
3000788 Intro to Comp Molec Biol

Lecture 7: Metagenomics

September 6, 2022



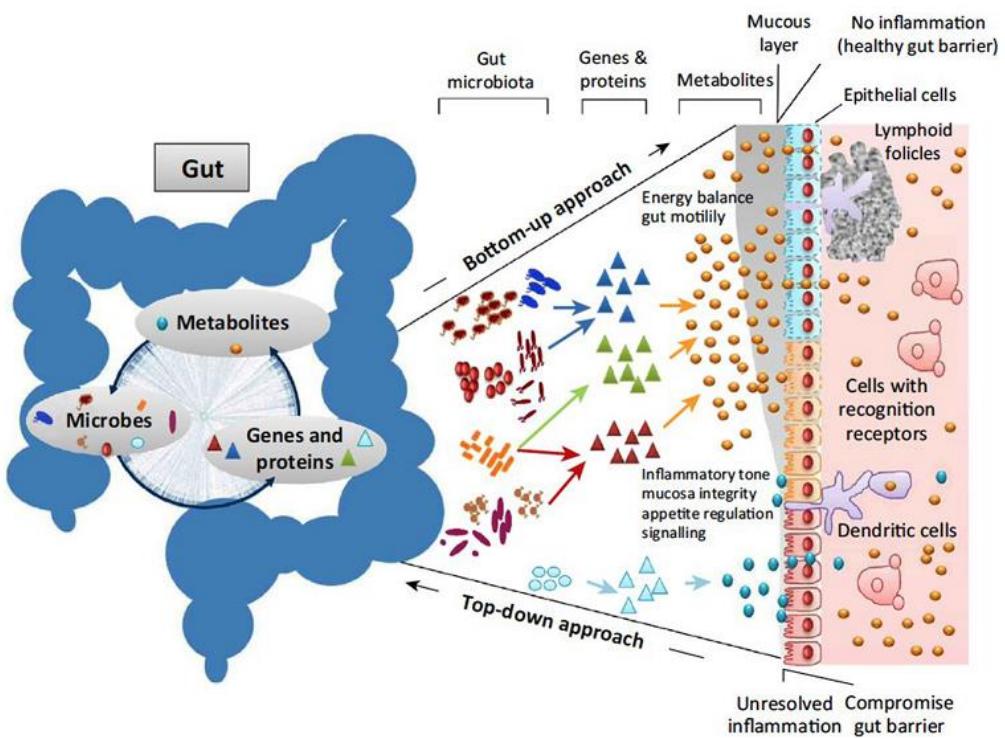
Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

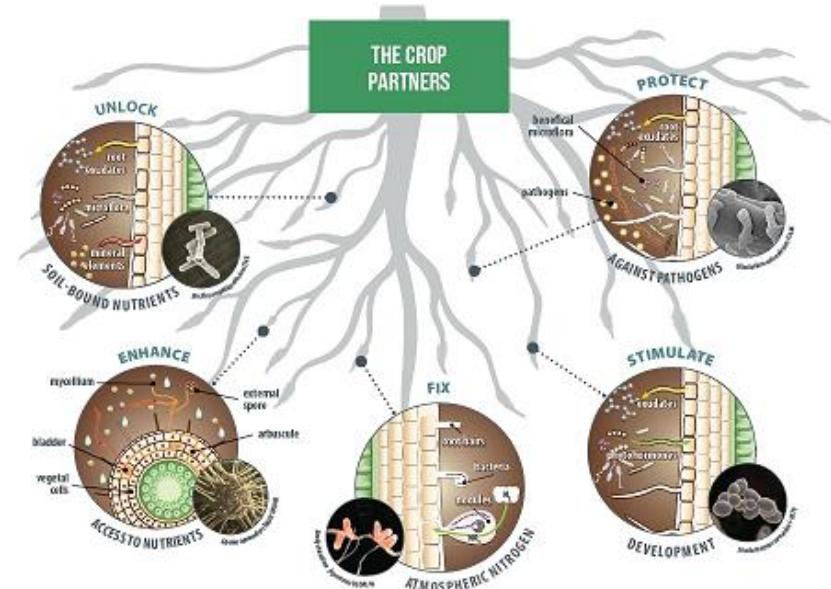


Microbiome and meta-omics

Microbiomes



Moya and Ferrer, Trends in Microbiology 24: 402-413 (2016)



<http://www.lallemandplantcare.com/en/our-solutions/rhizosphere-inoculants/>

Human microbiome

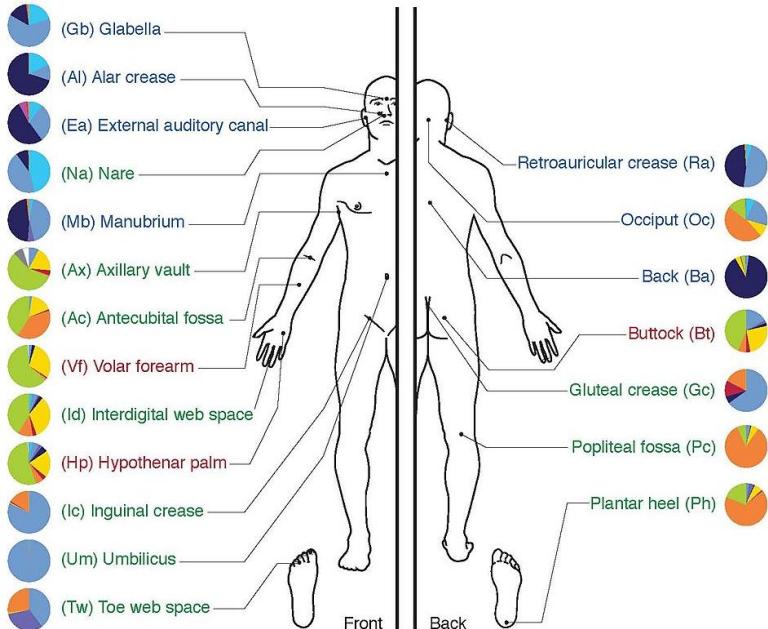
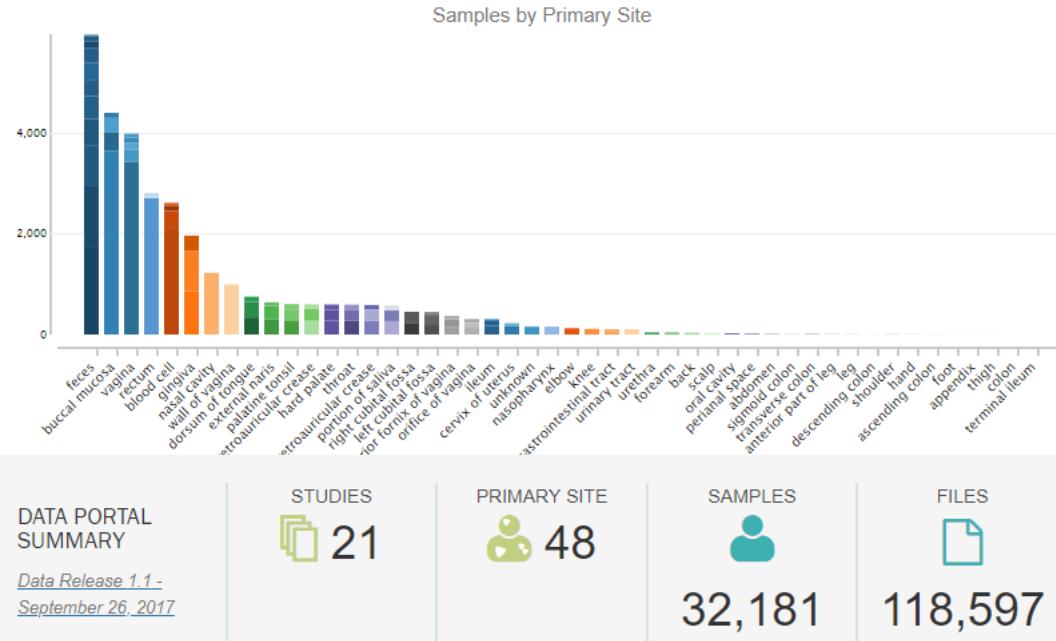
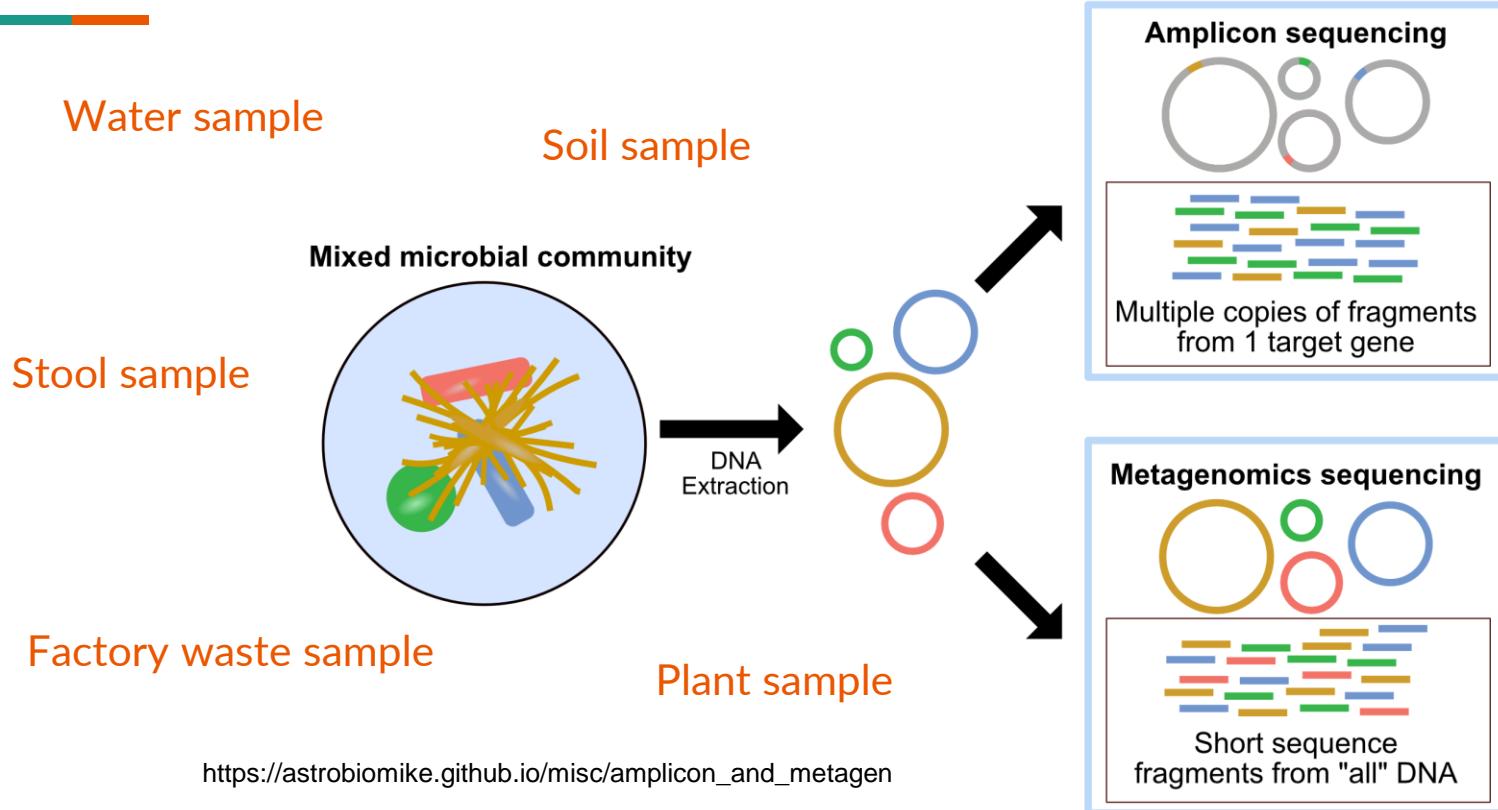


Image from Wikipedia.com



Omics analysis of a collection of organisms

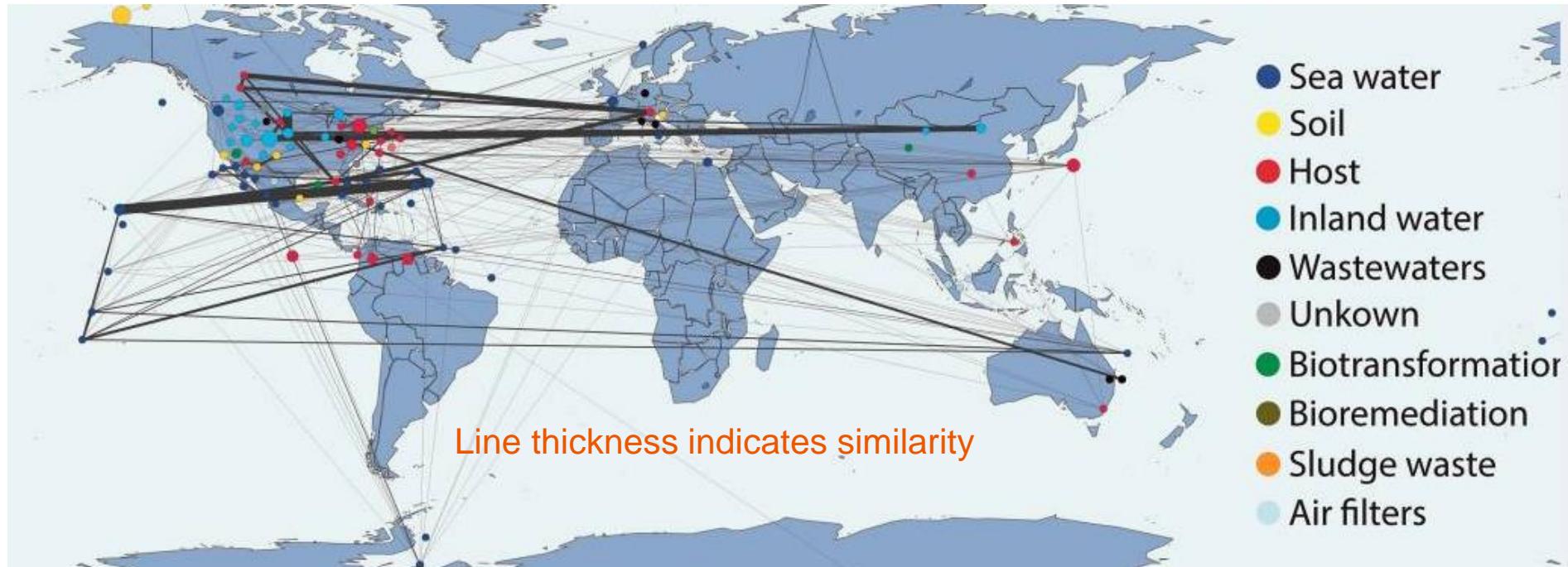


Why not other meta-omics?

