3000788 Intro to Comp Molec Biol

Lecture 8: Metagenomics

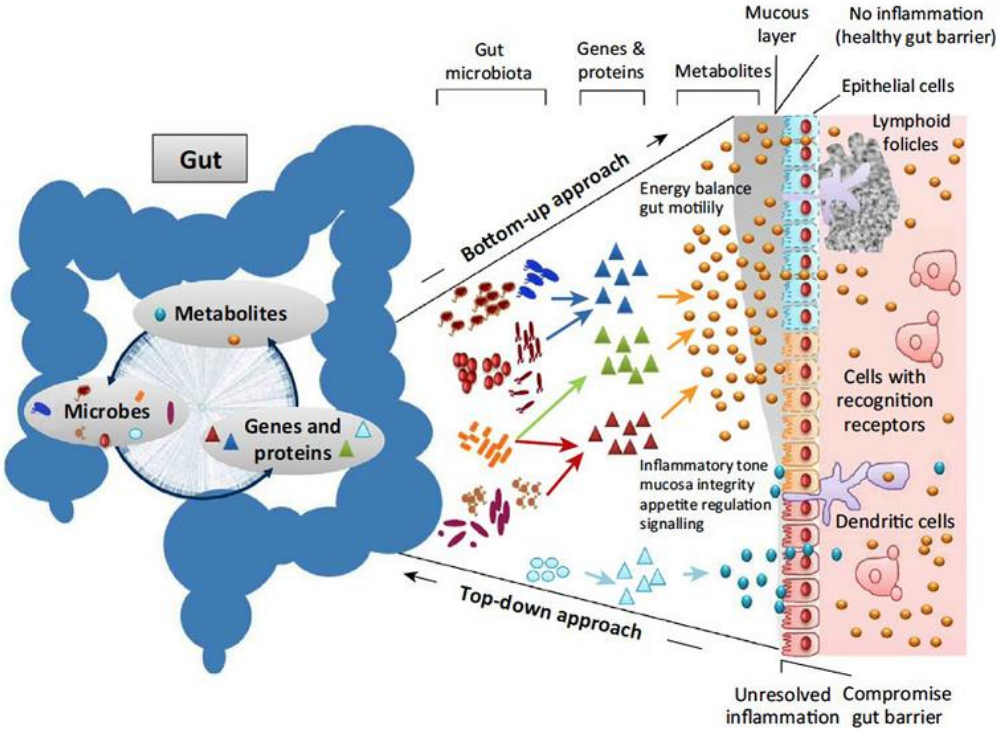
September 11, 2023

Sira Sriswasdi, PhD

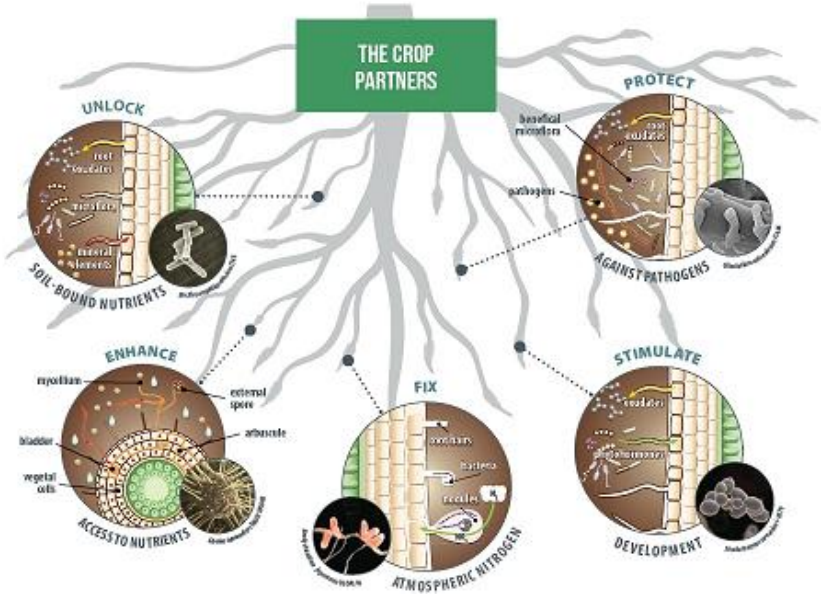
- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

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

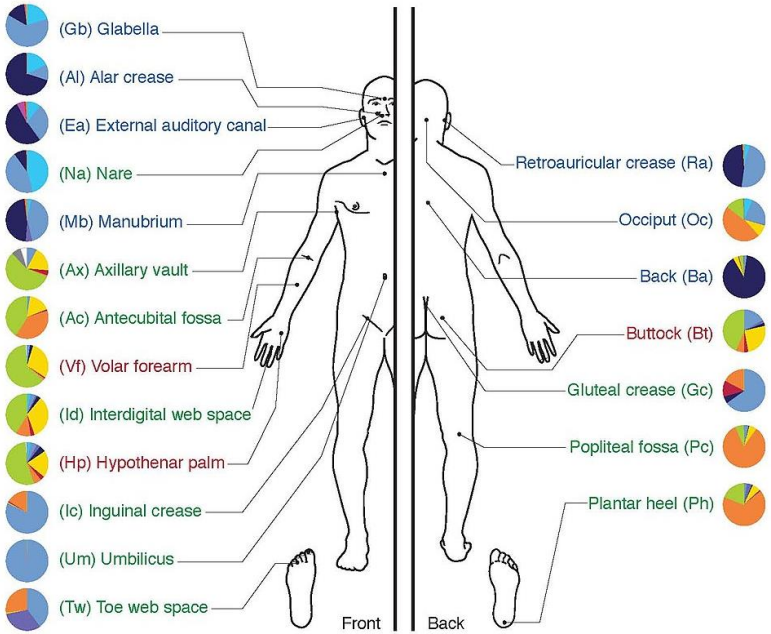
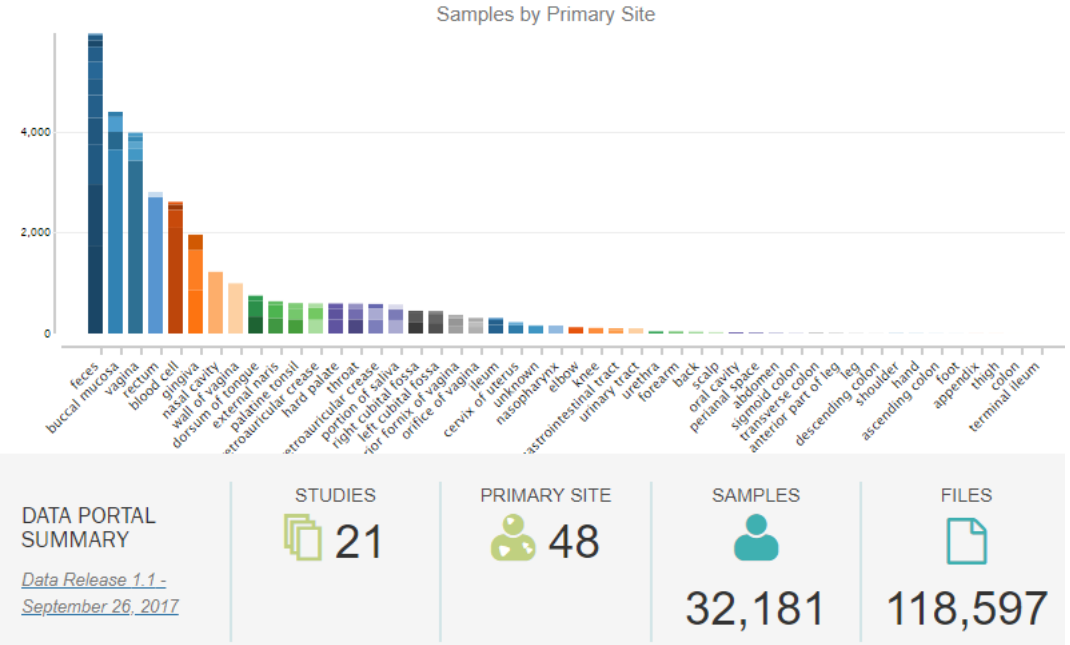


