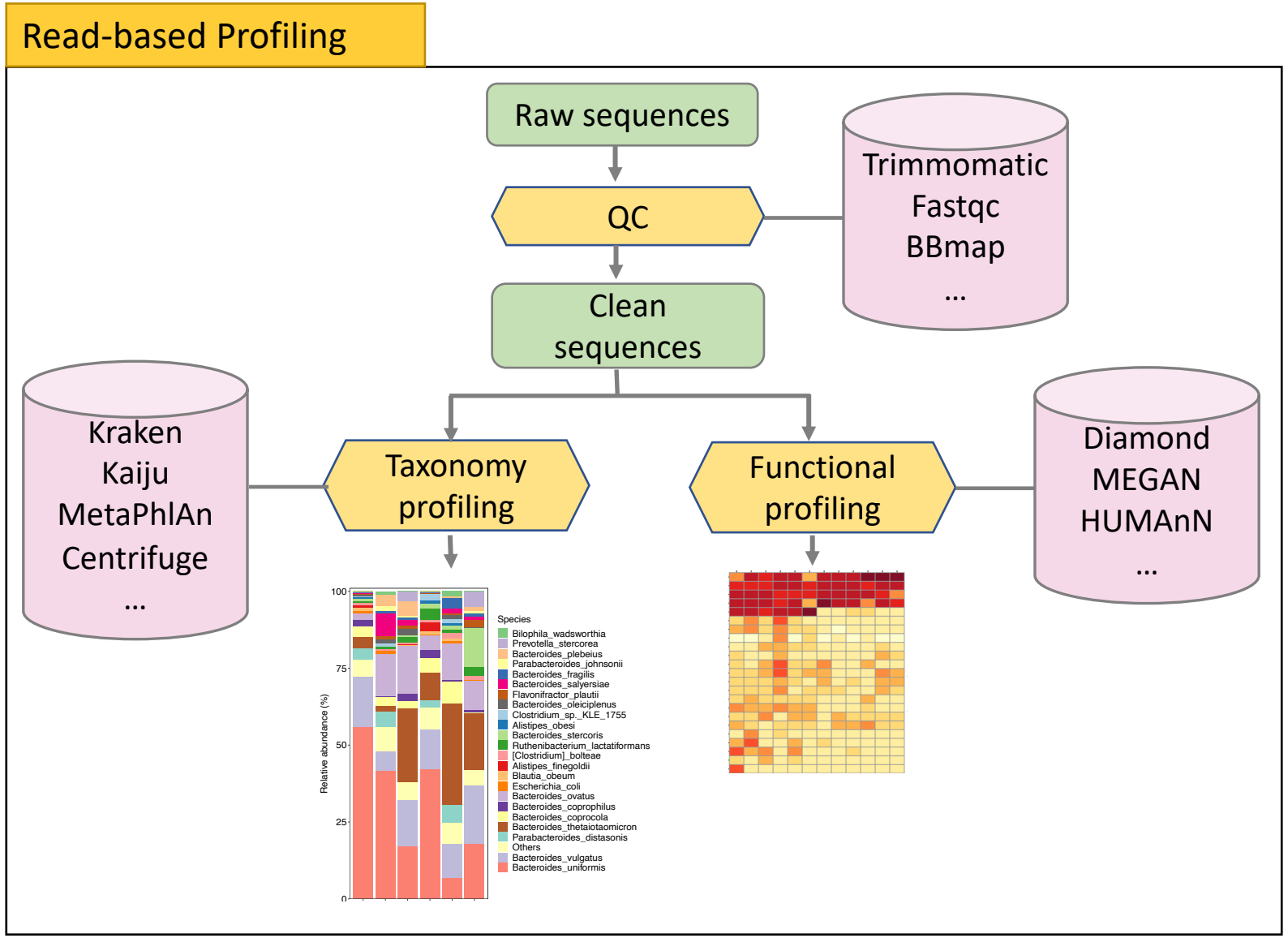
What?