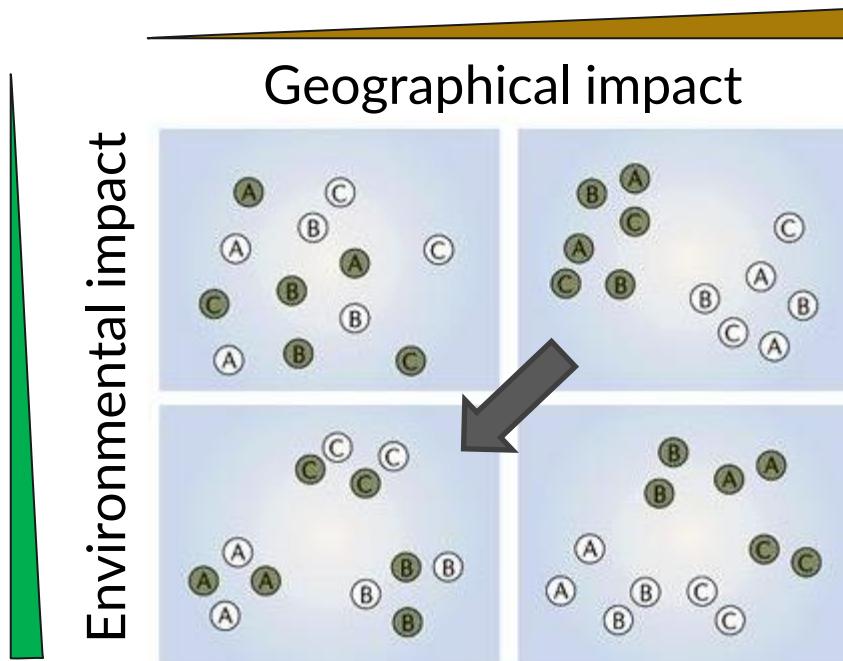


- Metatranscriptomics:
 - Difficult sample preparation
 - RNA are fragile
 - Challenging to determine whether a gene is ON or OFF in which subpopulations
- Metaproteomics:
 - Require reference protein database to interpret mass spectrometry data

Everything is everywhere, but the nature selects



Environment, not geography, defines microbiome

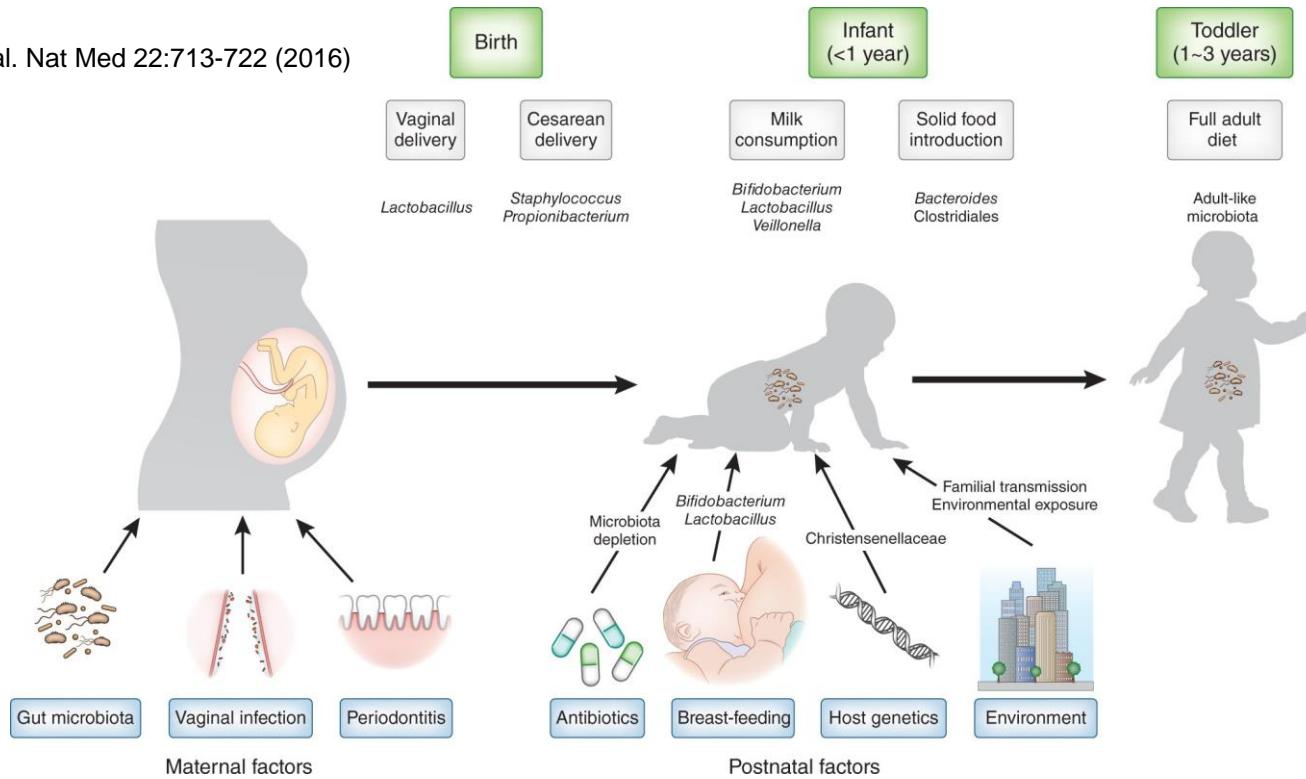


● ○ indicate same geographical locations

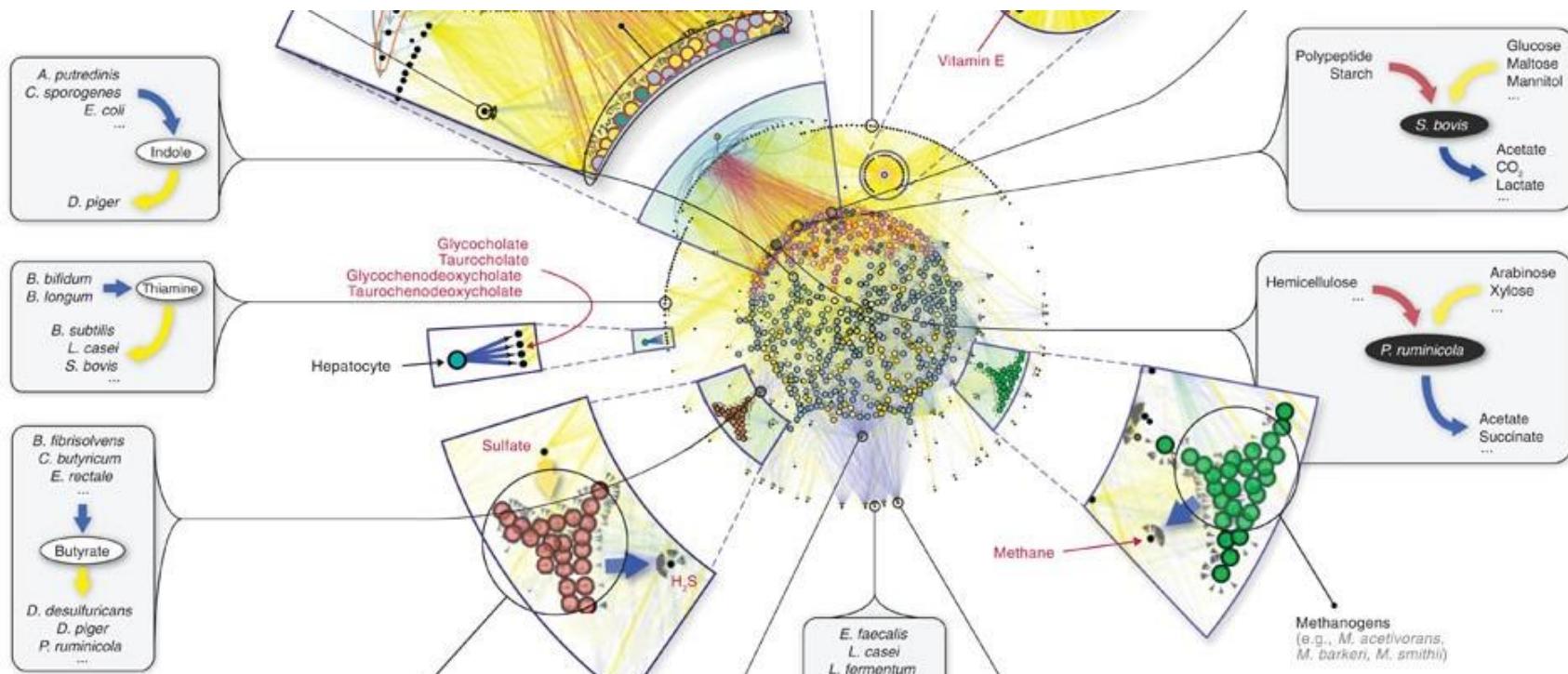
A, B, C indicate same environmental conditions

Microbiome is dynamics

Tamburini, S. et al. Nat Med 22:713-722 (2016)



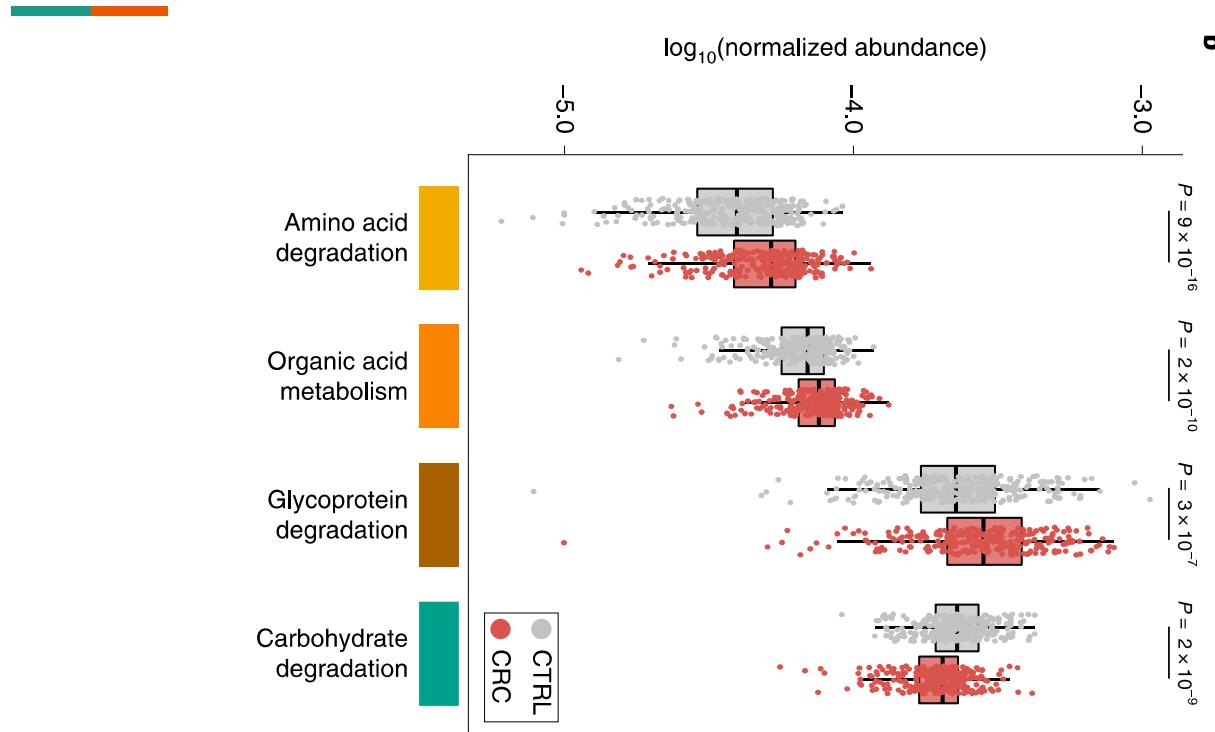
Microbiome is a network of cooperation and competition



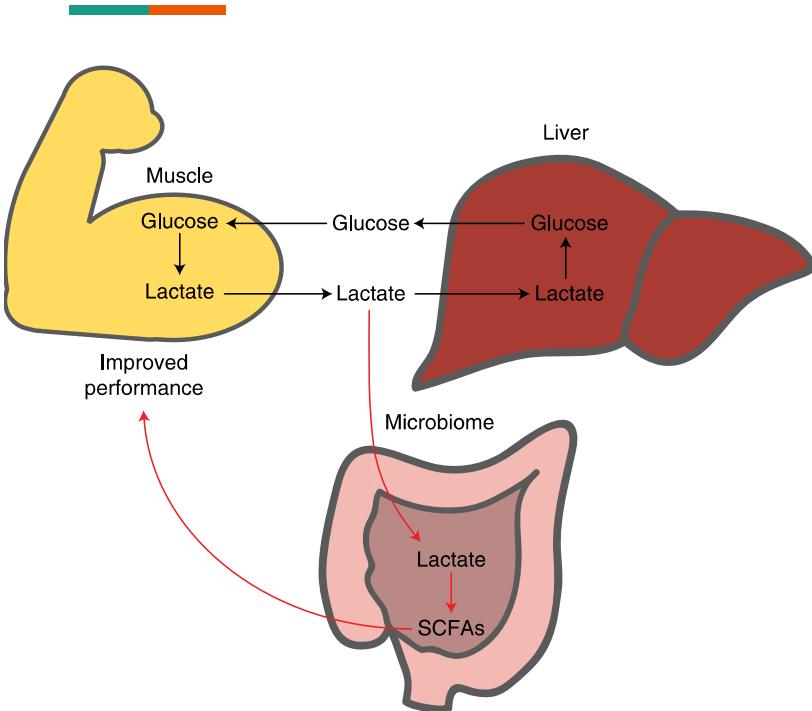


Applications of meta-omics

Microbiome link to health

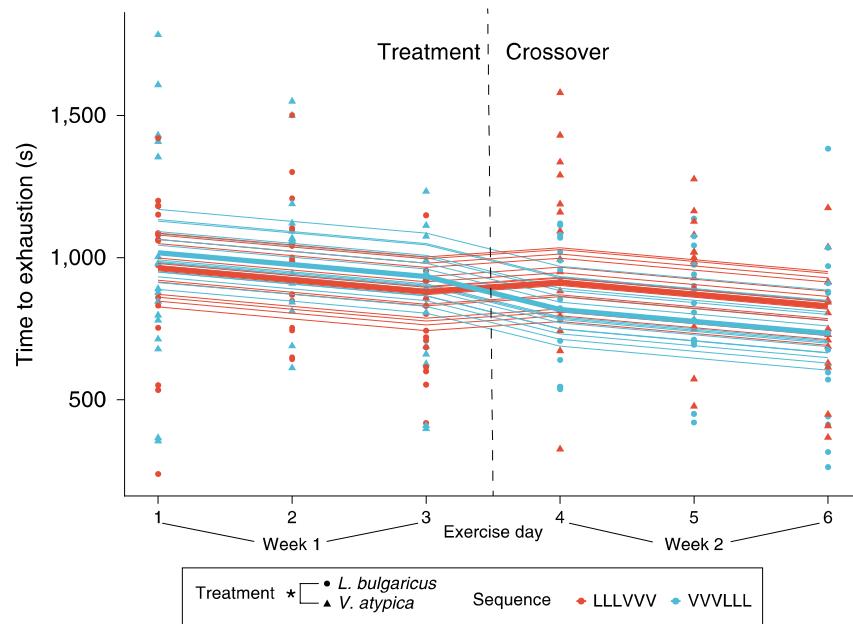


Lactate-utilizing bacteria in athlete guts



Scheiman et al. Nat Med 25:1104-1109 (2019)

Increased athletic ability in mice with transplanted microbiome



Wildlife conservation



Different
dietary plants
(nutrition,
isotopic
evidences)

Père David's deer and their gut microbiome

Dissimilarity

Gut microbial
composition

*Next-generation sequencing
bioinformatics' analysis*

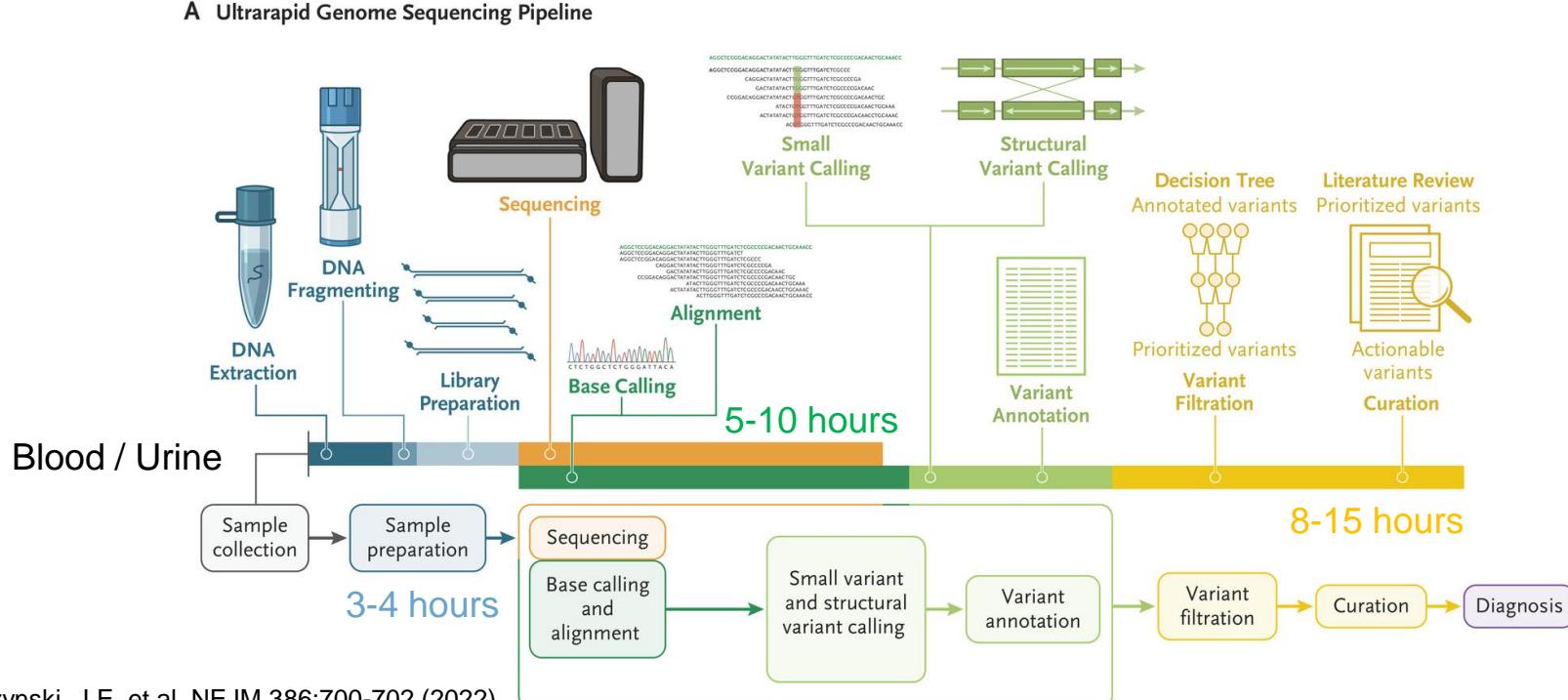
Gut microbial function
(cellulose digestion,
salt-related
metabolism)