Image from Wikipedia.com



Omics analysis of a collection of organisms

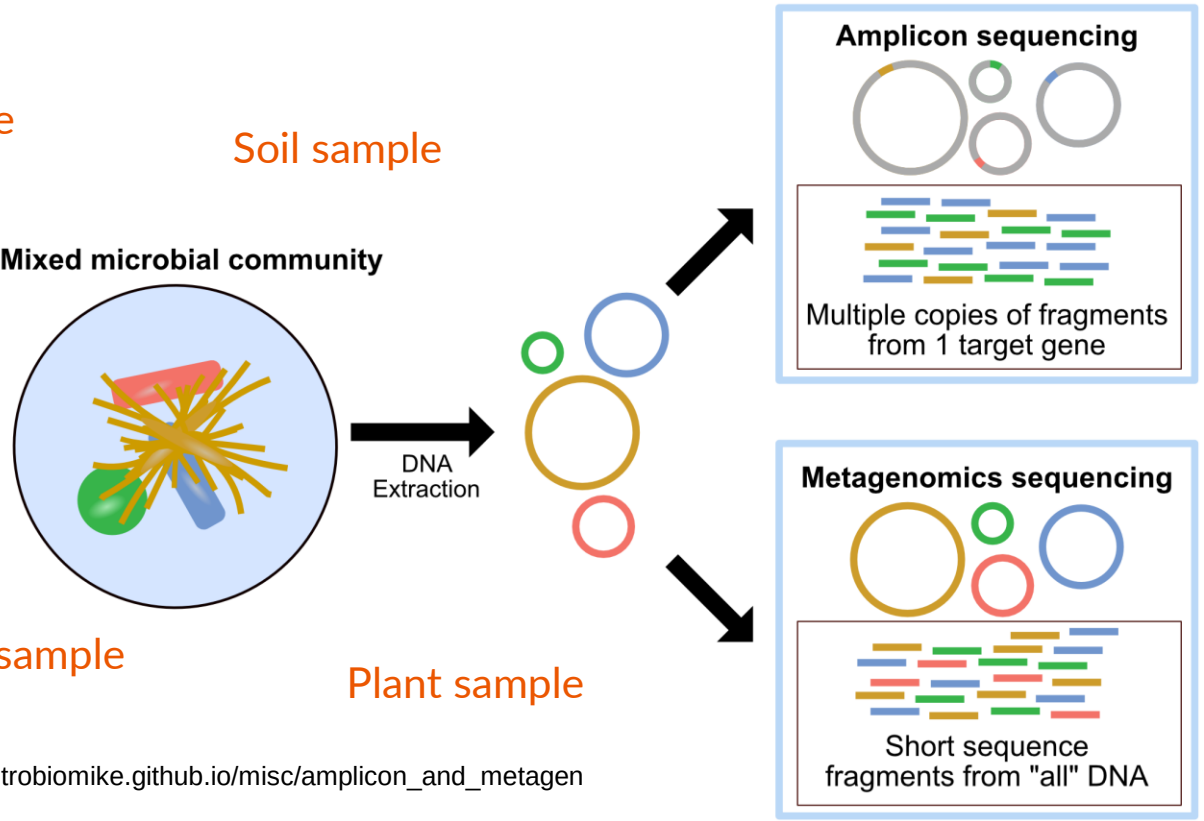
Water sample

Soil sample

Stool sample

Factory waste sample

Plant sample



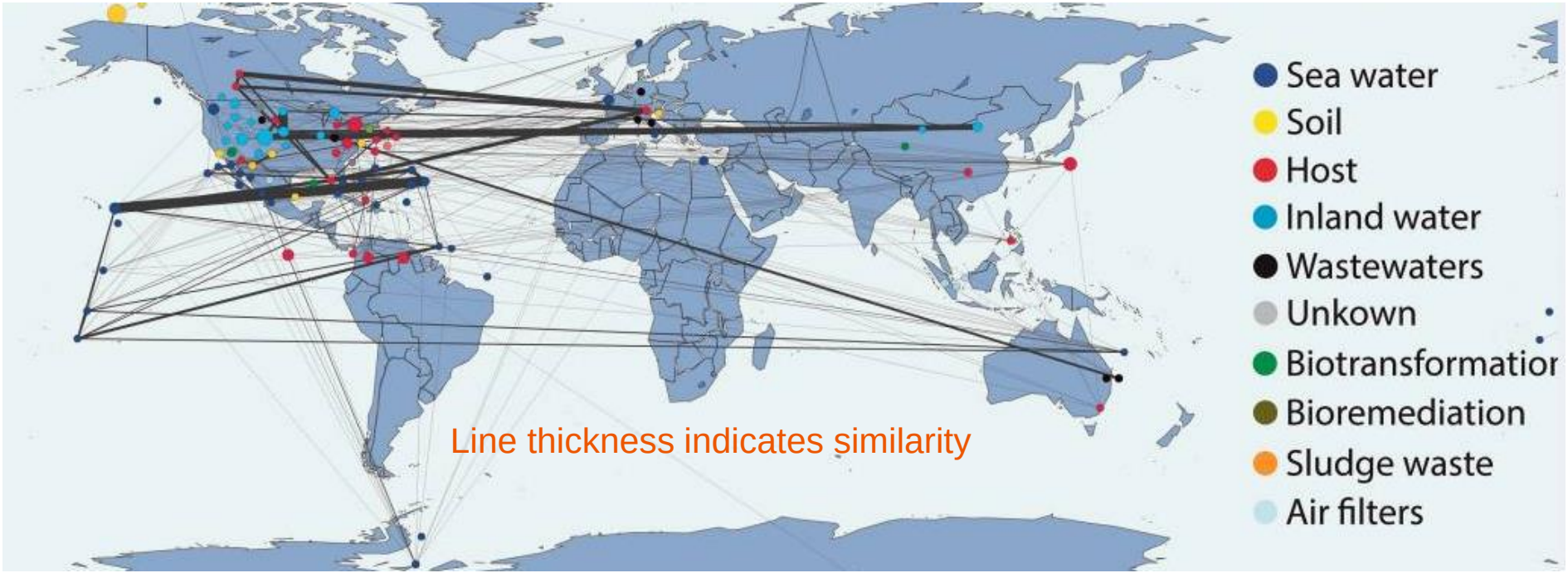
https://astrobiomike.github.io/misc/amplicon_and_metagen

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

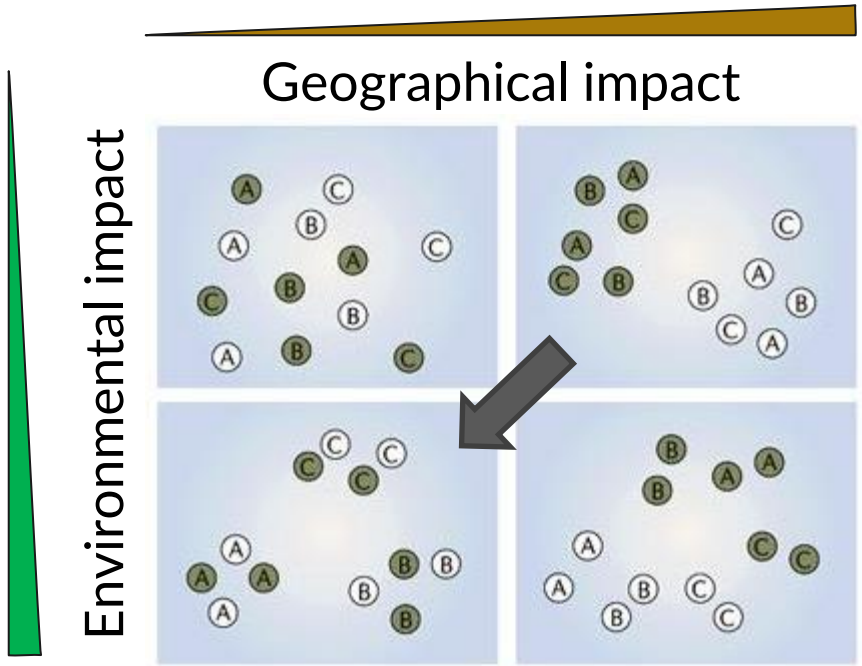
- Metaproteomics:
  - Require reference protein database to interpret mass spectrometry data