Shotgun Metagenomic Sequencing

How?



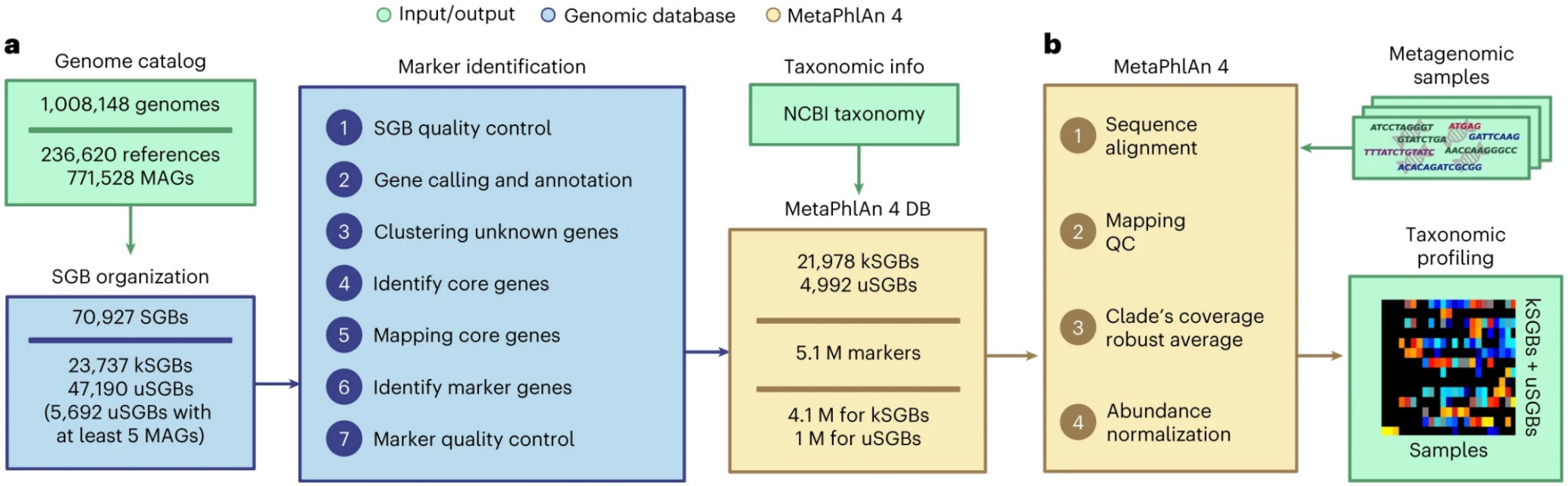
Shotgun Metagenomic Sequencing

Type	Classifier	Custom Databases	Generates Abundance Profile	Memory Required	Time Required
DNA	Bracken	yes	yes	<1 Gb	<1 min
	Centrifuge	yes	yes	20 Gb	7 min
	CLARK	yes	yes	80 Gb	2 min
	CLARK-S	yes	yes	170 Gb	40 min
	Kraken	yes	yes	190 Gb	1 min
	Kraken2	yes	yes	36 Gb	1 min
	KrakenUniq	yes	yes	200 Gb	1 min
	k-SLAM	yes	yes	130 Gb	2 h
	MegaBLAST	yes	no	61 Gb	4 h
	metaOthello	no	no	30 Gb	1 min
	PathSeq	yes[a]	no	140 Gb	5 min
	prophyle	yes	no	40 Gb	40 min
	taxMaps	yes	yes	65 Gb	25 min
Protein	DIAMOND	yes	no	110 Gb (varies)	10 min
	Kaiju	yes	yes	25 Gb	1 min
	MMseqs2	yes	no	85 Gb (varies)	9 h
Markers	MetaPhlAn2	no	yes	2 Gb	1 min
	mOTUs2	no	yes	2 Gb	1 min

Simon, H. Ye, et al. Cell (2019)