Conservation

Reflecting increased
evolutionary potential
and resilience in response
to environment changes

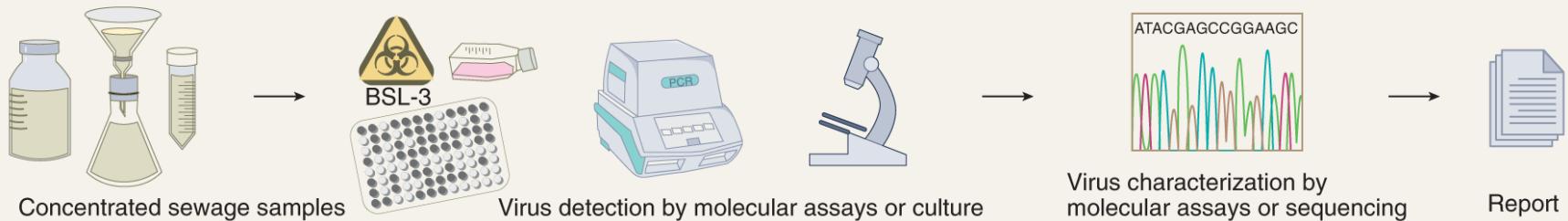
Helping us select a
putative translocation
region

Rapid pathogen detection for clinical decision



Pathogen surveillance

Sewage surveillance (for example, SARS-CoV-2)



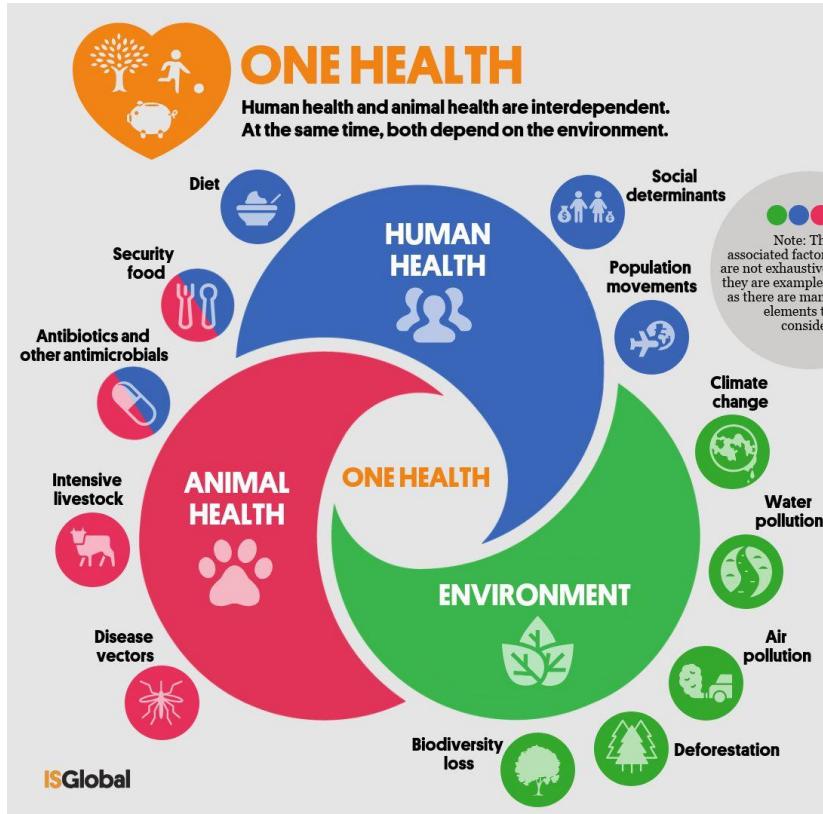
Ko, K.K.K. et al. Nat Micro 7:486-496 (2022)

- Direct strain identification
- Monitor evolution of active epidemic
- Detection of drug resistance gene
 - Resistome metagenomics: targeted sampling / amplicon sequencing

Research questions in meta-omics

- Health
 - Host-pathogen association
 - Drug resistance genes
 - Gut microbiome, cancer microbiome
- Ecology
 - Change in microbiome due to human actions
 - Factory and hospital wastes
 - Global warming
 - Microbiome of extreme conditions
- Agriculture = pathogens and yield
- Surveillance

Metagenomics for One Health



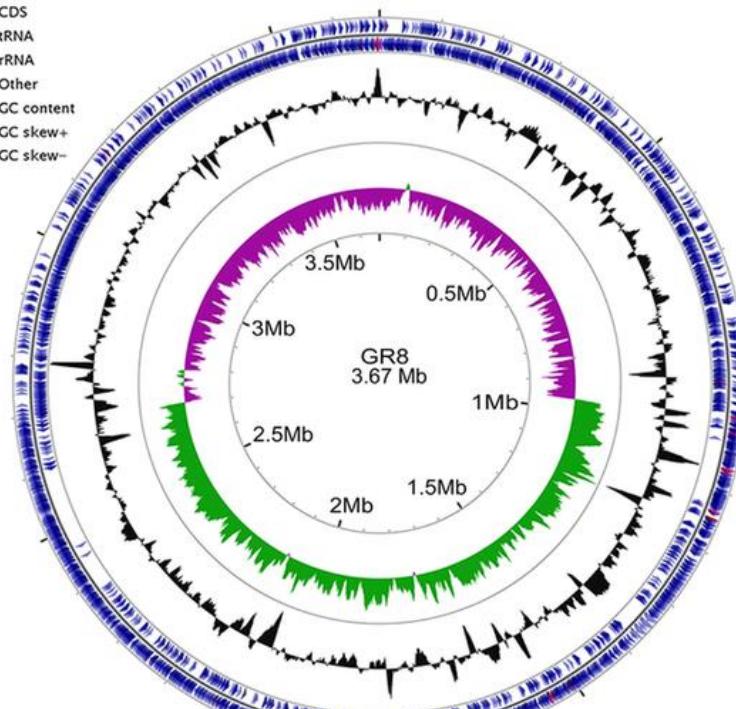


Prokaryotic genomes

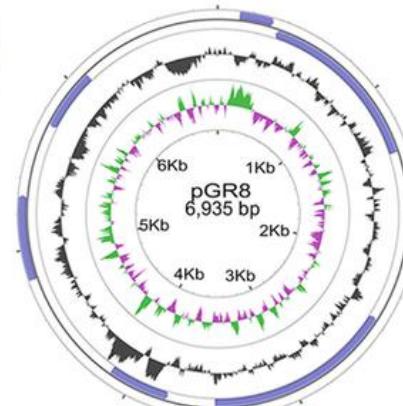
Circular chromosomes and plasmids



- [Blue square] CDS
- [Red square] tRNA
- [Purple square] rRNA
- [Grey square] Other
- [Black square] GC content
- [Green square] GC skew+
- [Purple square] GC skew-



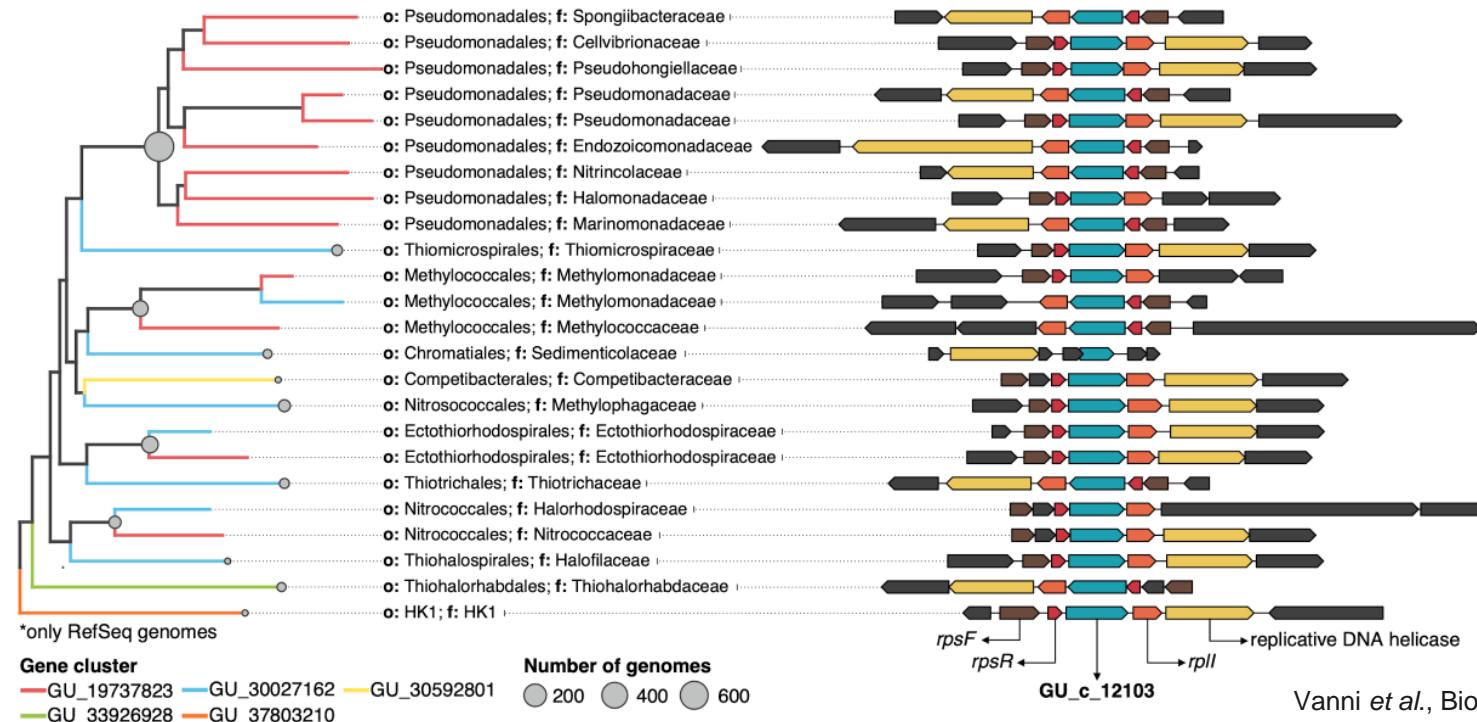
Yuan, Y. and Gao, M. Sci Rep 5:10259 (2015)



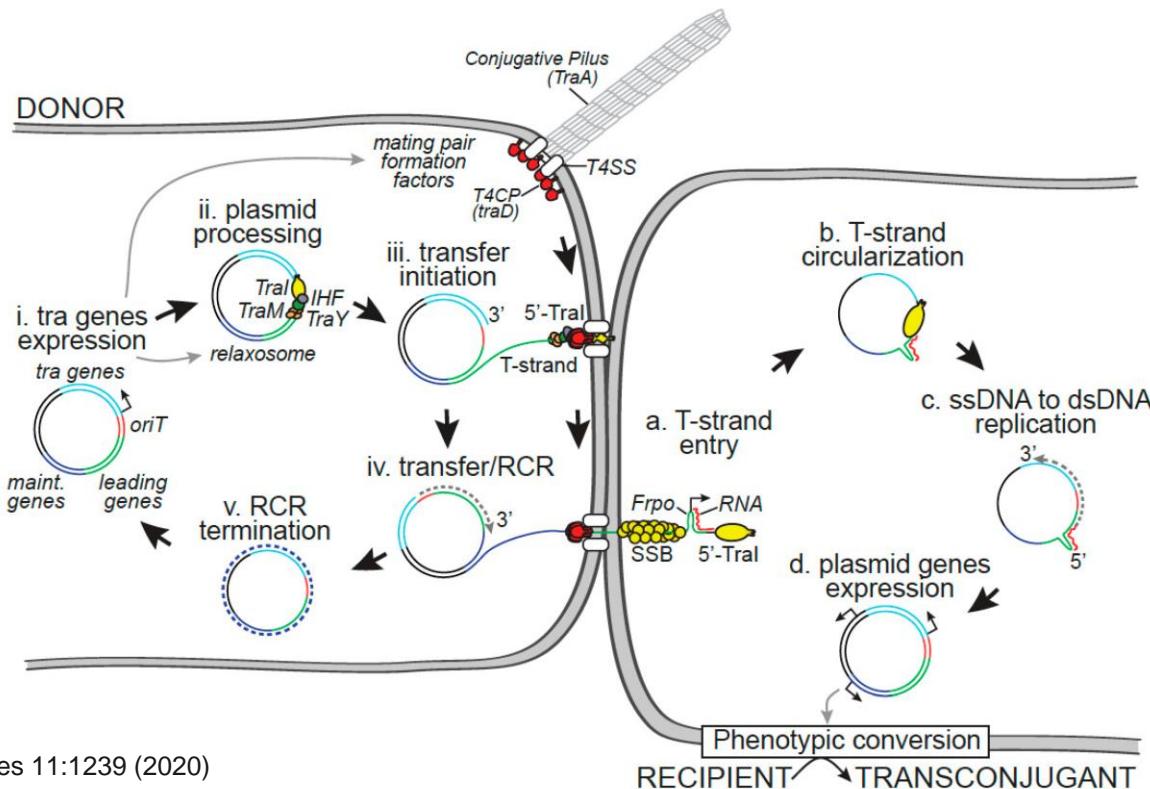
Gene operons

D

GTDB r86 | Bacteria; Proteobacteria; Gammaproteobacteria*



Plasmid transfer



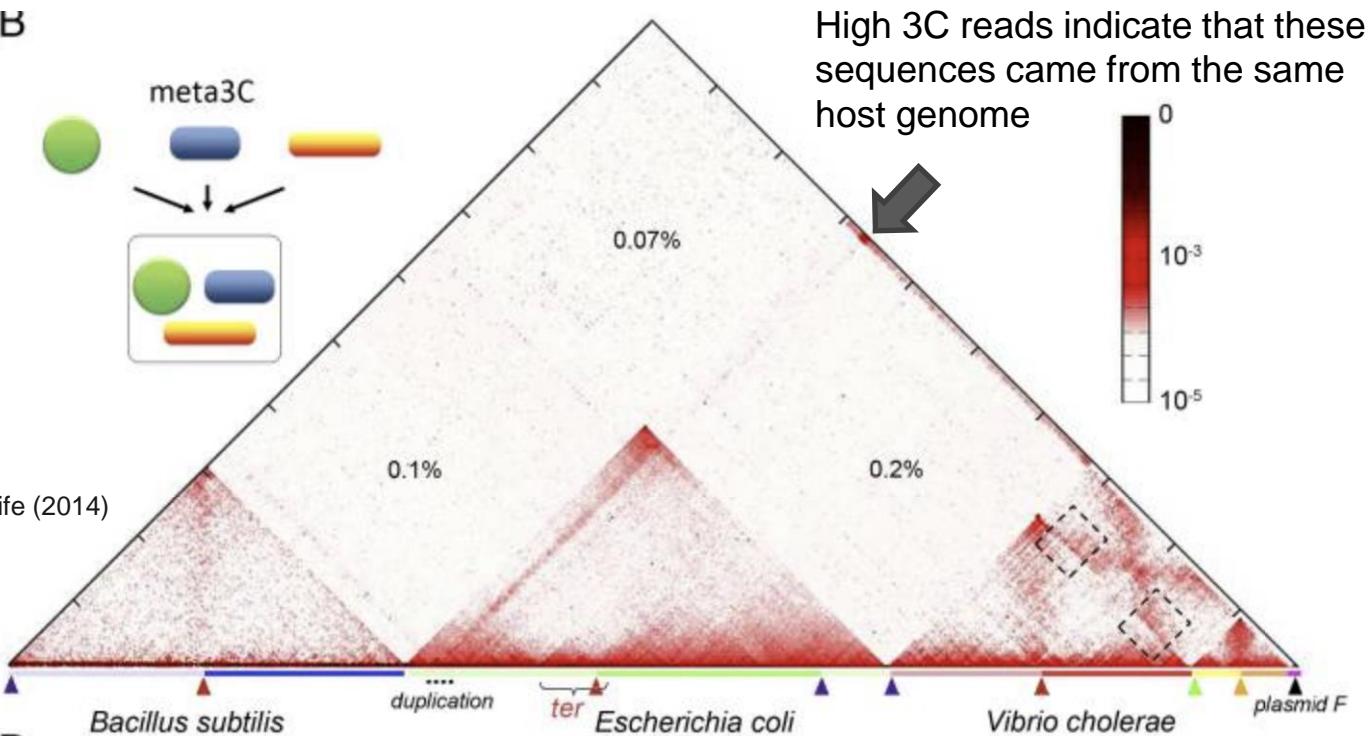
Challenges in metagenomics



- Grouping of DNA from the same organism is important
 - Interaction between genes in plasmid and chromosome
- Gene operon structure, not just sequence, is required for functional interpretation
 - Read assembly
- Small genomic differences across strains of the same species
 - Which similarity levels to call as SAME species, genus, etc. ?