Everything is everywhere, but the nature selects



Fondi *et al.*, Genome Biol Evol 8: 1388-1400 (2016)

Environment, not geography, defines micriobiome



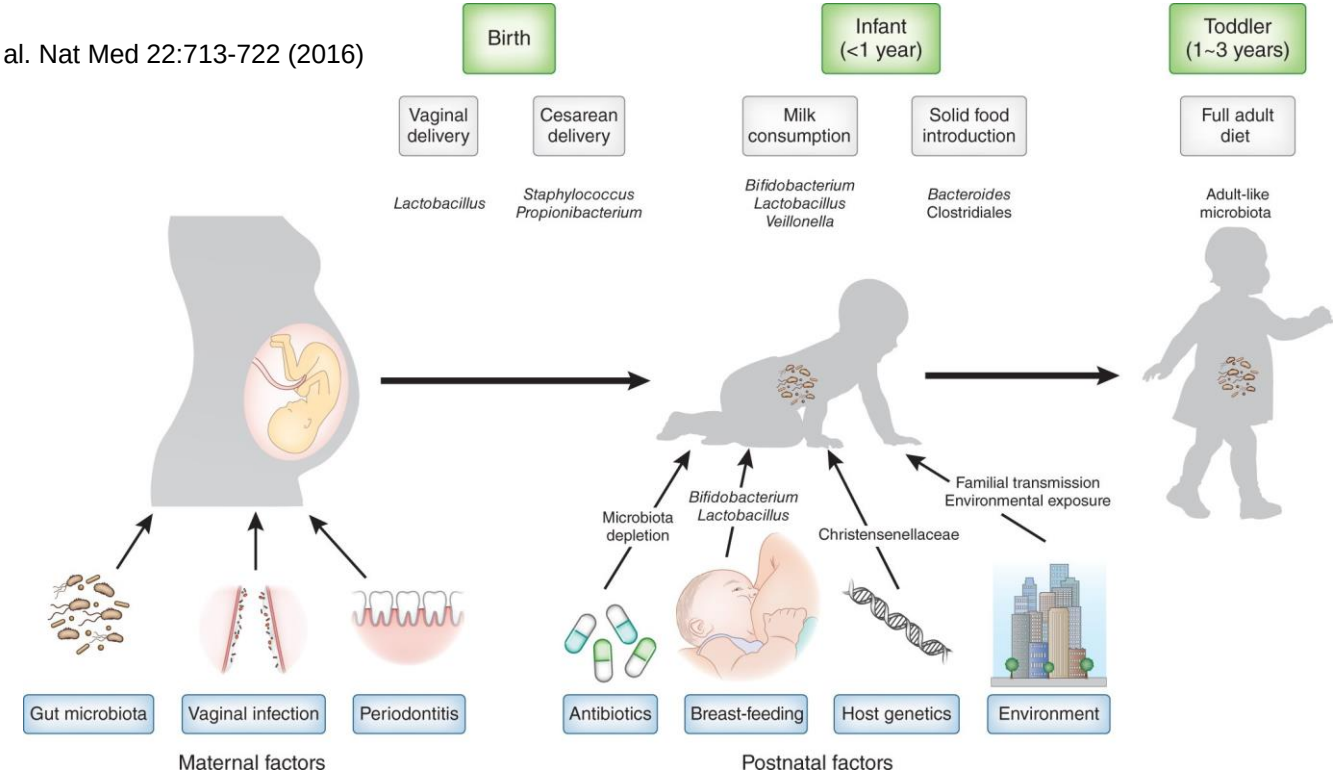
indicate same geographical locations

A, B, C indicate same environmental conditions

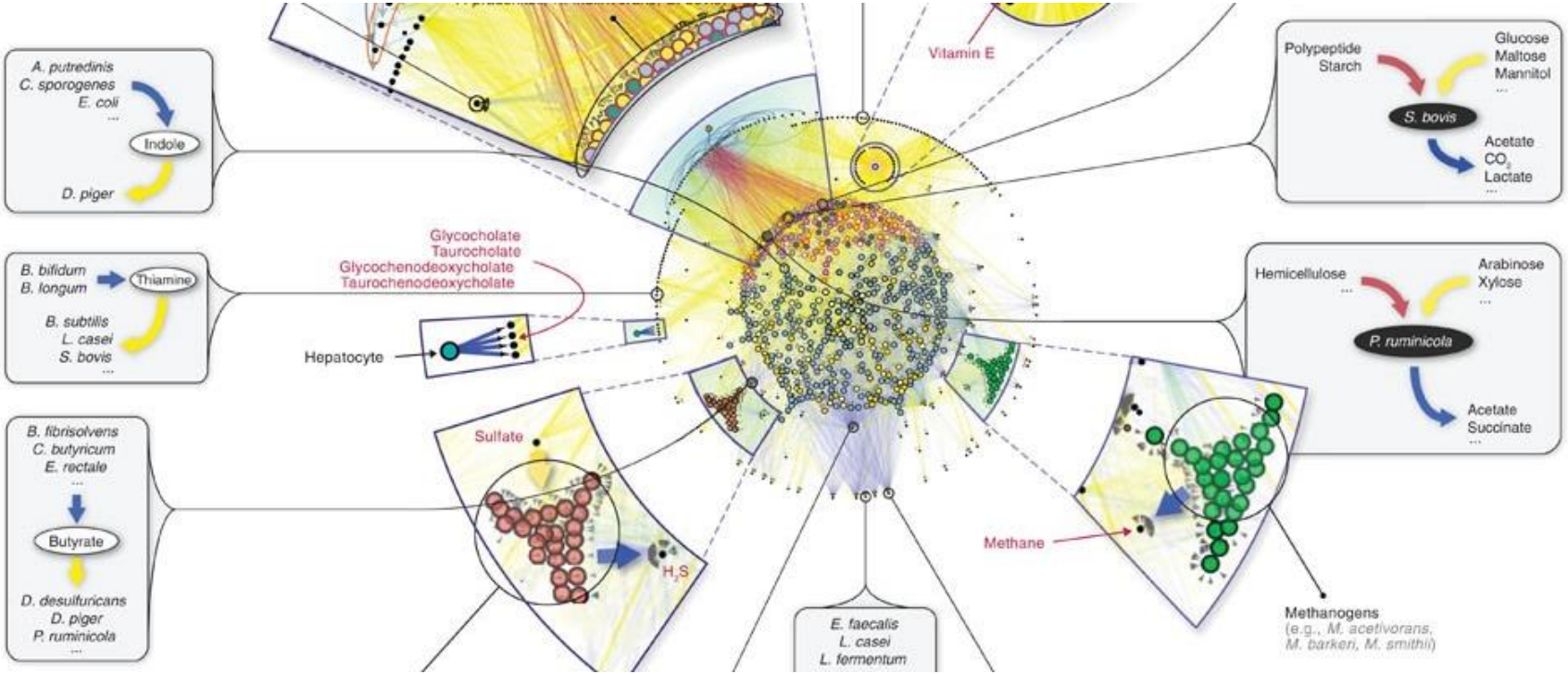
Adapted from Martiny *et al.*, Nature Review Microbiology 4: 101-112 (2006)