Shotgun Metagenomic Sequencing

MetaPhlAn4

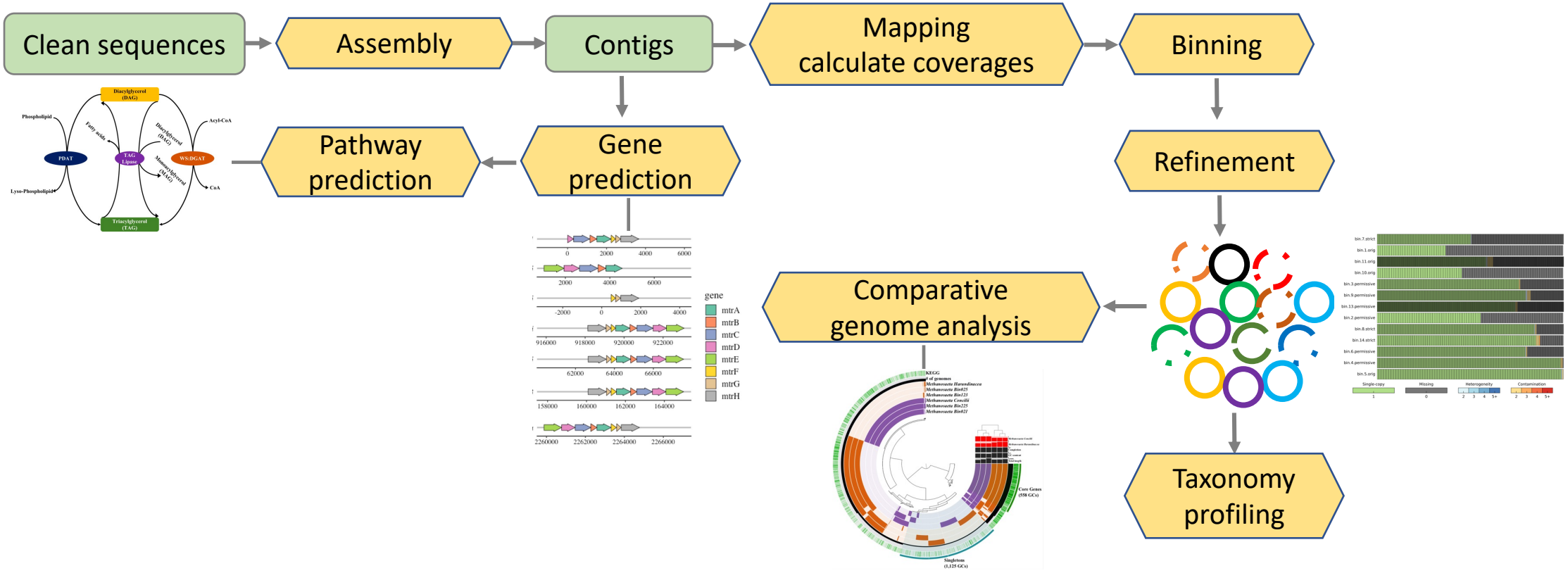


Blanco-Míguez, A., et al. Nature Biotechnology (2023)

Shotgun Metagenomic Sequencing

How?