Chromatin conformation capture for metagenomics

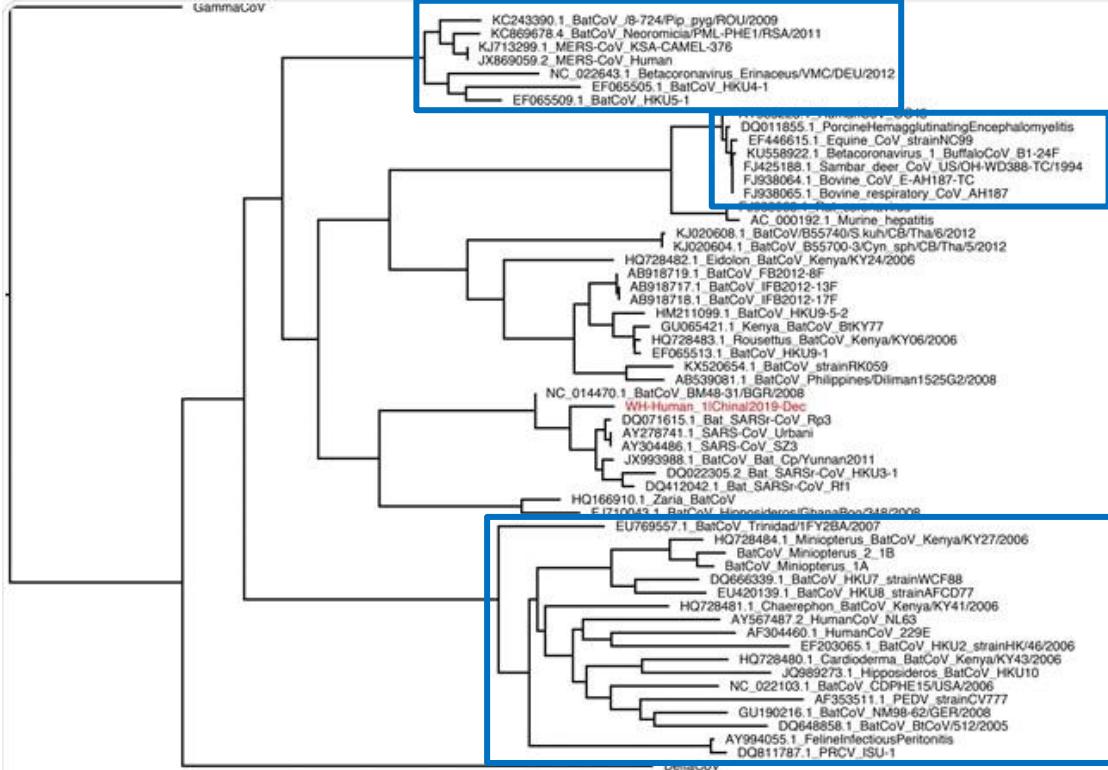
B





Operational texonomic unit (OTU)

Cluster of sequences with high similarity



Rough OTU similarity thresholds

nature reviews microbiology

Published: 14 August 2014

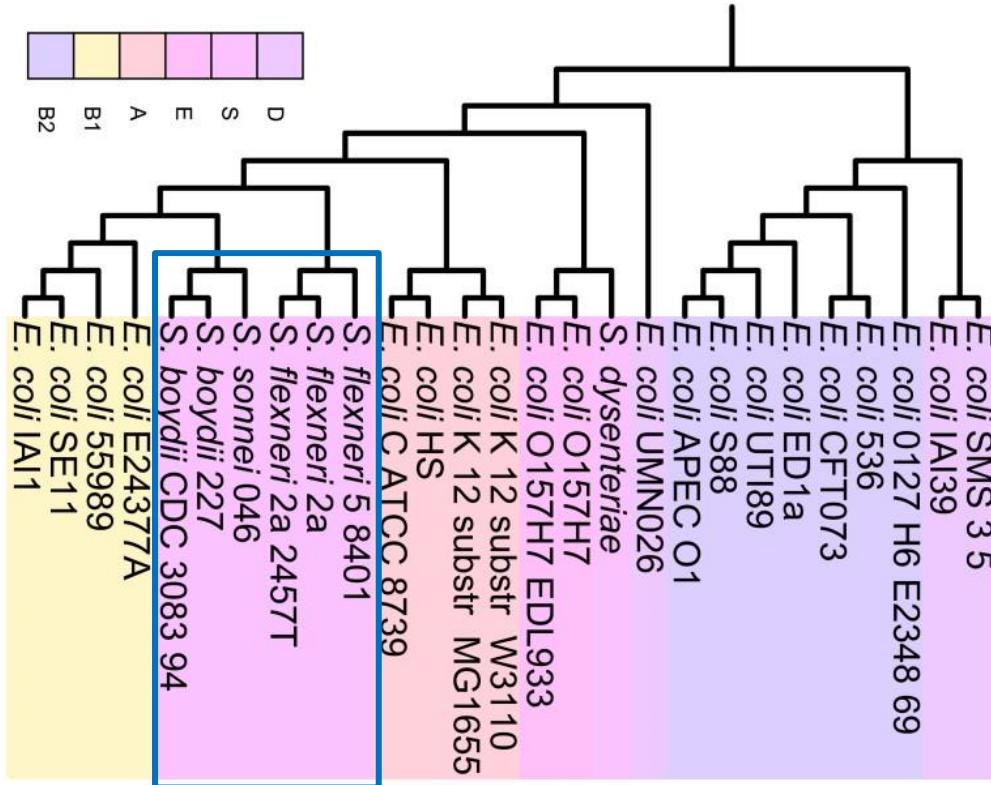
Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences

[Pablo Yarza](#) , [Pelin Yilmaz](#), [Elmar Pruesse](#), [Frank Oliver Glöckner](#), [Wolfgang Ludwig](#), [Karl-Heinz Schleifer](#),
[William B. Whitman](#), [Jean Euzéby](#), [Rudolf Amann](#) & [Ramon Rosselló-Móra](#) 

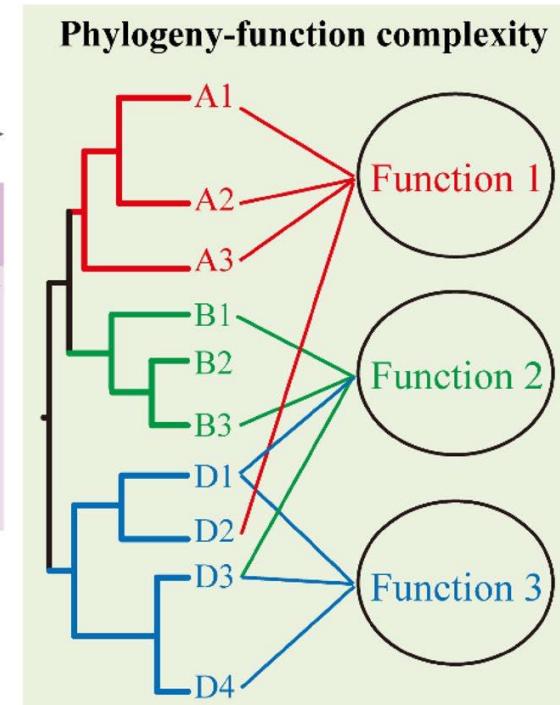
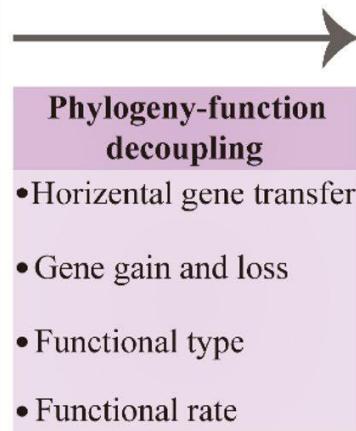
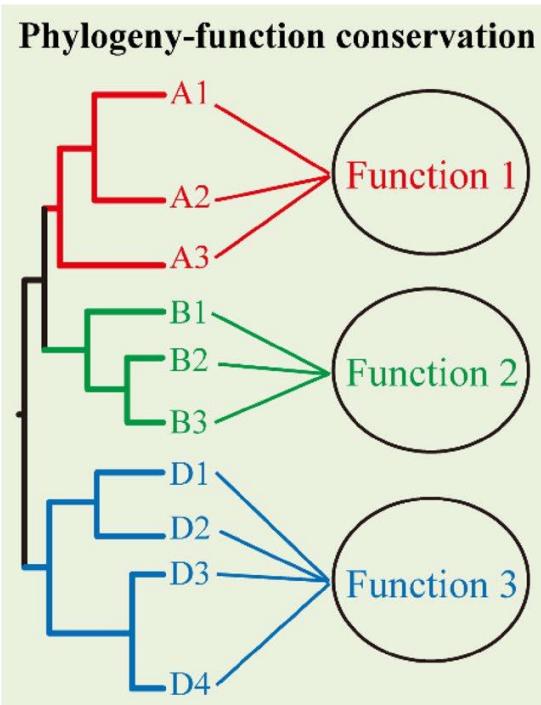
- Based on annotated 16S rRNA genes in database
- 94.5% similarity for genus, 86.5% for family, 75% for phylum
- Near-complete sequences are required to get accurate classification

An exception

- Genus *Shigella* are pathogens that evolved from an *E. coli* ancestor
- 80-90% similarity to some *E. coli* clades
- Definition of taxonomy may require both genotypes and phenotypes