Microbiome is dynamics

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



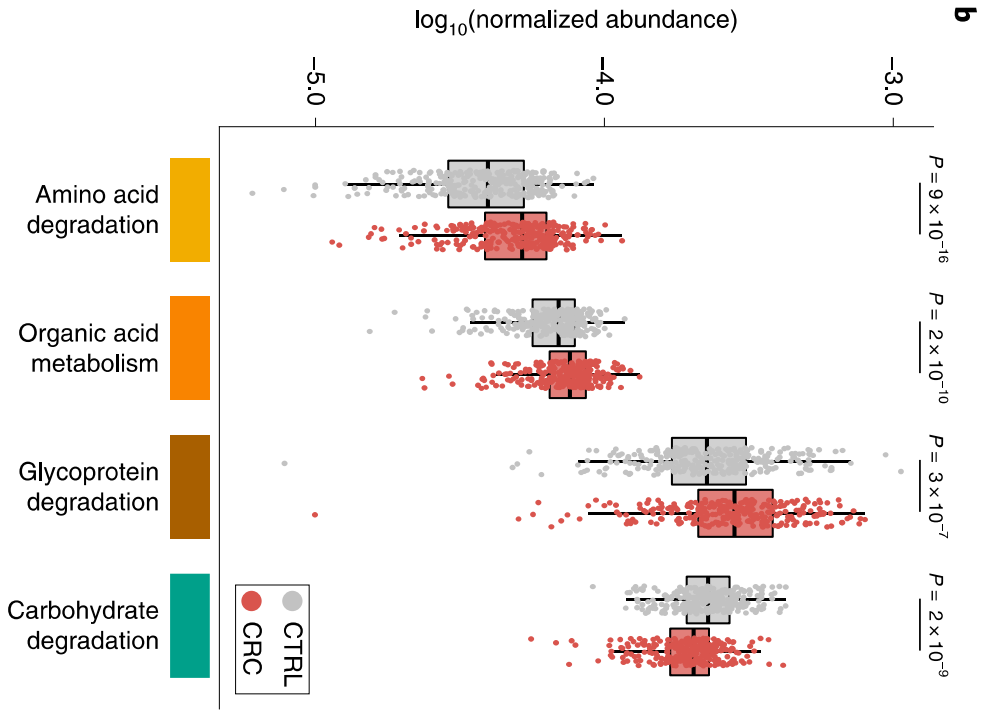
Microbiome is a network of cooperation and competition



Sung, J. et al. Nat Comm 8:15393 (2017)

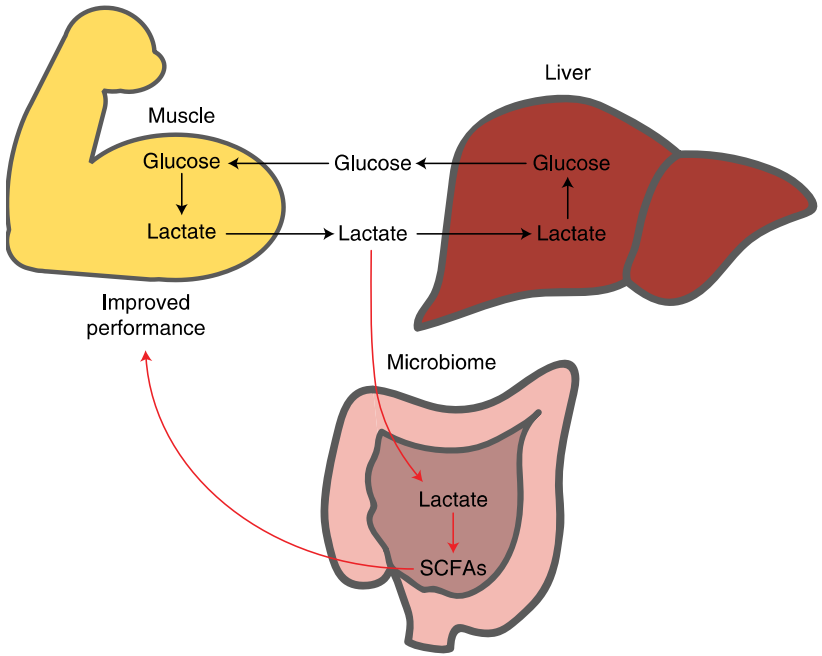
Applications of meta-omics

Microbiome link to health



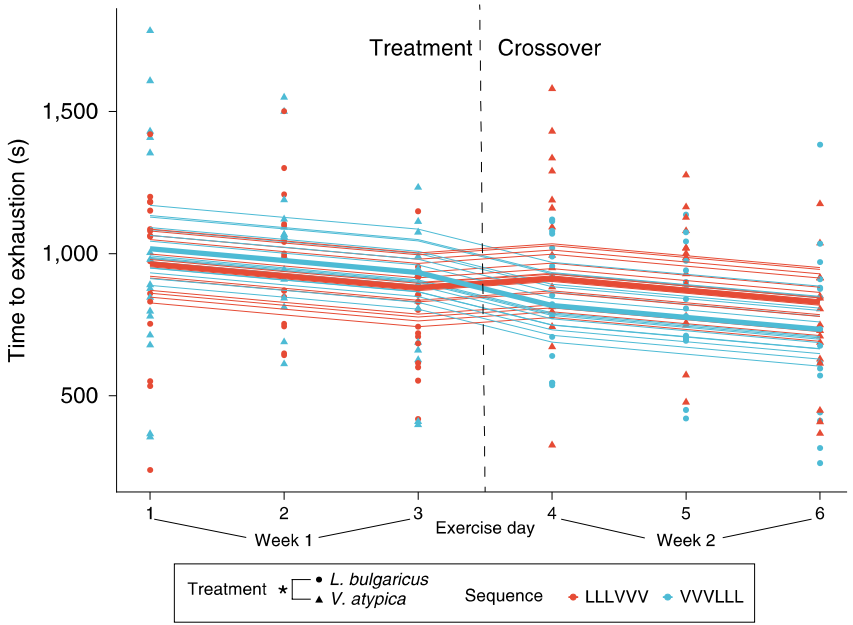
Wirbel *et al*. Nat Med 25:679-689 (2019)

Lactate-utilizing bacteria in athelete guts

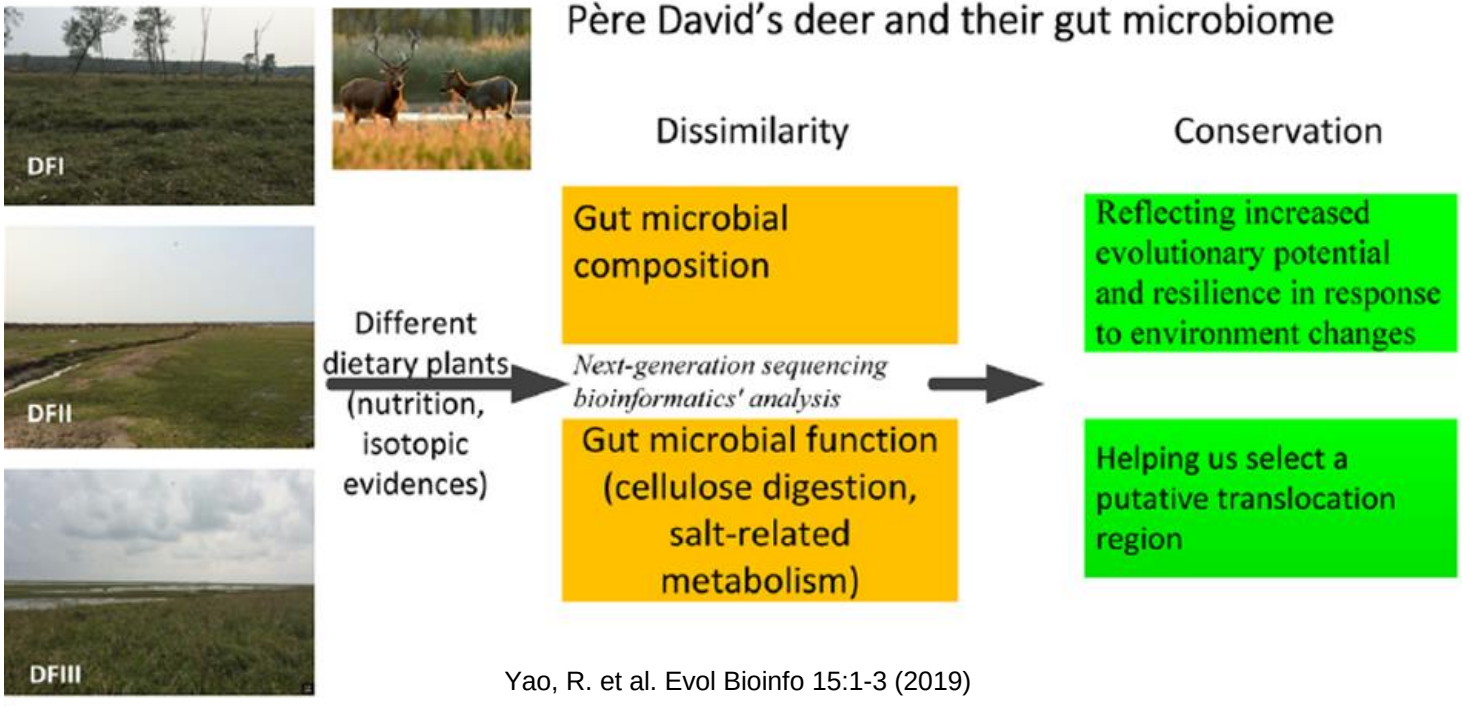


Scheiman *et al*. Nat Med 25:1104-1109 (2019)