Assembly-based Analysis

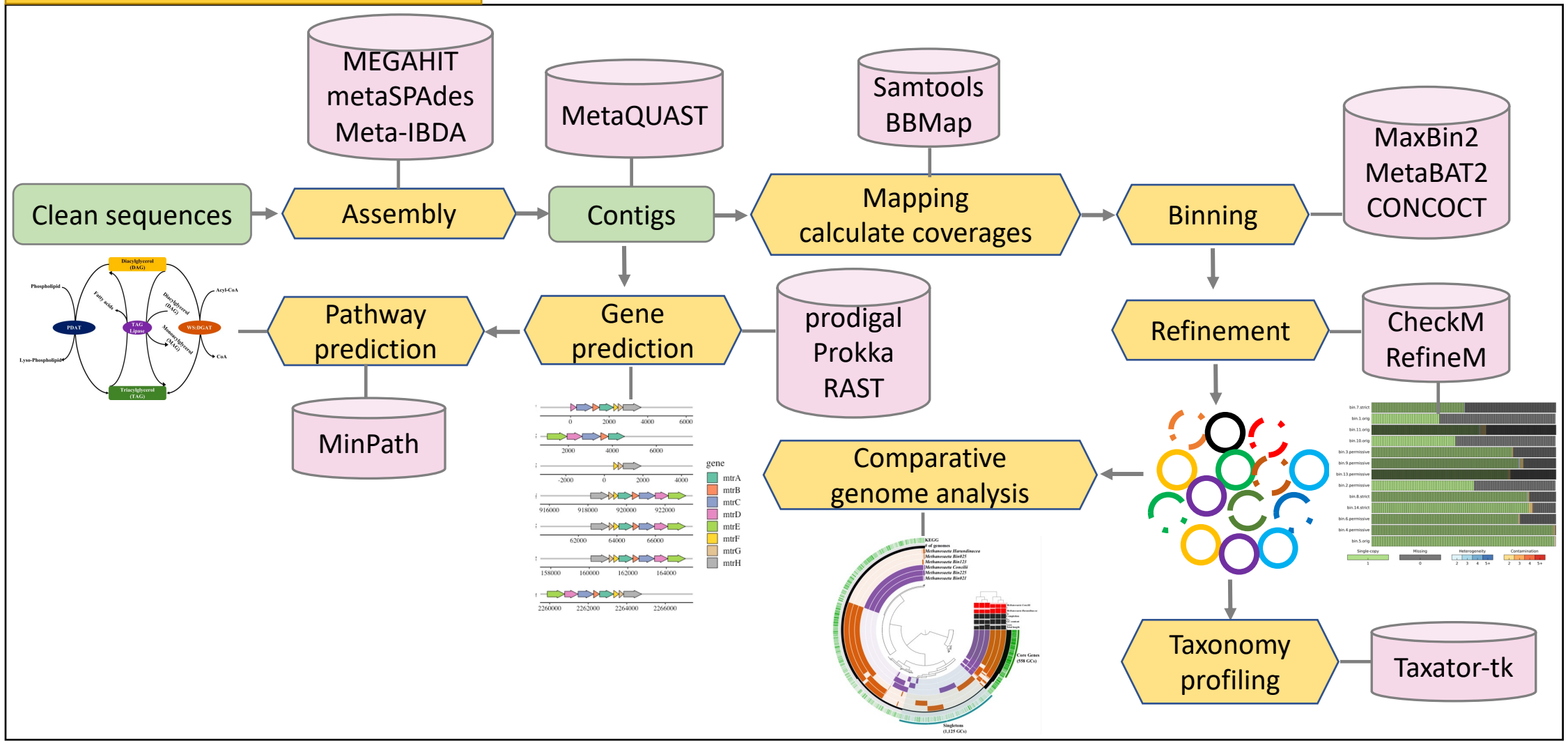


Shotgun Metagenomic Sequencing

MAYO CLINIC

How?