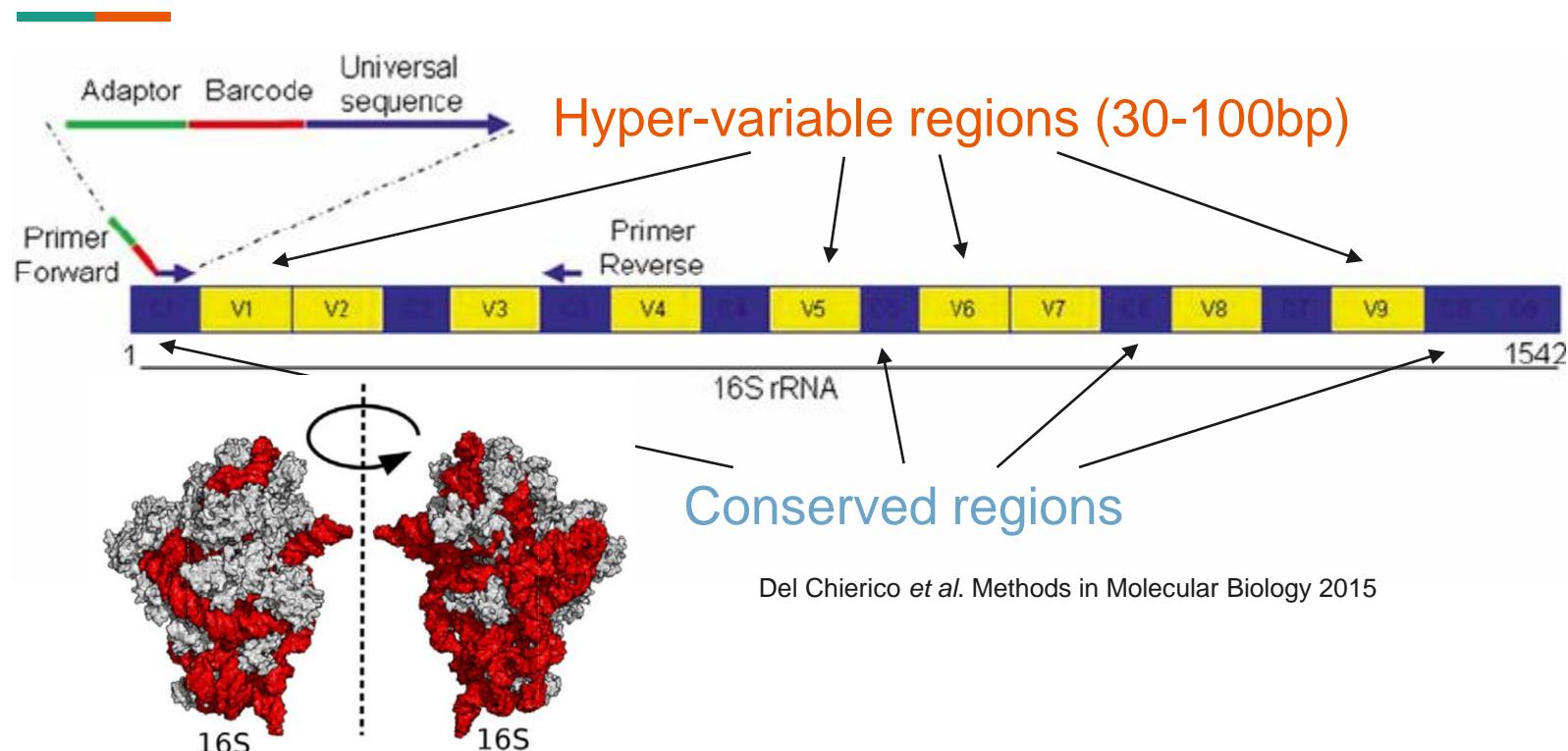
Functional view of taxonomy





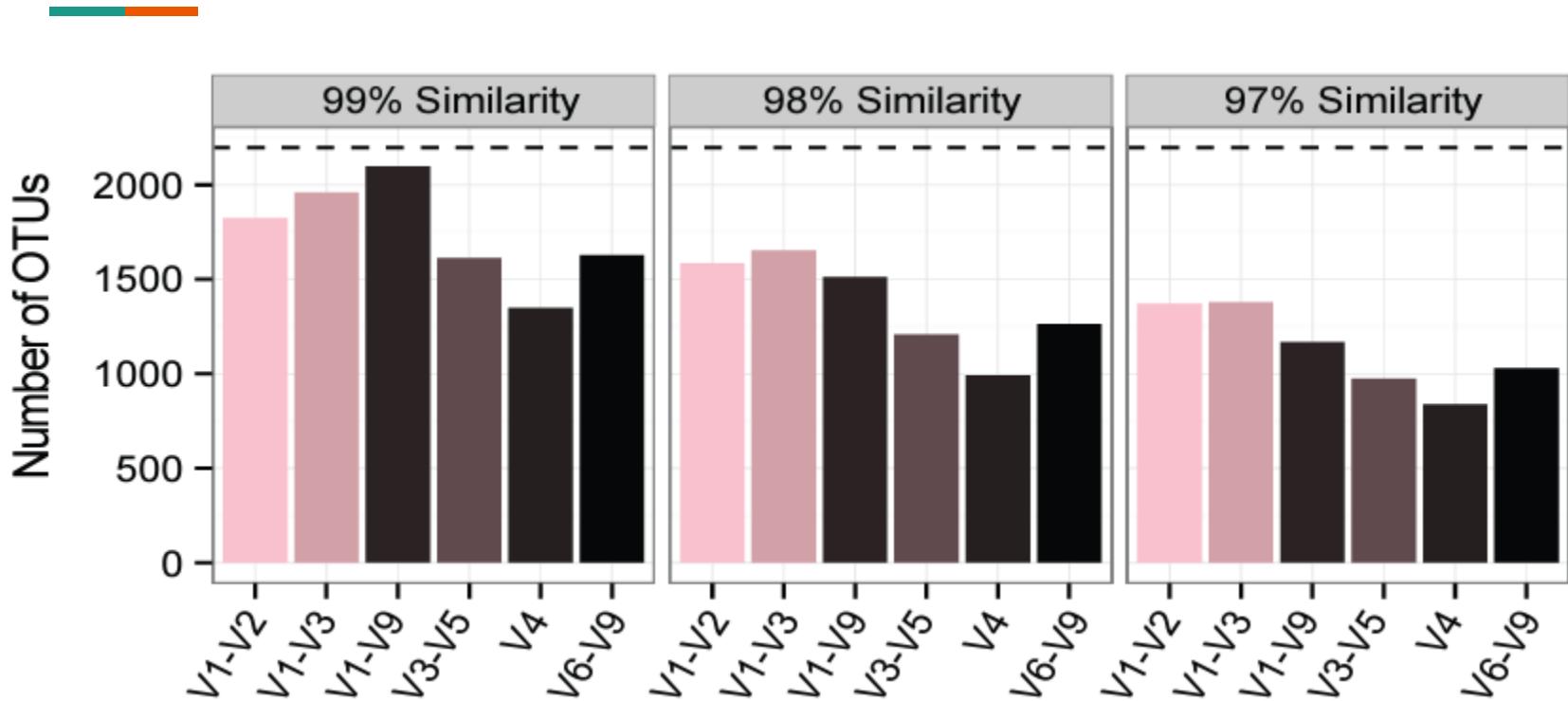
Taxonomy analysis via rRNA loci

16S rRNA in prokaryotes

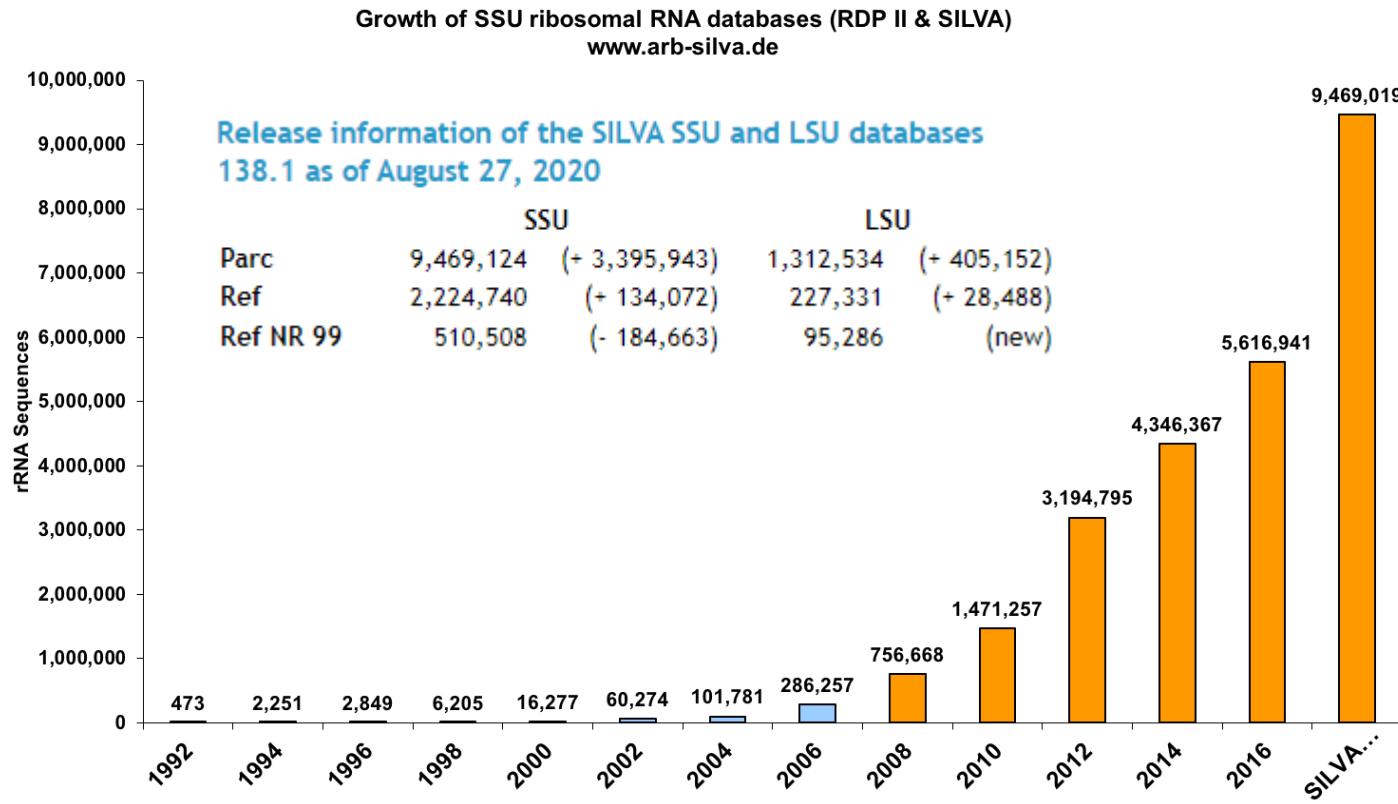


Del Chierico *et al.* Methods in Molecular Biology 2015

Impact of hypervariable region choices



Known rRNA sequences



rRNA databases



high quality ribosomal RNA databases

Home SILVangs Browser Search ACT Download Doc

SILVA

Welcome to the SILVA rRNA database project

A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.

SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (*Bacteria*, *Archaea* and *Eukarya*).

SILVA are the official databases of the software package ARB.

RDP News
12/12/2018 *RDP and Fungene Pipelines are back*

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

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

The Greengenes Database

Browse links below to download versions of the Greengenes 16S rRNA gene database or experimental datasets created with the PhyloChip 16S rRNA microarray. Beware that these publicly available versions of the Greengenes database utilize taxonomic terms proposed from phylogenetic methods applied years ago between 2012 and 2013. Since then, a variety of novel phylogenetic methods have been proposed for Archaea and Bacteria. For a recent example see Yokono 2018.

unite
community

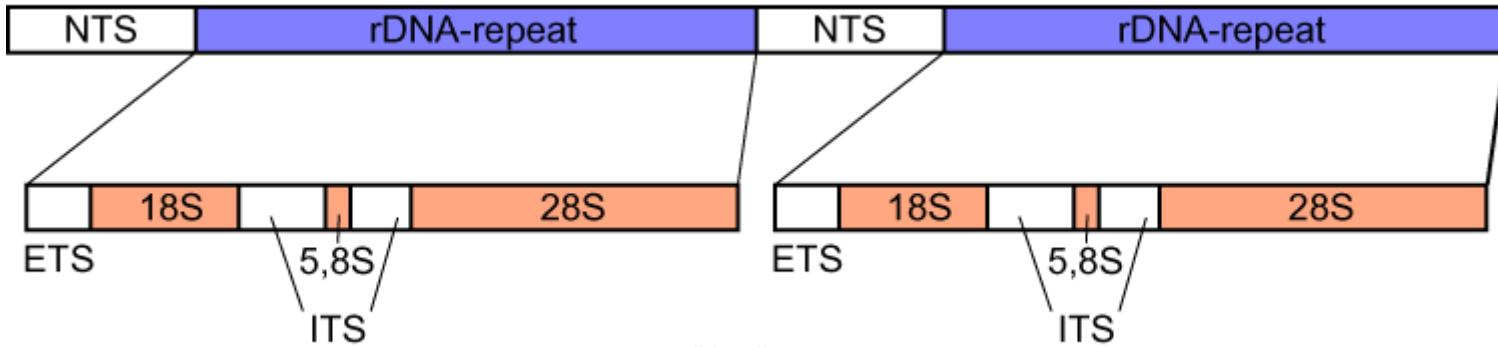
rDNA ITS based identifica



Current version: 8.2; Last updated: 2020-01-15 ([read more](#))

Number of ITS sequences (UNITE+INSD): 2 480 043; Number of UNITE fungal S

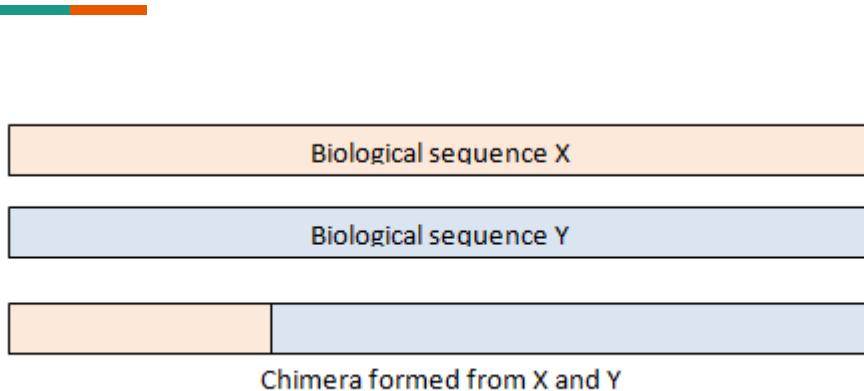
Internal transcribed spacer (ITS)



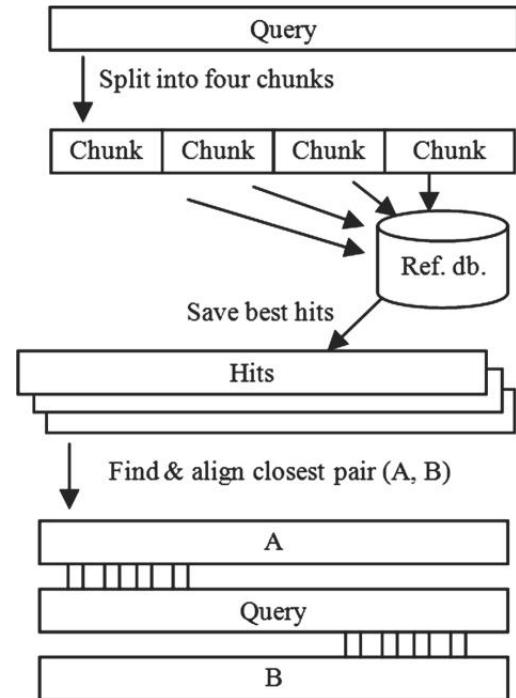
wikipedia.com

- Located between rRNA repeats
- ITS1 and ITS2
- 400-1000 bp
- Phylogenetics analysis of fungi and algae

Chimeric reads in amplicon sequencing



- Produced during PCR amplification
- Detected by alignment different portion of the reads to rRNA databases
- Mismatch of hits = chimeric reads



rRNA BLAST

The screenshot shows the 'Choose Search Set' interface for rRNA BLAST. It has three main sections: Database, Organism, and Exclude.

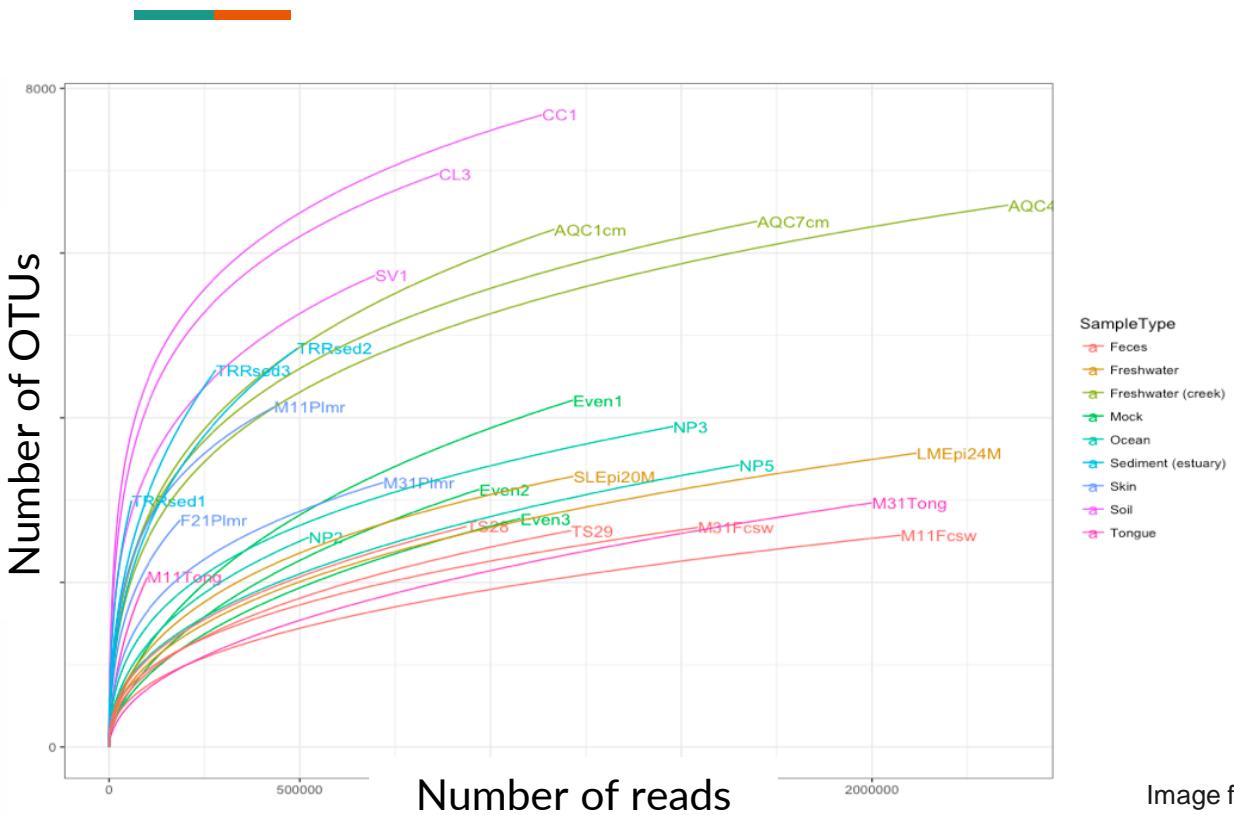
Database: Radio buttons for Standard databases (nr etc.), rRNA/ITS databases (selected), Genomic + transcript databases, and Betacoronavirus.

Organism: A dropdown menu showing '16S ribosomal RNA sequences (Bacteria and Archaea)' (selected) and other options: 18S ribosomal RNA sequences (SSU) from Fungi type and reference material, 28S ribosomal RNA sequences (LSU) from Fungi type and reference material, and Internal transcribed spacer region (ITS) from Fungi type and reference material. There is also a 'Targeted Loci Project Information' link and an 'Add organism' button.

Exclude: Checkboxes for Models (XM/XP) and Uncultured/environmental sample sequences.

- Endpoint of rRNA amplicon analysis is taxonomic assignment
- Abundance profiles of taxa can be correlated to environment condition or disease status

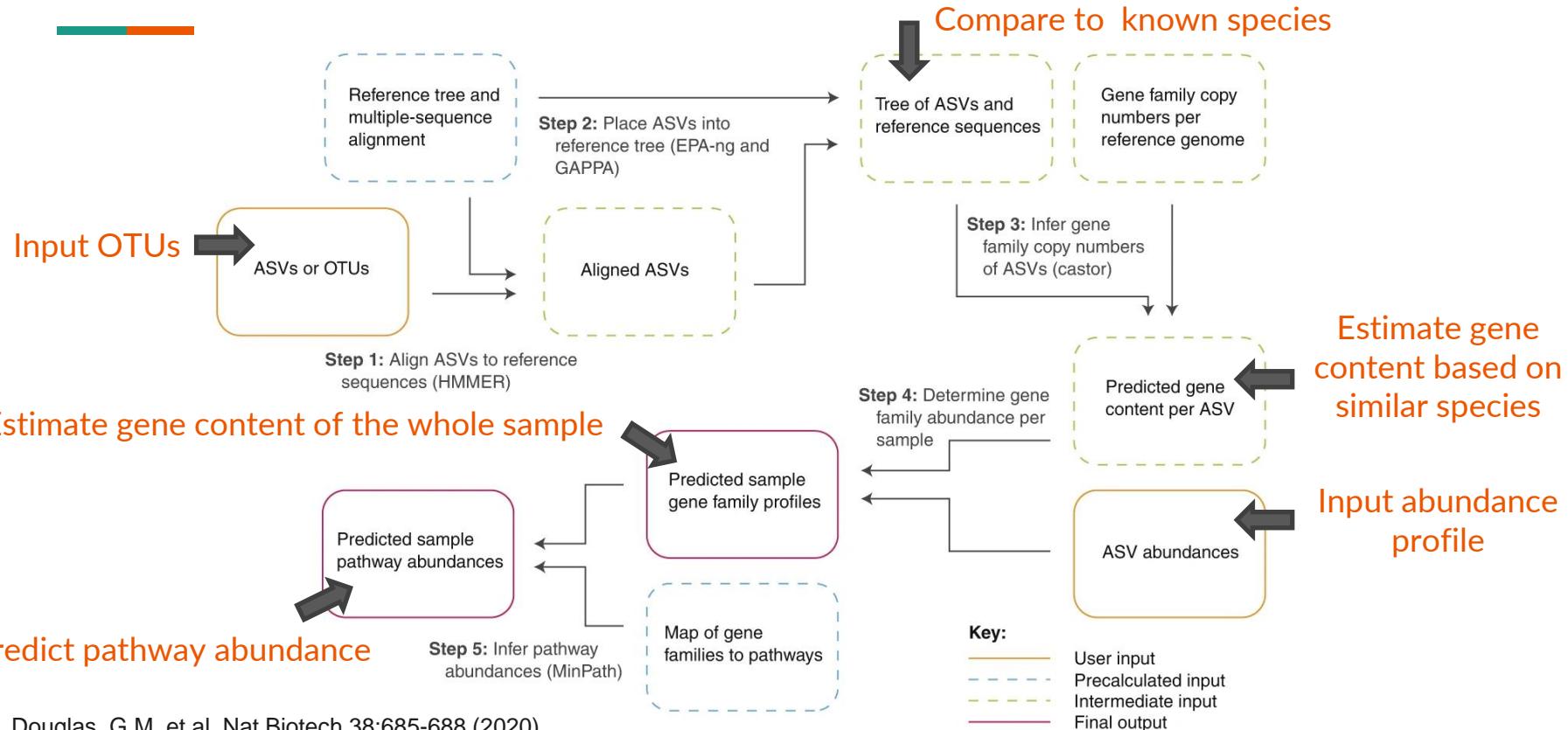
Rarefaction curve for evaluating depth of sequencing