Increased athletic ability in mice with transplanted microbiome

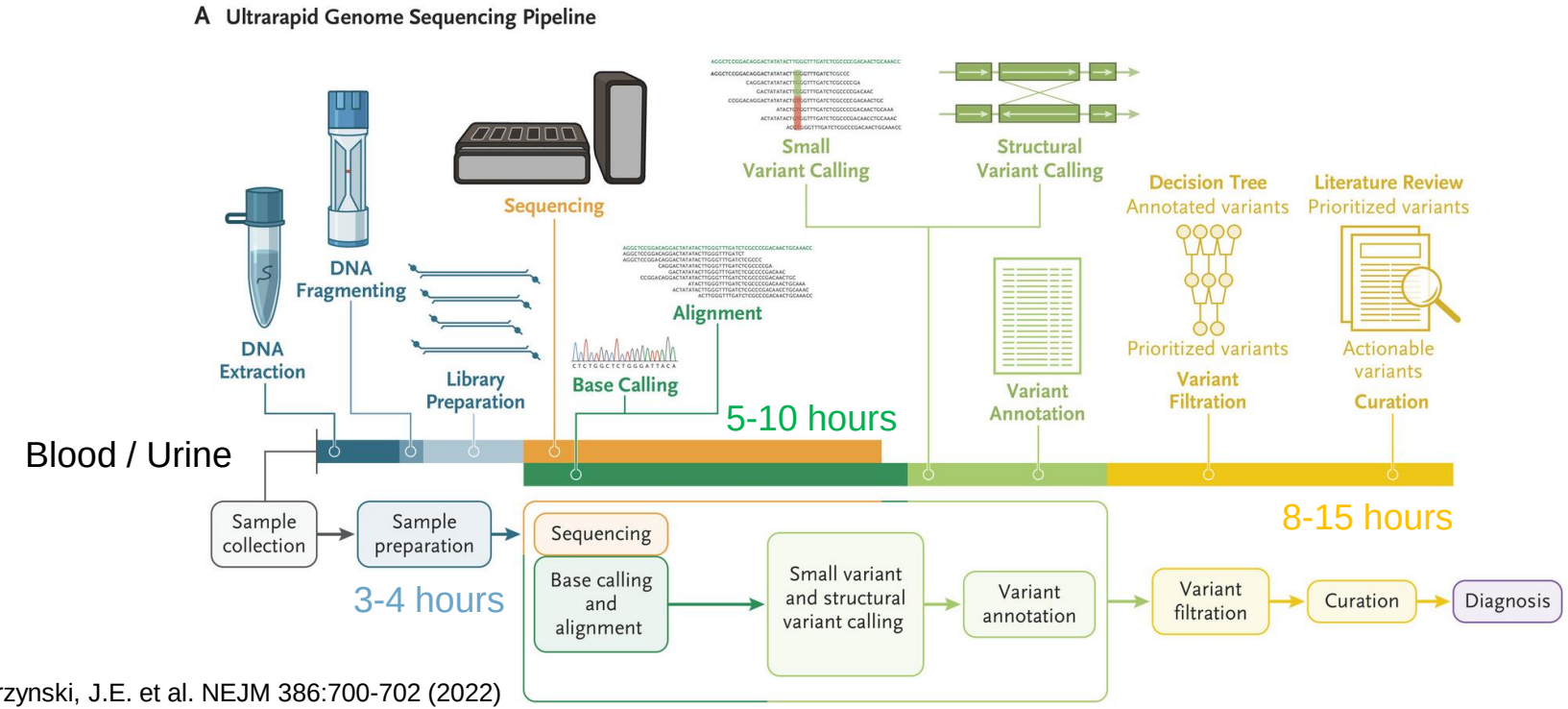


Wildlife conservation



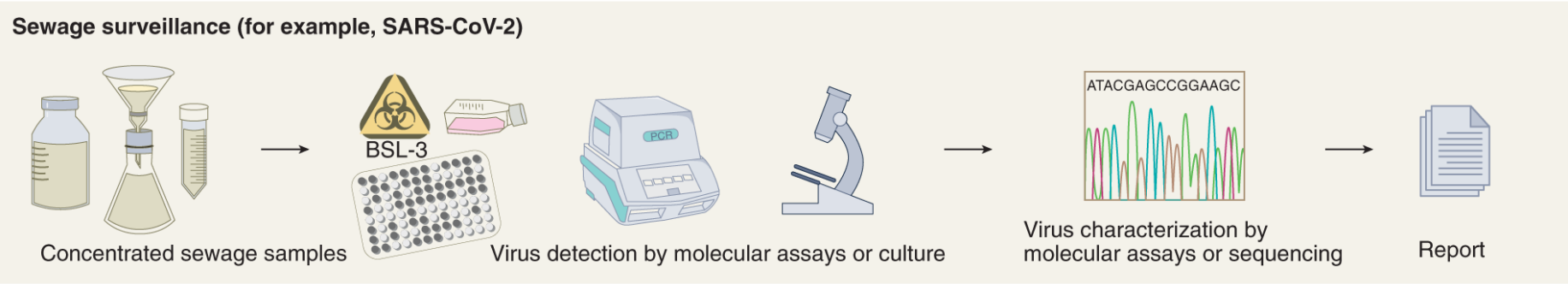
Yao, R. et al. Evol Bioinfo 15:1-3 (2019)

Rapid pathogen detection for clinical decision



Gorzynski, J.E. et al. NEJM 386:700-702 (2022)

Pathogen surveillance



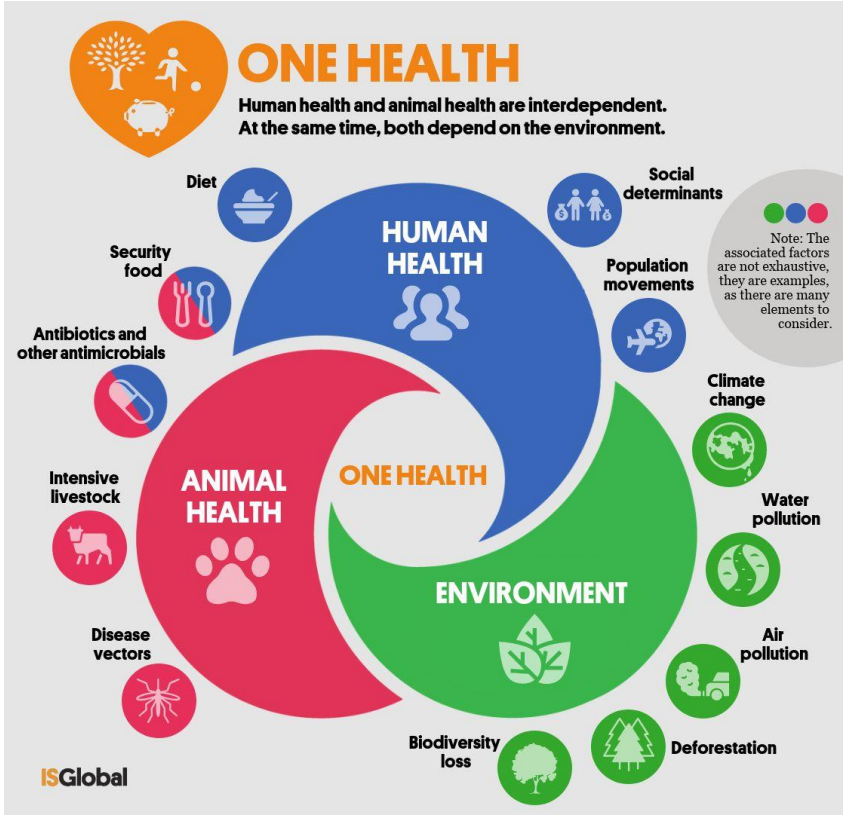
Ko, K.K.K. *et al*. Nat Micro 7:486-496 (2022)