Assembly-based Analysis

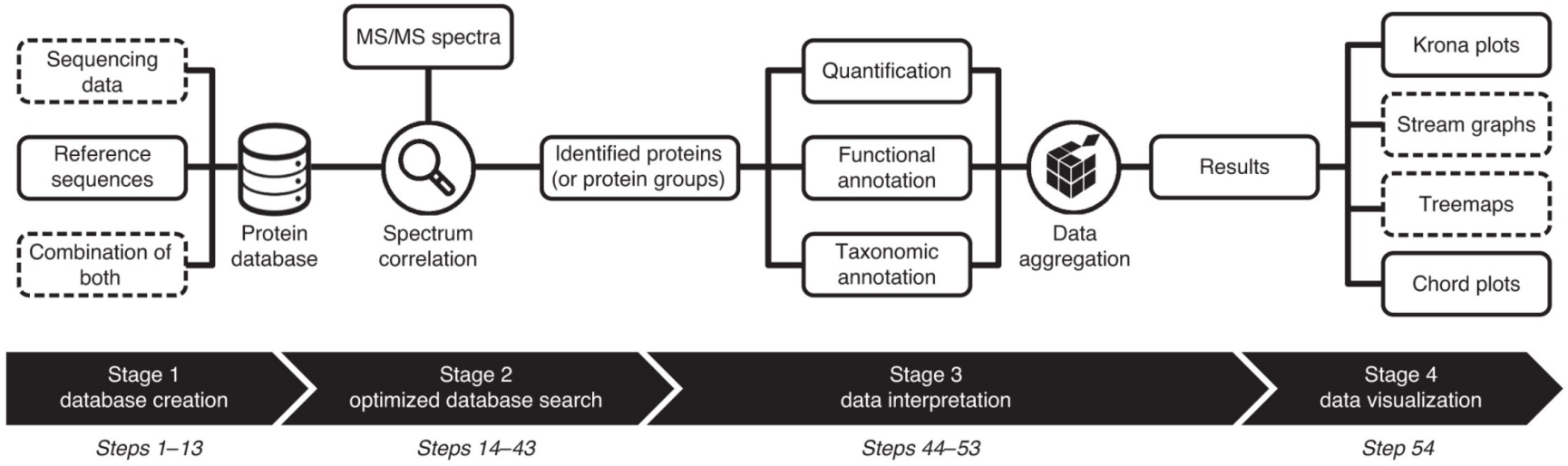


Other Approaches

Metaproteomics

- The study of the collective protein composition of multi-organism systems.
- **Goals**: Identify and quantify all proteins present in complex samples, such as the human gut.
- **Insights**: provides deep insights into the biodiversity of microbial communities and the complex functional interplay between microbes and their hosts or environment.