- Subsample some reads and count the number of OTUs
- If depth is enough, not many additional OTUs will be found as the number of reads increases

Image from PhyloSeq GitHub

From taxonomy to functional pathway



Complexity of microbiome composition

Diversity indices/ Parameters	Description	Formula
Shannon diversity index (H)	Estimator of species richness and species evenness: more weight on species richness	$H = -\sum_{i=1}^s (p_i \ln p_i)$ <p>where s is the number of OTUs and p_i is the proportion of the community represented by OTU i.</p>
Simpson's index (D)	Estimator of species richness and species evenness: more weight on species evenness	$D = \frac{1}{\sum_{i=1}^s p_i^2}$ <p>where s is the total number of species in the community and p_i is the proportion of community represented by OTU i.</p>

Kim, B.-R. et al. J Microbiol Biotechnol 27:2089-2093 (2017)

- Richness = number of distinct species
- Evenness = no dominant species

Behavior of Shannon entropy

Diversity indices/ Parameters	Description	Formula
Shannon diversity index (H)	Estimator of species richness and species evenness: more weight on species richness	$H = -\sum_{i=1}^s (p_i \ln p_i)$
Simpson's index (D)	Estimator of species richness and species evenness: more weight on species evenness	$D = \frac{1}{\sum_{i=1}^s p_i^2}$ <p>where s is the total number of species in the community and p_i is the proportion of community represented by OTU i.</p>

Kim, B.-R. et al. J Microbiol Biotechnol 27:2089-2093 (2017)

- Entropy is maximized when $p_i = \frac{1}{s} \rightarrow H_{max} = \ln(s)$
- Entropy is minimized when there is one dominant species $H_{min} = 0$

Behavior of Simpson's index

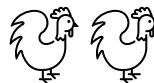
Diversity indices/ Parameters	Description	Formula
Shannon diversity index (H)	Estimator of species richness and species evenness: more weight on species richness	$H = -\sum_{i=1}^s (p_i \ln p_i)$ <p>where s is the number of OTUs and p_i is the proportion of the community represented by OTU i.</p>
Simpson's index (D)	Estimator of species richness and species evenness: more weight on species evenness	$D = \frac{1}{\sum_{i=1}^s p_i^2}$ <p>where s is the total number of species in the community and p_i is the proportion of community represented by OTU i.</p>

Kim, B.-R. et al. J Microbiol Biotechnol 27:2089-2093 (2017)

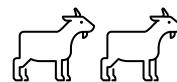
- D is maximized when $p_i = \frac{1}{s} \rightarrow D_{max} = s$
- D is minimized when there is only dominant species $D_{min} = 1$

Comparing microbiome composition

Sample 1



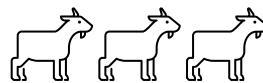
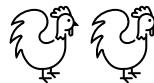
Sample 2



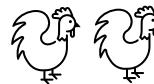
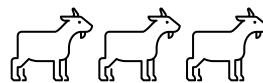
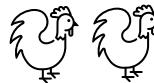
- $S_1 = \# \text{ of individuals in sample 1} = 6$
- $S_2 = \# \text{ of individuals in sample 2} = 5$
- Overlap = $1 + 2 + 1 = 4$
- Bray-Curtis dissimilarity = $1 - \frac{2 \times \text{Overlap}}{S_1 + S_2} = 1 - \frac{8}{11} = \frac{3}{11}$

Impact of sequencing depth

Sample 1

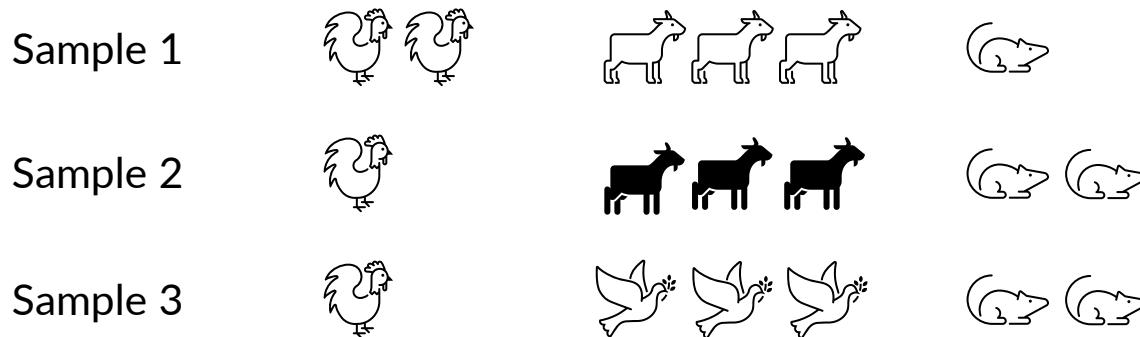


Sample 2



- Bray-Curtis is suitable between samples with similar sequencing depths

Impact of taxonomic similarity

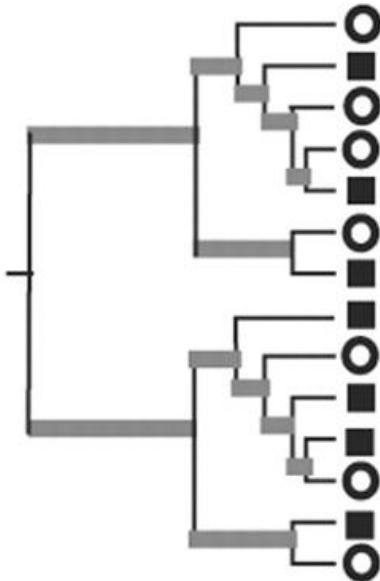


- Bray-Curtis does not take into account taxonomic similarity

UniFrac distance

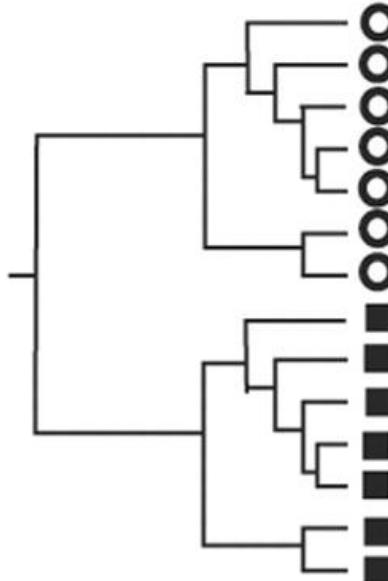
Lozupone, C. and Knight, R. Applied and Environmental Microbiology 71 (2005)

A.



○ Sample 1
■ Sample 2

B.

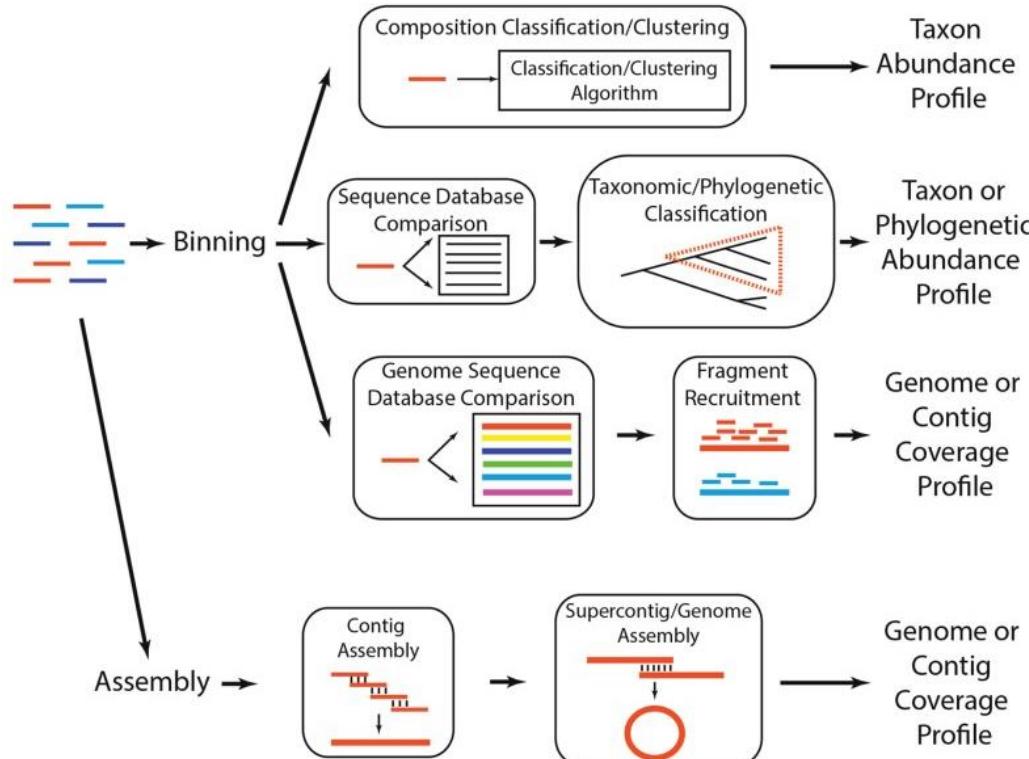


- UniFrac = fraction of shared phylogenetic branches between samples
- Can be weighted or unweighted by taxa abundances



Shotgun metagenomics

Key steps in shotgun metagenomics

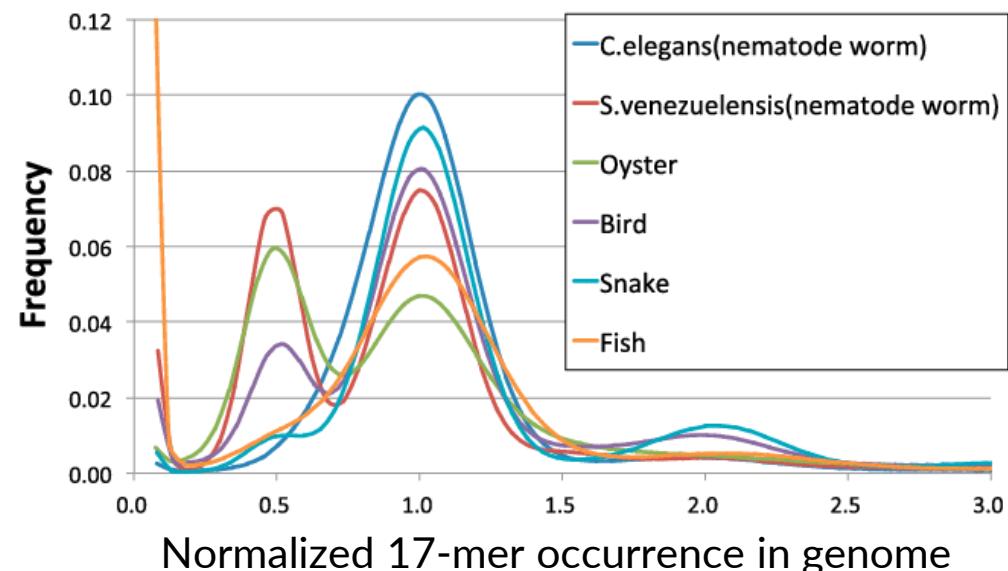


- Dealing with **contamination**
 - Host DNA
- **Binning** = grouping DNA/RNA from the same host organisms together
- Direct assembly is possible for abundant species

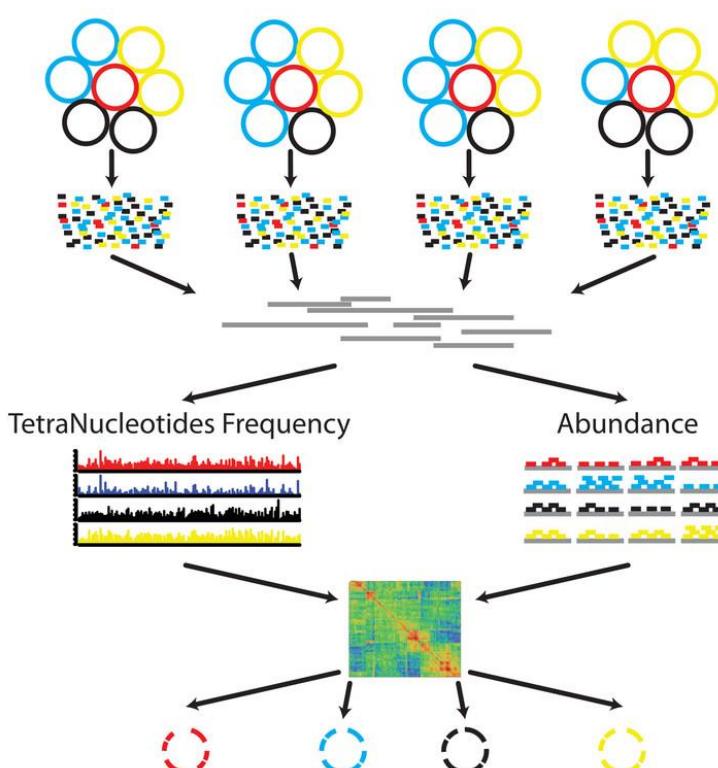
DNA k -mer can serve as species signature

k-mers for GTAGAGCTGT

<i>k</i>	<i>k</i> -mers
1	G, T, A, G, A, G, C, T, G, T
2	GT, TA, AG, GA, AG, GC, CT, TG, GT
3	GTA, TAG, AGA, GAG, AGC, GCT, CTG, TGT
4	GTAG, TAGA, AGAG, GAGC, AGCT, GCTG, CTGT
5	GTAGA, TAGAG, AGAGC, GAGCT, AGCTG, GCTGT
6	GTAGAG, TAGAGC, AGAGCT, GAGCTG, AGCTGT
7	GTAGAGC, TAGAGCT, AGAGCTG, GAGCTGT
8	GTAGAGCT, TAGAGCTG, AGAGCTGT
9	GTAGAGCTG, TAGAGCTGT
10	GTAGAGCTGT

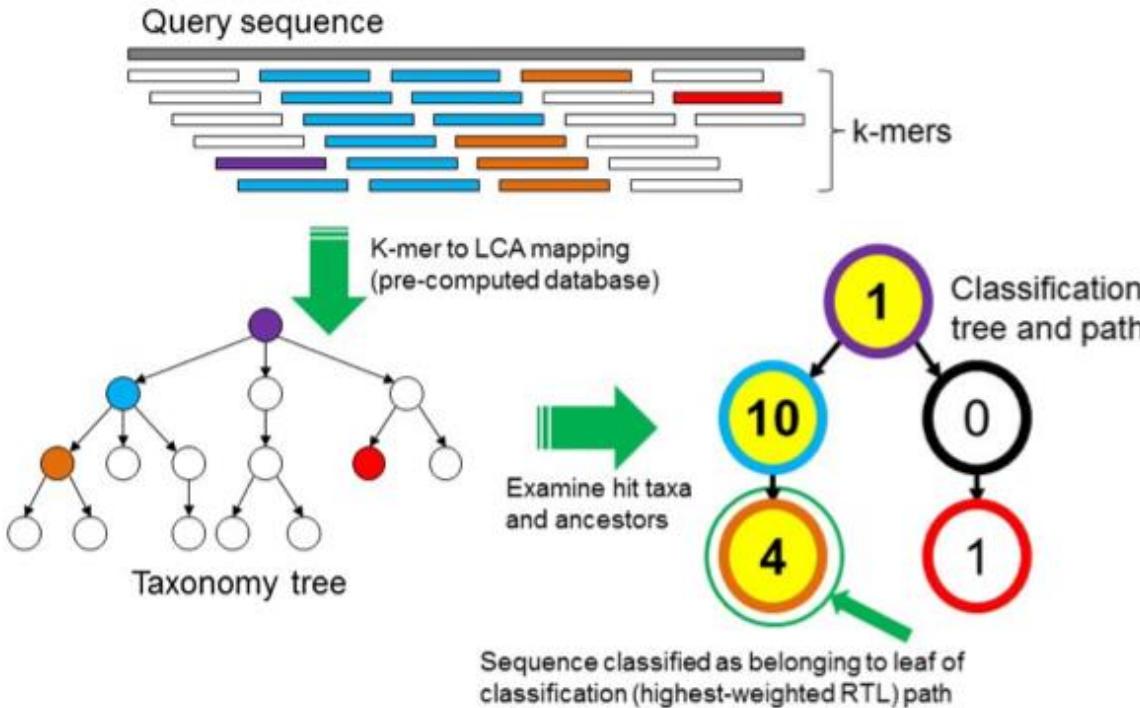


Read binning strategies



- Reads originated from the same species should have **similar k-mer profile**
- Pairs of reads originated from the same species should have **highly correlated abundances across samples** (because their abundances correlated with species abundance)