- 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

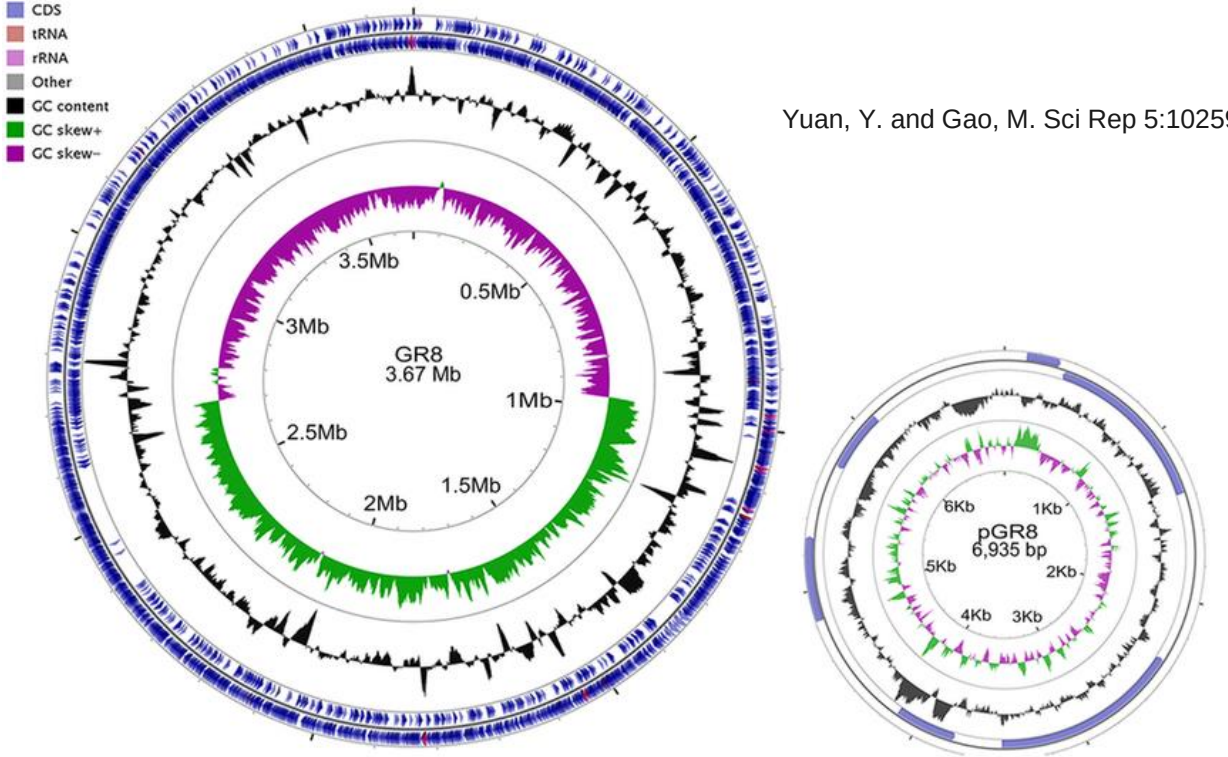
Metagenomics for One Health



https://www.isglobal.org/en/healthisglobal

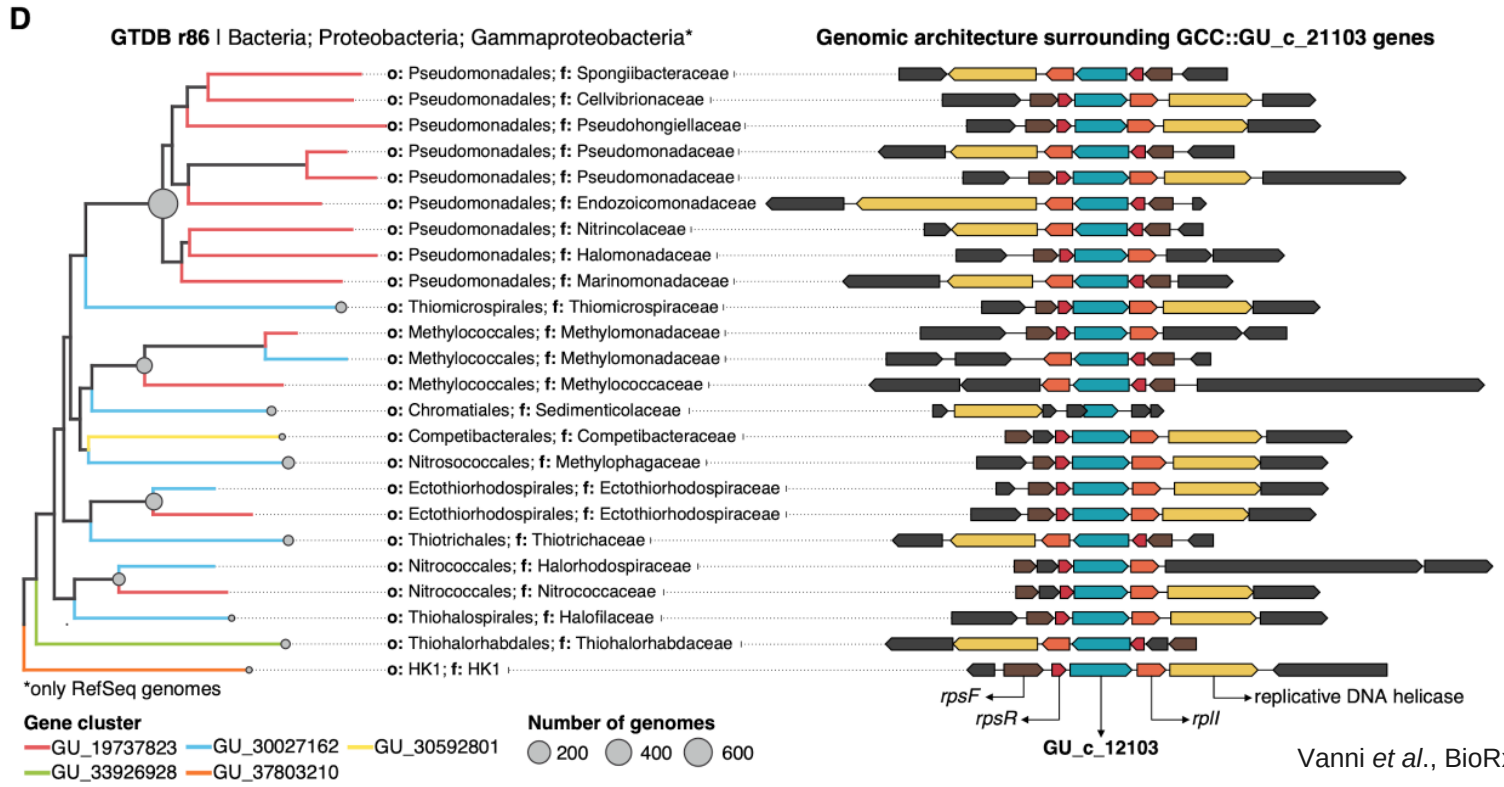
Prokaryotic genomes

Circular chromosomes and plasmids