From: A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and Prophane

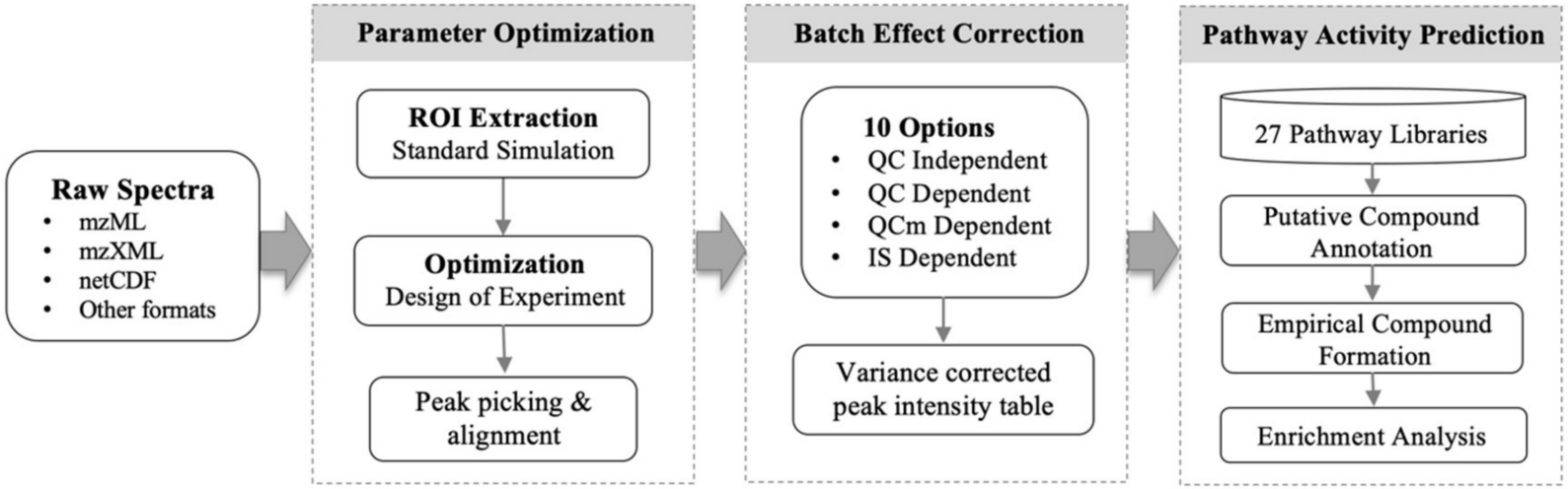


Schiebenhoefer, H. et al. Nature Protocol (2020)

Other Approaches

Metabolomics

- Large-scale study of small molecules, commonly known as metabolites, within cells, biofluids, tissues or organisms.
- **Goals**: Identify and quantify the complete set of metabolites in a biological sample.
- **Insights**: Provides biochemical activity and physiological state insights, aiding in the understanding of metabolic pathways, disease mechanisms, and the effects of drugs or environmental changes.