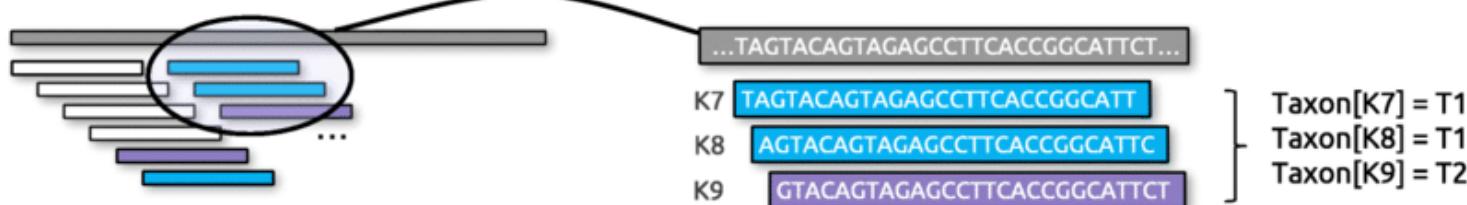
k-mer matching to predict taxonomy



- Much faster than alignment
- Select reads from taxonomy of interest
 - Virus
- Remove reads from contamination / host
 - Abundant bacteria
 - Human

k-mer matching to predict taxonomy

A Read *k*-mers are looked-up in the database and assigned to taxa:



C K-mer count and coverage in taxonomic report show evidence behind classifications:

reads	kmers	dup	cov	taxID	rank	name
122	112	144	0.0004	11855	species	<i>Clostridioides difficile</i>
9650	7129	74.5	0.192	10632	species	Human polyomavirus 2
15	1570	1	0.0002	7643	species group	<i>Mycobacterium tb</i> complex

Bad classification with few k-mers
Good classification, reads cover genome

Number of distinct k-mers for taxon, and coverage of the taxon's k-mers



Resources for metagenomics

Pre-built phylogenetics

cellular organisms

6 Show comments

Legend Zoom tree view + -

cellular organisms

Bacteria

- 1013-28-CG33
- 196up
- aaa34a10
- AEGEAN-245
- Aquamonas hayward
- aquifer1
- aquifer2
- ARKDMS-49
- ARKICE-90
- bacterium NTL235
- bacterium NTL237
- bacterium NTL338
- bacterium NTL344
- bacterium S13Fe21
- BD3-1
- BD72BR169
- BJGMM-U27
- Candidatus Benitsukem
- CK-1C2-67
- CK-1C4-19
- CS-B046
- EC3
- Elev-16S-509
- endophytic bacterium
- endosymbiont of Col
- endosymbiont of Col
- endosymbiont of Form

more... (Arenicellales - Sodalis-like symbiont of Nip)

more... (SPOTSOCT00m83 - Yokenella)

Open Tree of Life

Node properties

Source taxonomy
NCBI: 131567

Reference taxonomy
OTT: 93302

Node id in synthetic tree
ott93302

Descendant tips
2,391,916

Download subtree as Newick string
This tree is too large to download through the browser. See the [release notes](#) for download links

Search EOL for 'cellular organisms'

Browse cellular organisms in OneZoom

Add a comment

NCBI's GenBank genomics databases

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information

Overview (55471); Eukaryotes (12807); Prokaryotes (266319); Viruses (40989); Plasmids (23040); Organelles (16974) Filters Download

#	Organism Name	Organism Groups	Strain	BioSample
1	'Brassica napus' phytoplasma	Bacteria;Terrabacteria group;Tenericutes	TW1	SAMN09083457
2	'Candidatus Kapabacteria' thiocyanatum	Bacteria;FCB group;Bacteroidetes/Chloro group	59-99	SAMN05660602
3	'Catharanthus roseus' aster yellows phytoplasma	Bacteria;Terrabacteria group;Tenericutes	De Villa	SAMN10923938
4	'Chrysanthemum coronarium' phytoplasma	Bacteria;Terrabacteria group;Tenericutes	OY-V	SAMD00018609
5	'Cynodon dactylon' phytoplasma	Bacteria;Terrabacteria group;Tenericutes	LW01	SAMN12727363
6	'Echinacea purpurea' witches'-broom phytoplasma	Bacteria;Terrabacteria group;Tenericutes	NCHU2014	SAMN04017316
7	'Massilia aquatica' Holochova et al. 2020	Bacteria;Proteobacteria;Bet	CCM 8693	SAMN12721297
8	'Massilia aquatica' Lu et al. 2020	Bacteria;Proteobacteria;Bet	FT127W	SAMN13678849
9	'Nostoc azollae' 0708	Bacteria;Terrabacteria group;Cyanobacteria/Melain	0708	SAMN02598476

Table 5. Number of entries in commonly used reference databases

Domain	Level	Draft genomes		Complete genomes ¹	
		GenBank	RefSeq	GenBank	RefSeq
Archaea	Entries	859	351	260 (20)	225 (12)
	Species	695	204	209 (14)	178 (7)
Bacteria	Entries	89 730	78 783	7314 (1346)	6973 (1066)
	Species	19 078	11 217	2677 (542)	2586 (406)
Fungi	Entries	1897	191	28 (414)	7 (38)
	Species	997	190	17 (68)	7 (36)
Protists	Entries	430	47	2 (49)	2 (27)
	Species	226	47	2 (38)	2 (26)
Viruses	Entries	3	3	0 (0)	7214 (22)
	Species	1	3	0 (0)	7073 (22)

A comprehensive review of metagenomics tools

MICROBIAL GENOMICS

REVIEW

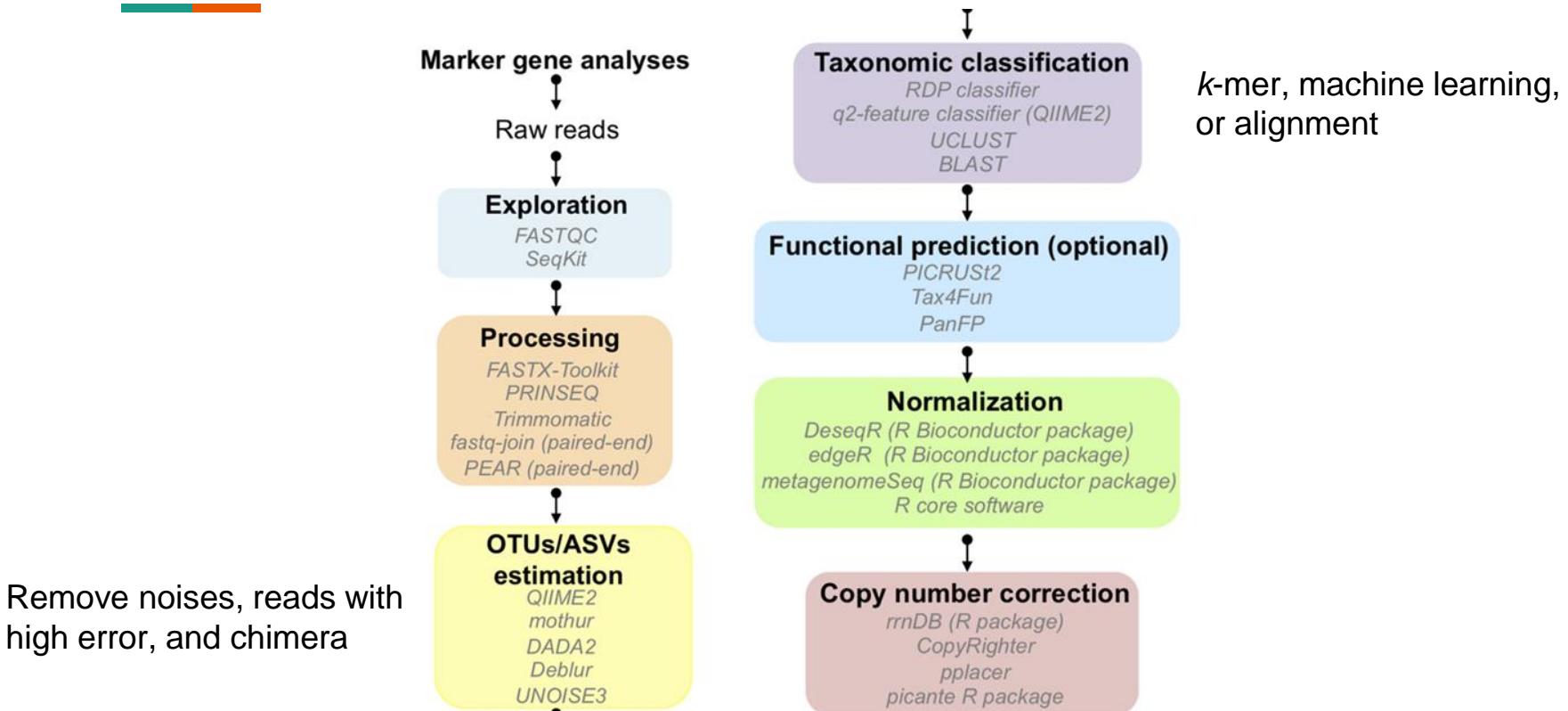
Pérez-Cobas *et al.*, *Microbial Genomics* 2020;6

DOI 10.1099/mgen.0.000409

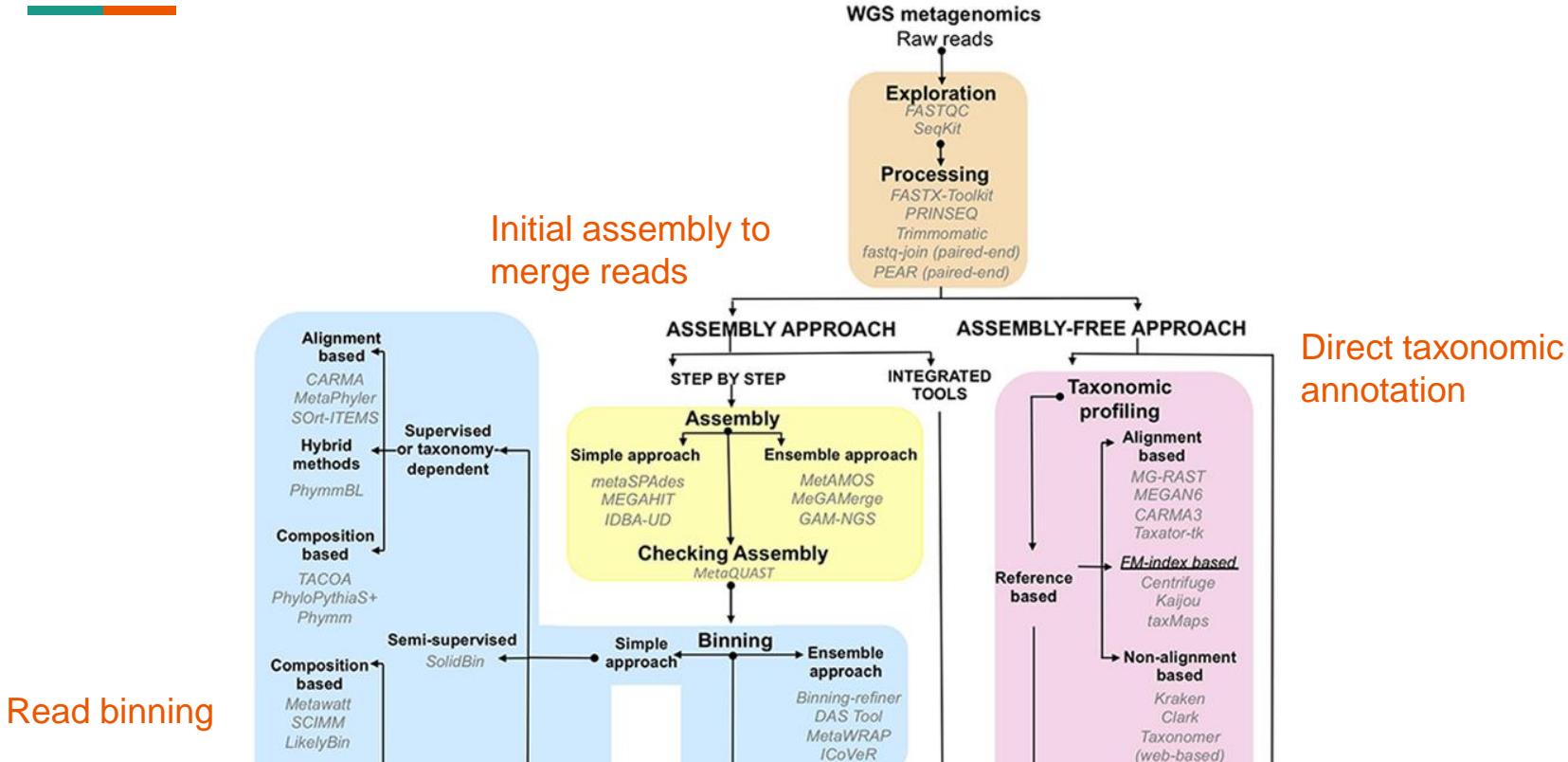
Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses

Ana Elena Pérez-Cobas, Laura Gomez-Valero and Carmen Buchrieser*

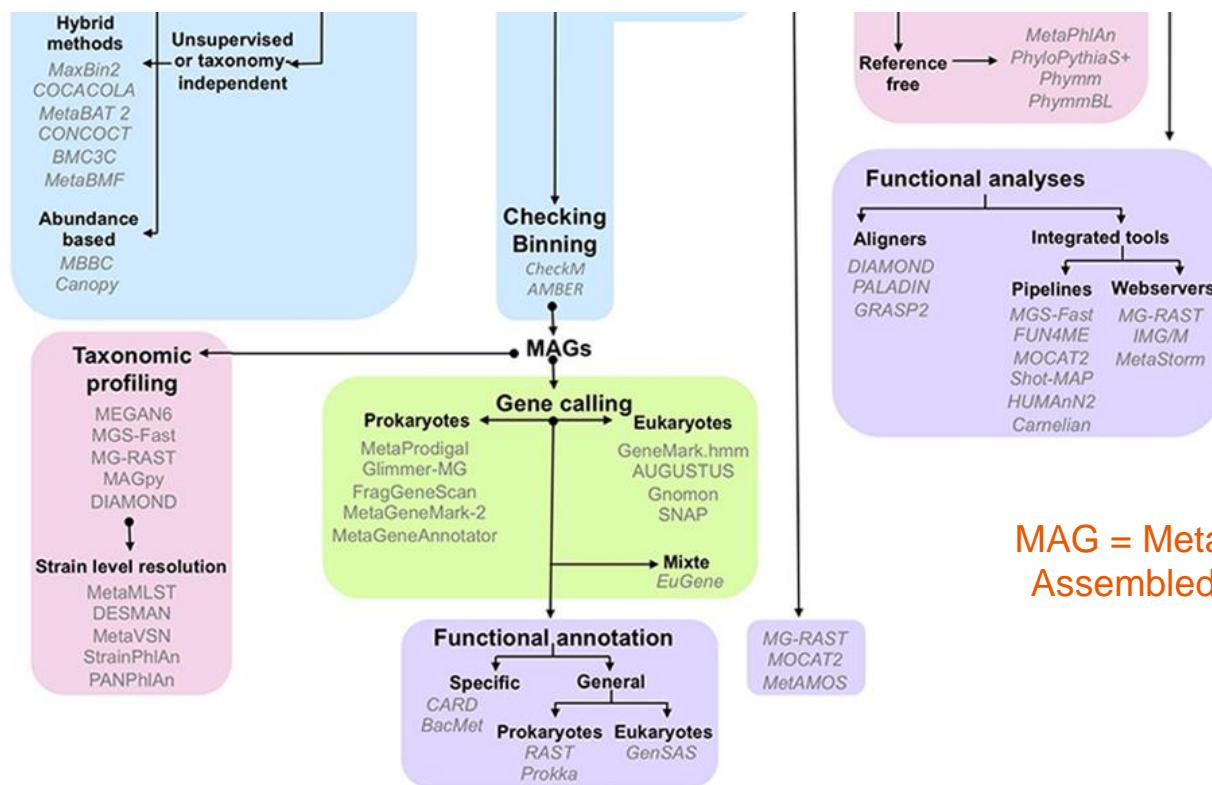
Amplicon analysis pipeline (16S rRNA)