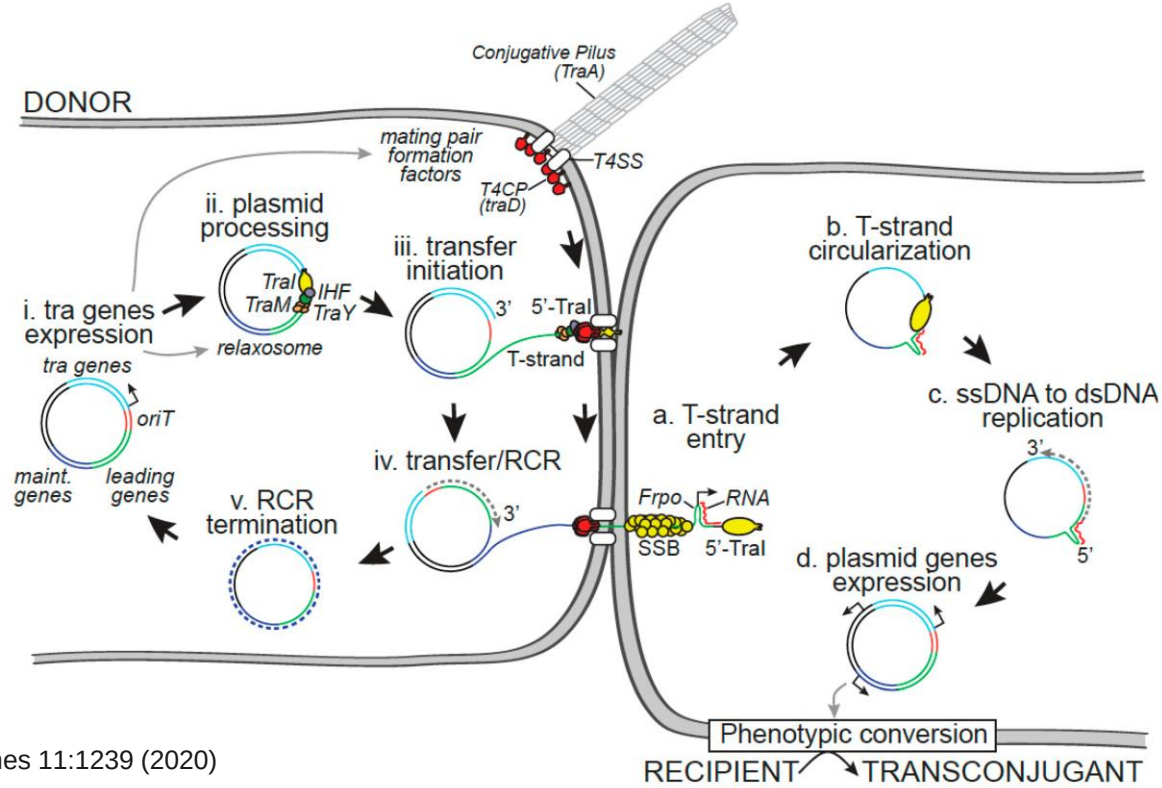
Yuan, Y. and Gao, M. Sci Rep 5:10259 (2015)

Gene operons



Vanni *et al.*, BioRxiv (2020)

Plasmid transfer

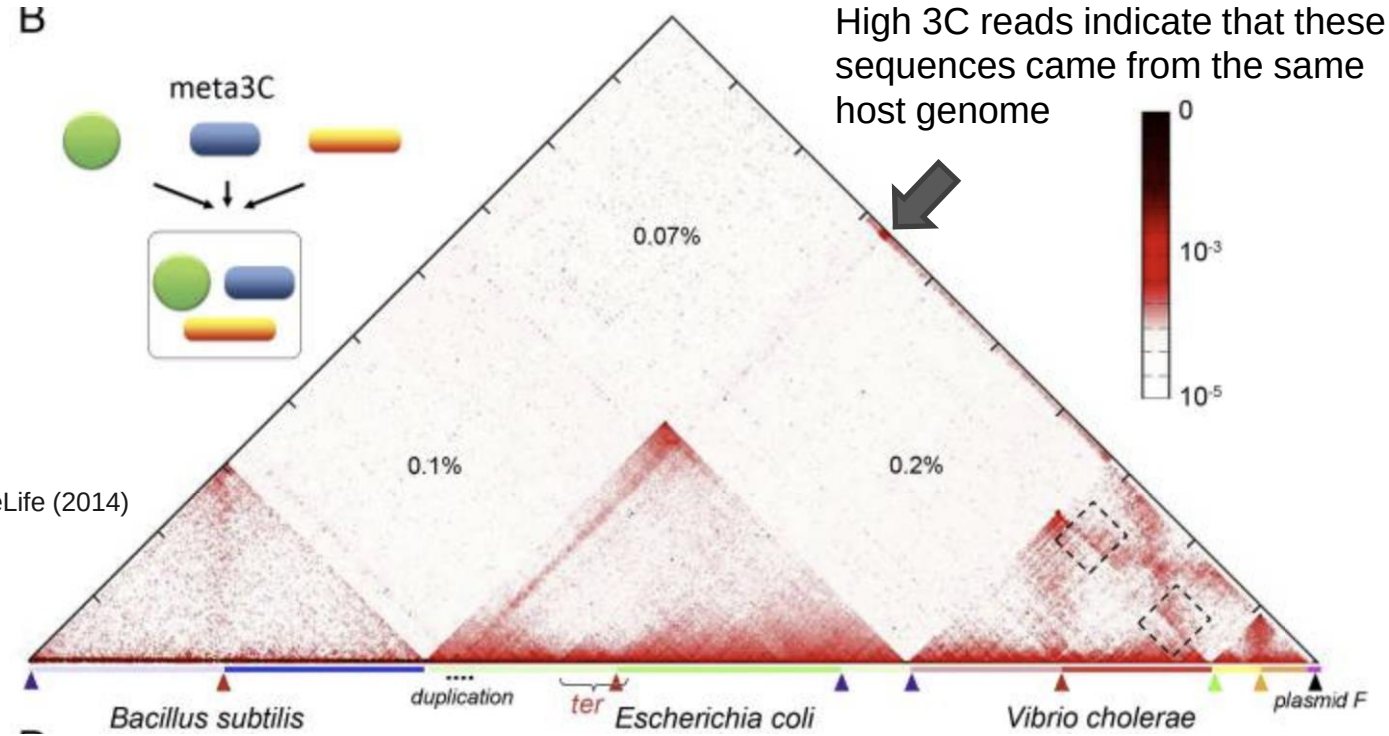


Virolle, C. *et al.*, Genes 11:1239 (2020)