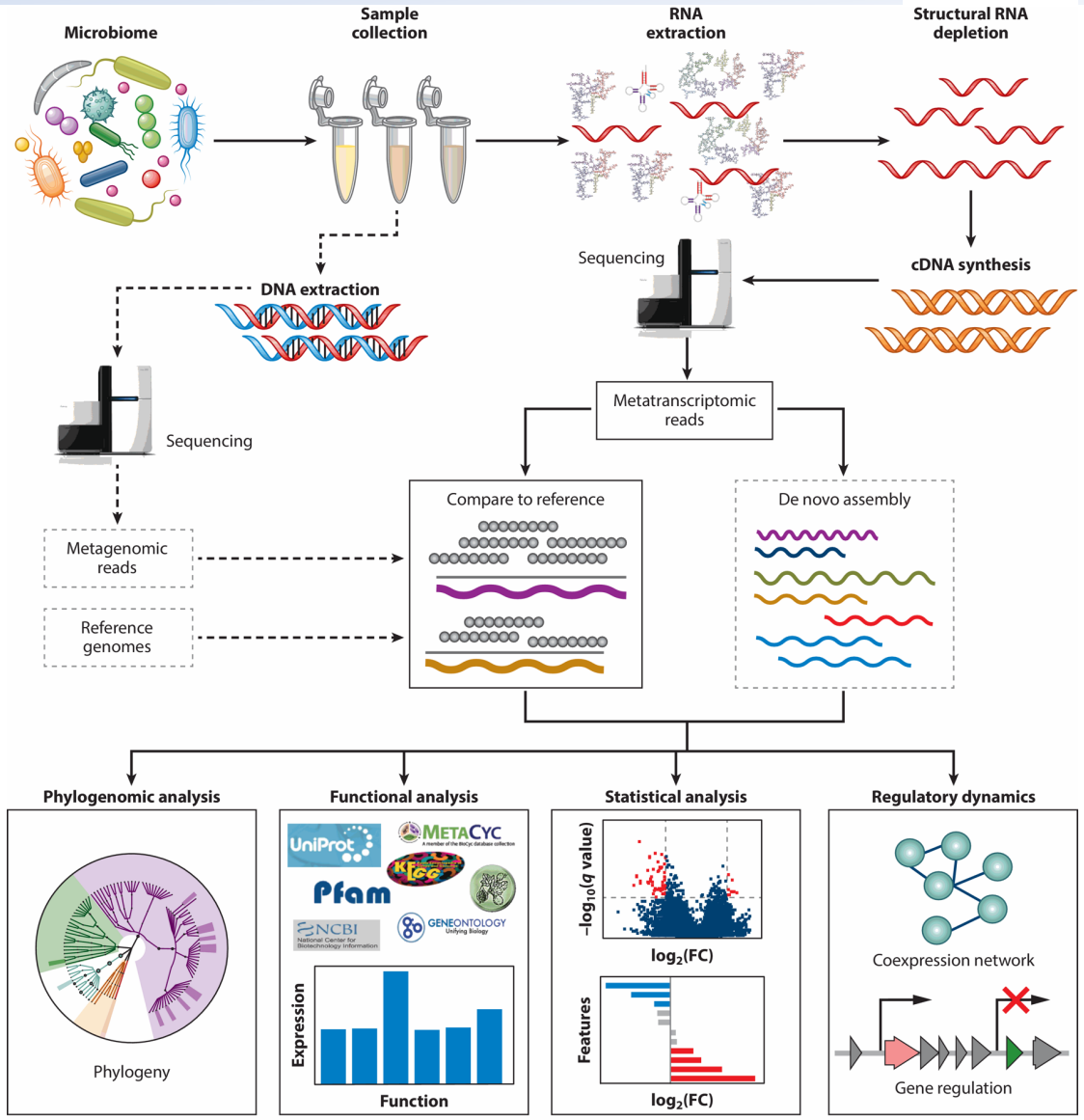


Pang, Z. et al., Metabolites (2020)

Other Approaches

Metatranscriptomics

- Survey microbial community gene function and regulation at scale.
- **Goals**: Understand the active metabolic and regulatory processes by analyzing expressed genes.
- **Insights**: Provides information on the functional dynamics and interactions of the community under various conditions, helping to link gene expression with microbial functions and responses.



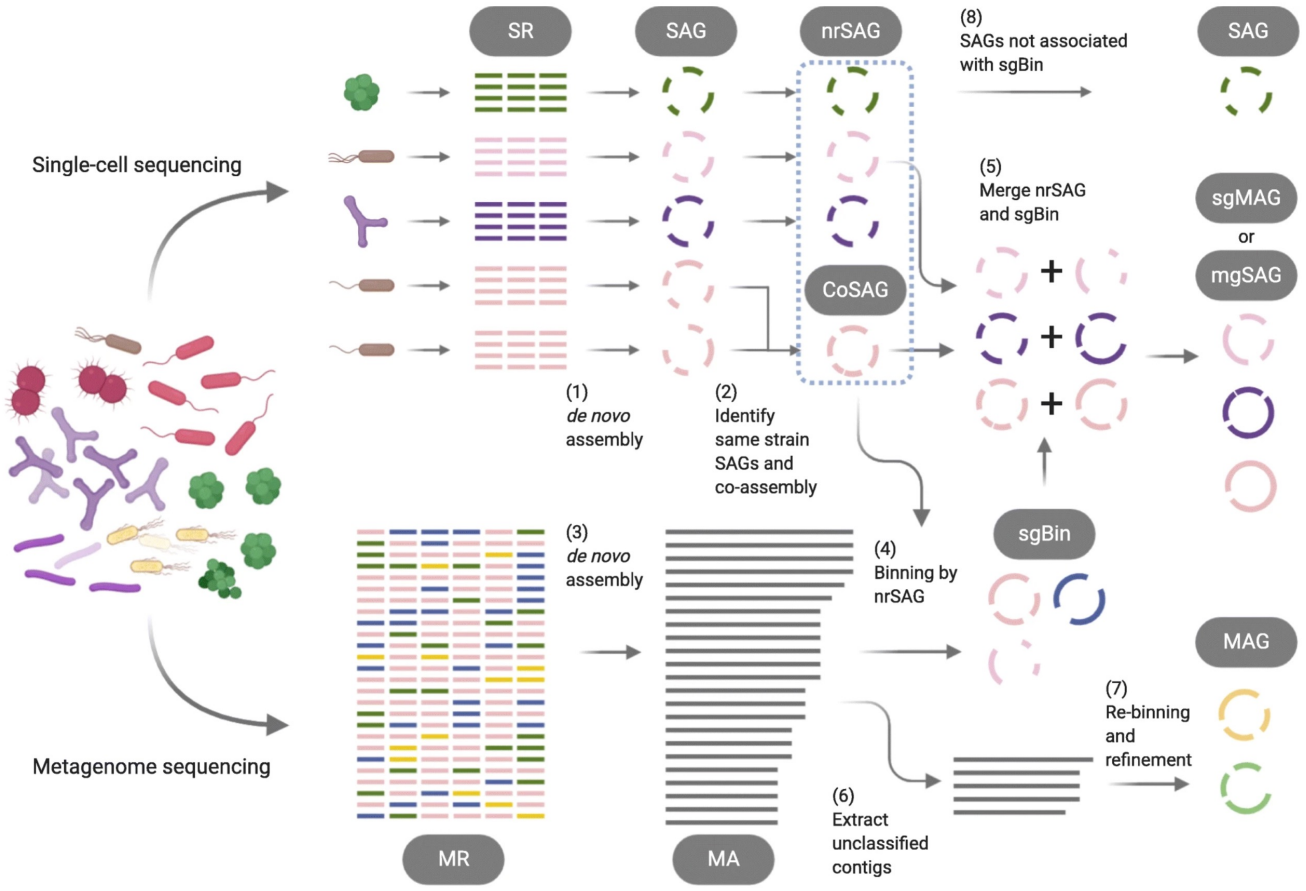
Zhang Y, et al. 2021
Annu. Rev. Biomed. Data Sci. 4:279–311