Shotgun analysis pipeline



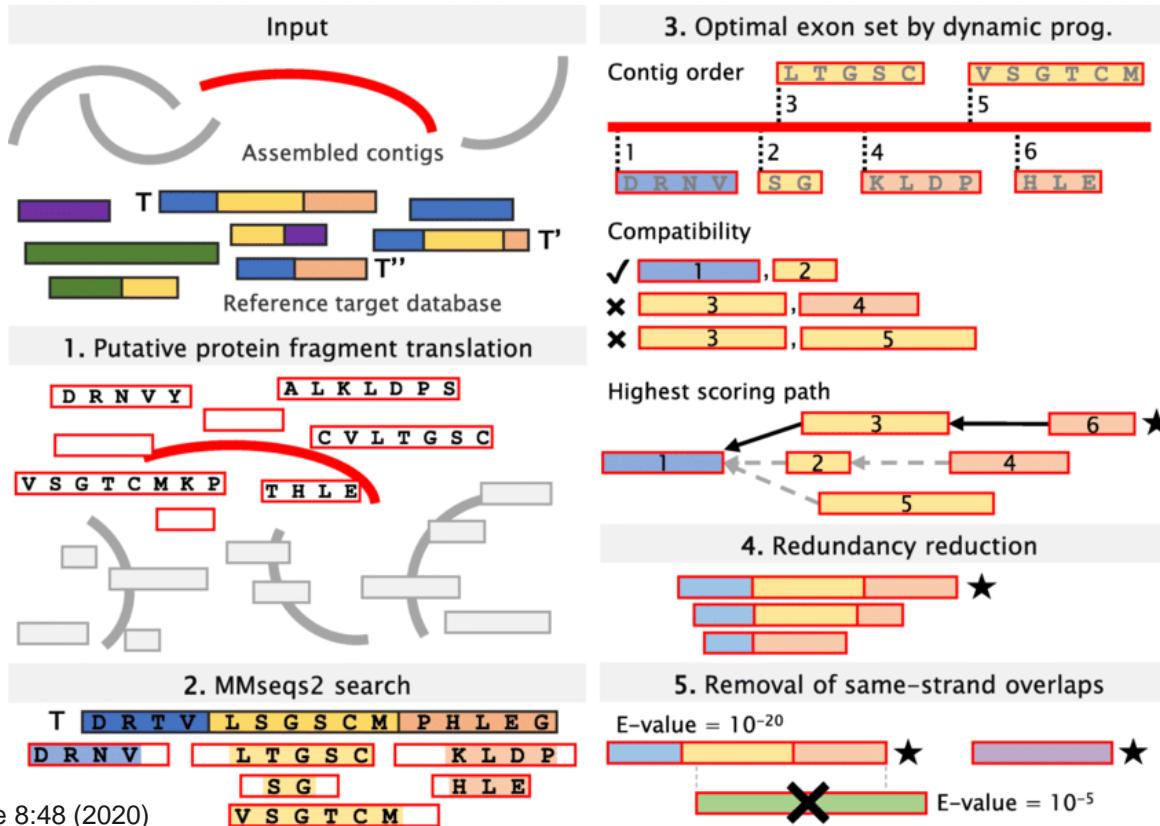
Shotgun analysis pipeline



Gene prediction

Six-frame translation

Align to known proteins

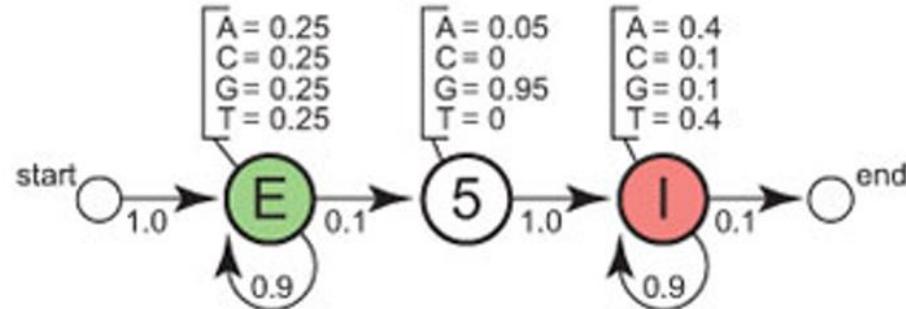


Order exons

Remove duplicated genes

Remove contradicting predictions

Hidden Markov Model for gene prediction



sequence: CTTCATGTGAAAGCAGACGTAAGTCA

state path: E E E E E E E E E E E E E E E 5 I I I I I I I

log P
-41.22

parsing:
-43.90
-43.45
-43.94
-42.58
-41.71

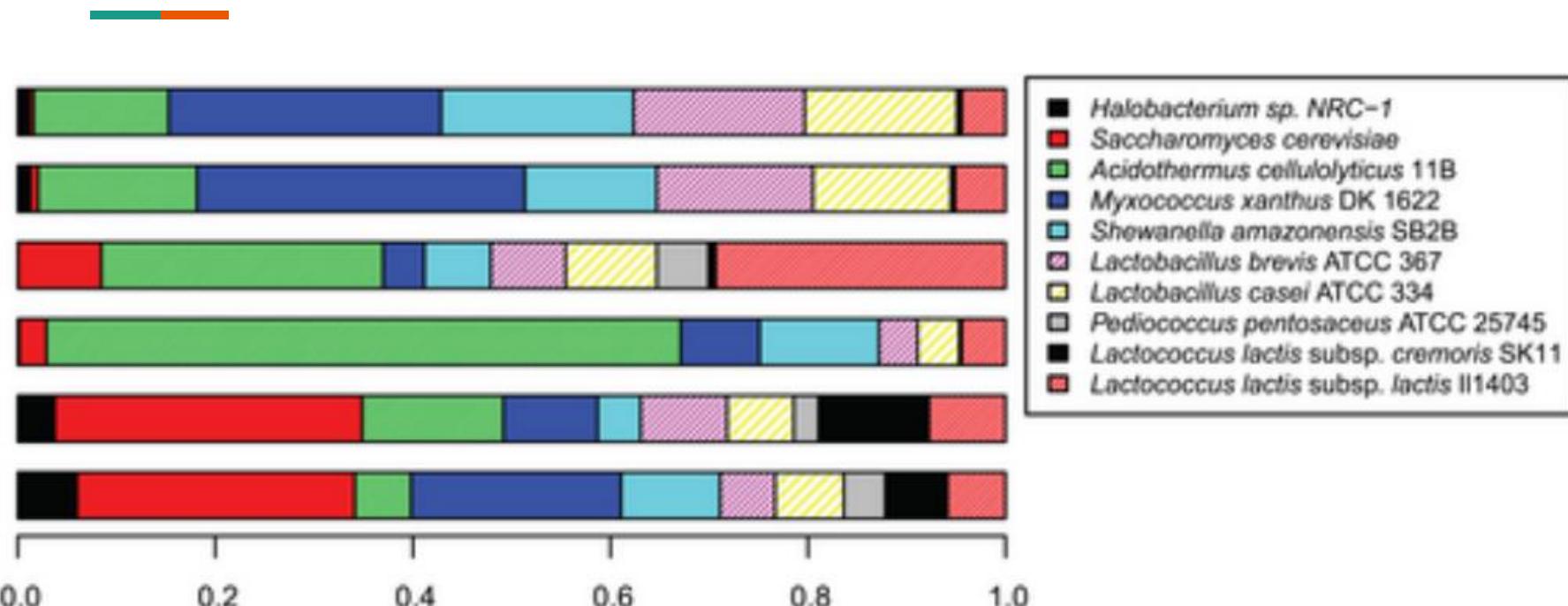
Highest likelihood

- Codon usage frequency
- Patterns in upstream / downstream sequences



Variability of microbiome analysis

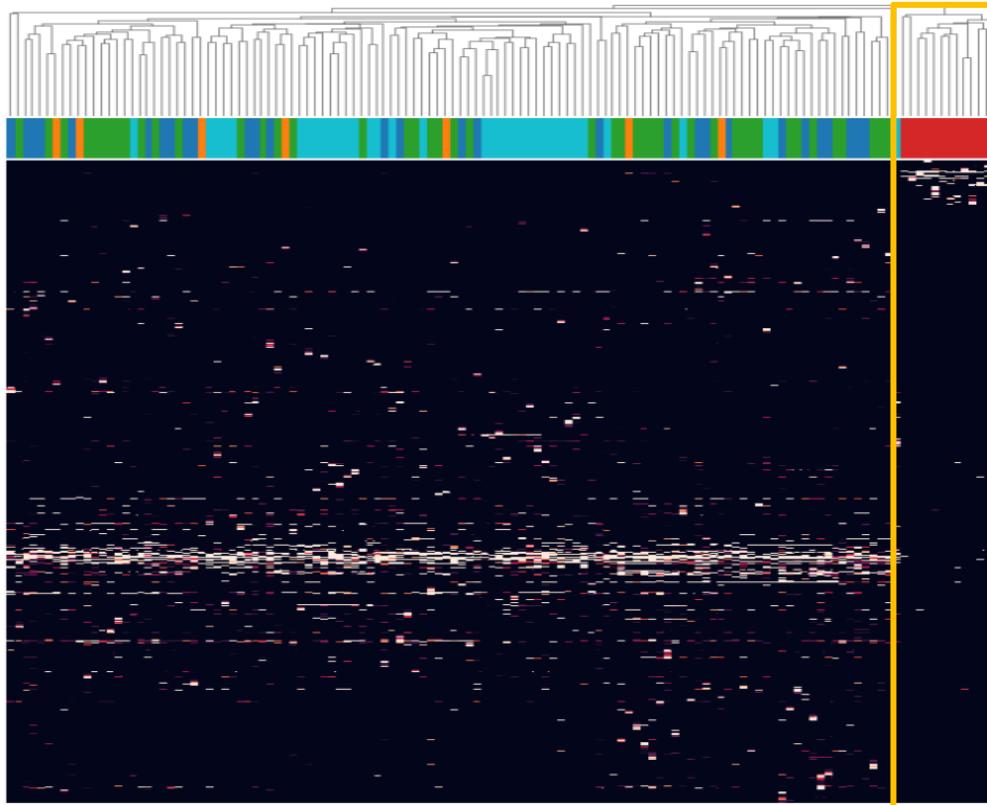
Same sample, different profiles



Source: <http://www.cbs.dtu.dk/courses/27626>

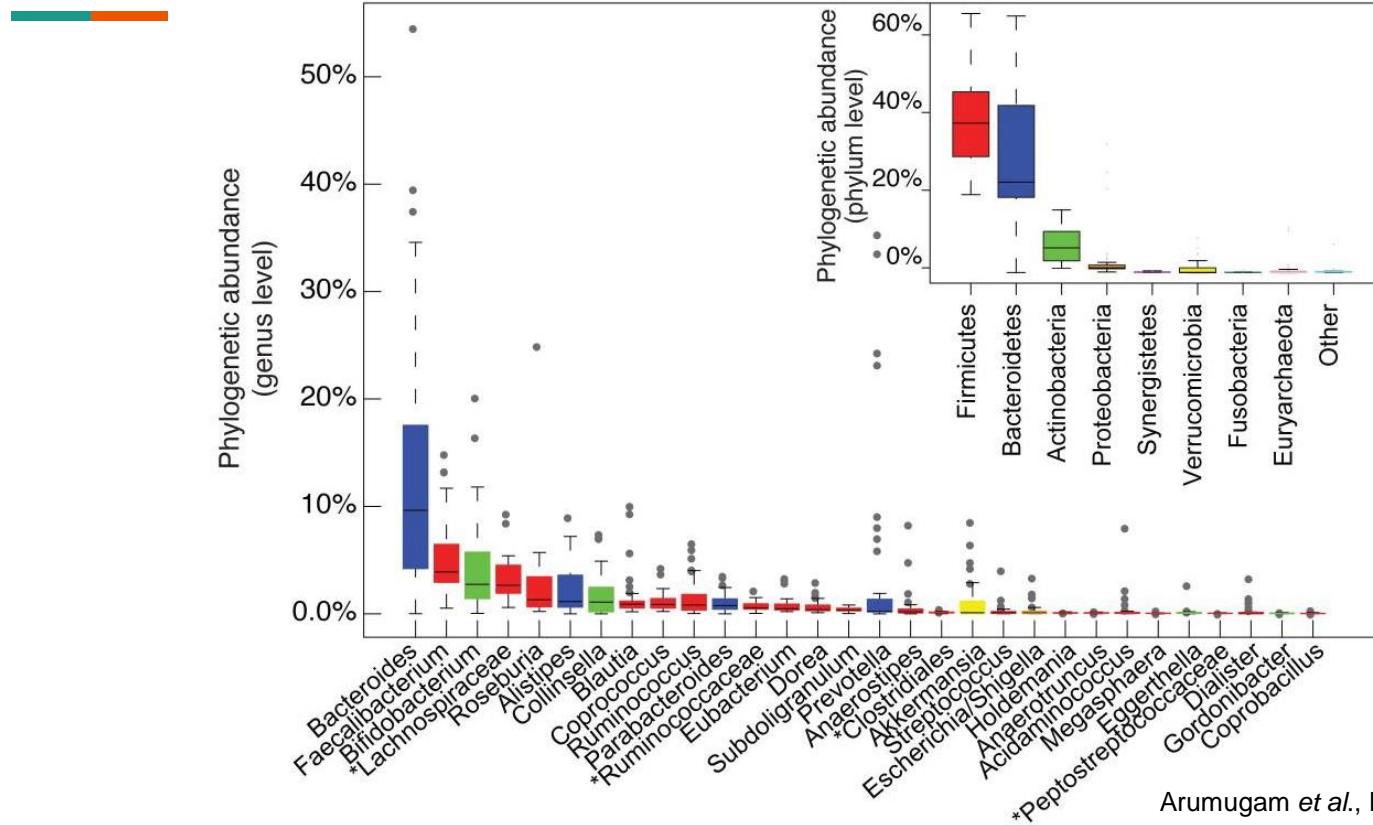
Batch effects

Operational Taxonomic Unit

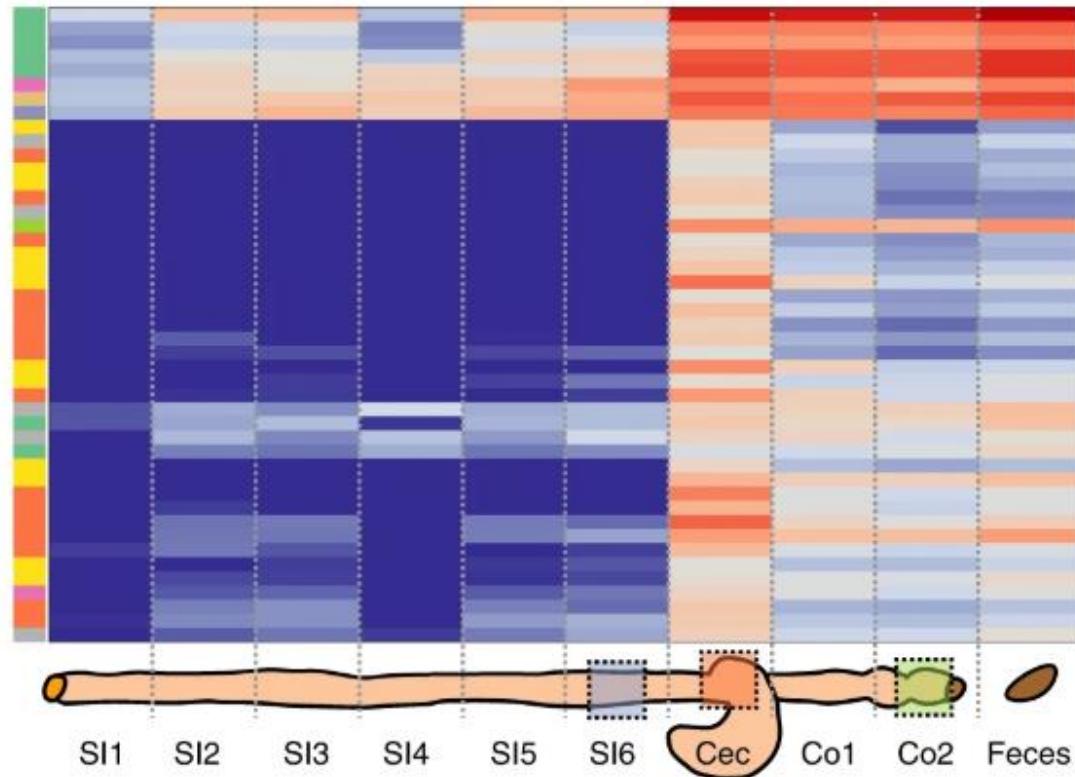
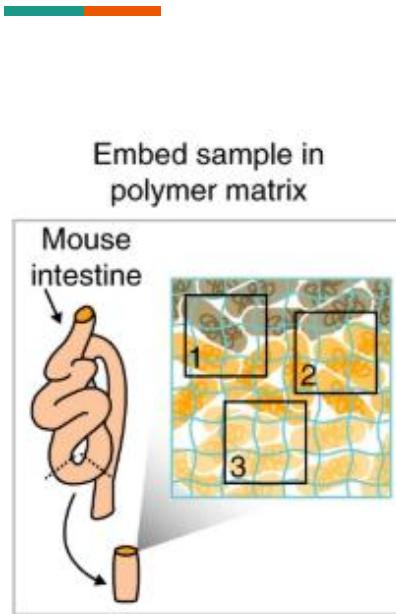


Red = same batch

Variability of human gut microbiomes



Spatial variability



Sheth et al., Nature Biotech 37: 877-883 (2019)

Any question?

- See you on September 8th 1-2:30pm