Other Approaches

Single-Cell genomics

- Study of individual microbial genomes, focuses on analyzing the genomes of individual microbial cells within a microbiome.
- **Goals**: Understand the genetic diversity, cellular heterogeneity, and functional capabilities of microbial communities at the single-cell level.
- **Insights:** Provides detailed information on the genetic composition and functional potential of individual microbes, allowing for the identification of rare or novel species, understanding of microbial interactions, and elucidation of community structure and dynamics.

From: Recovery of strain-resolved genomes from human microbiome through an integration framework of single-cell genomics and metagenomics



Arikawa, K. et al. Microbiome (2021)

References

- Shreiner, A. B., Kao, J. Y., & Young, V. B. (2015). The gut microbiome in health and in disease. *Current opinion in gastroenterology*, *31*(1), 69-75.
- Kinross, J. M., Darzi, A. W., & Nicholson, J. K. (2011). Gut microbiome-host interactions in health and disease. *Genome medicine*, *3*, 1-12.
- Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in plant science*, *5*, 86894.
- Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K. H., ... & Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, *12*(9), 635-645.
- Shahi, S. K., Freedman, S. N., & Mangalam, A. K. (2017). Gut microbiome in multiple sclerosis: The players involved and the roles they play. *Gut microbes*, *8*(6), 607-615.
- Simon, H. Y., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell*, *178*(4), 779-794.
- Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., ... & Segata, N. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*, *41*(11), 1633-1644.
- Schiebenhoefer, H., Schallert, K., Renard, B. Y., Trappe, K., Schmid, E., Benndorf, D., ... & Fuchs, S. (2020). A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and Prophane. *Nature protocols*, *15*(10), 3212-3239.
- Pang, Z., Chong, J., Li, S., & Xia, J. (2020). MetaboAnalystR 3.0: toward an optimized workflow for global metabolomics. *Metabolites*, *10*(5), 186.
- Arikawa, K., Ide, K., Kogawa, M., Saeki, T., Yoda, T., Endoh, T., ... & Hosokawa, M. (2021). Recovery of strain-resolved genomes from human microbiome through an integration framework of single-cell genomics and metagenomics. *Microbiome*, *9*, 1-16.
- https://www.researchgate.net/profile/Emilio_Laserna-Mendieta/publication
- https://igcbioinformatics.github.io/biomeshinycourse/pages/dada2/Biodata.ptCrashCourses.html
- https://docs.qiime2.org/2024.5/tutorials/overview

Happy Mining ATGC!