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

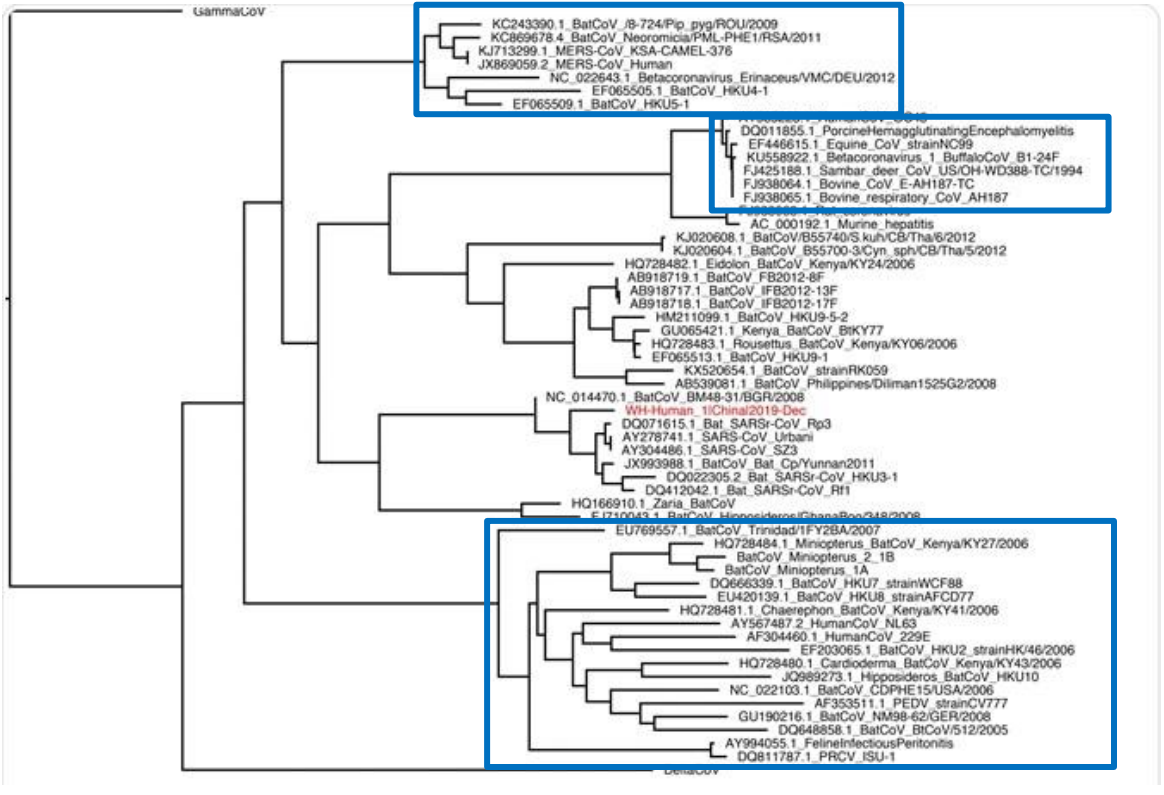


High 3C reads indicate that these sequences came from the same host genome

Marbouty *et al.*, eLife (2014)

Operational texonomic unit (OTU)

Cluster of sequences with high similarity



Rough OTU similarity thresholds

nature reviews microbiology

Published: 14 August 2014

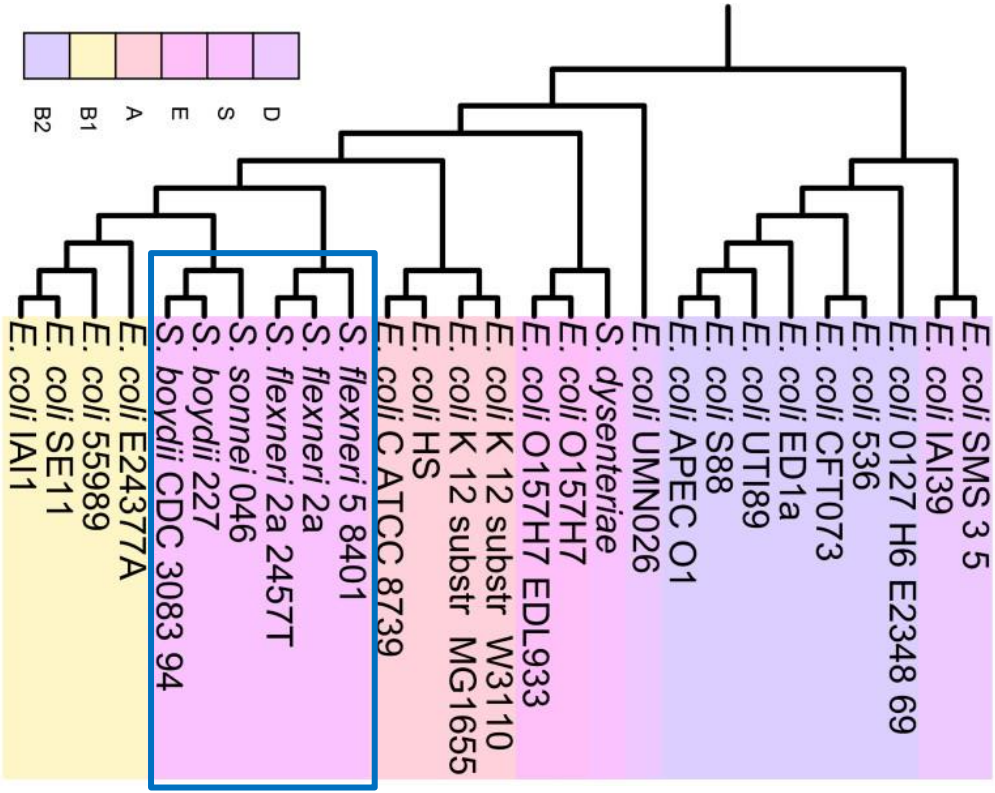
Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences

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

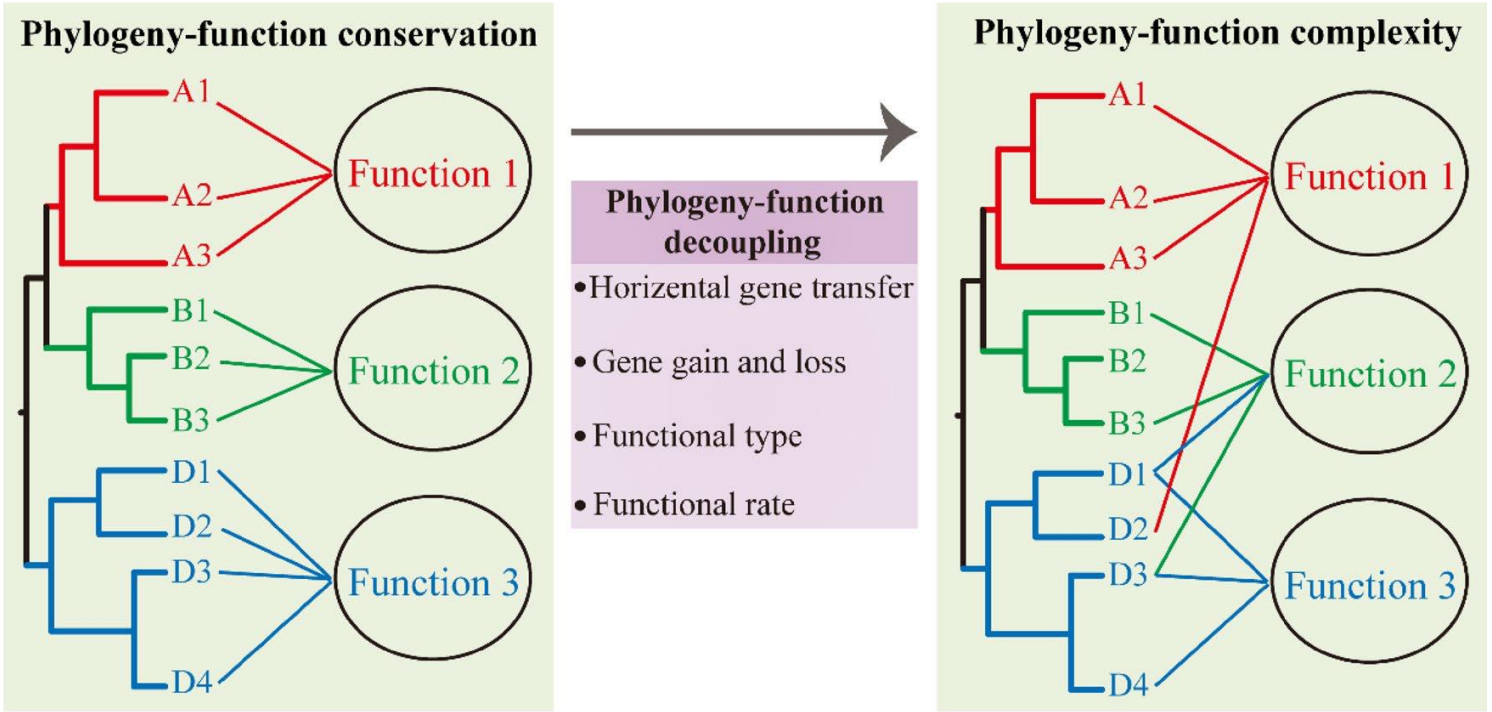
An exception

- 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



Skippington and Ragan, BMC Genomics 12:532 (2011)

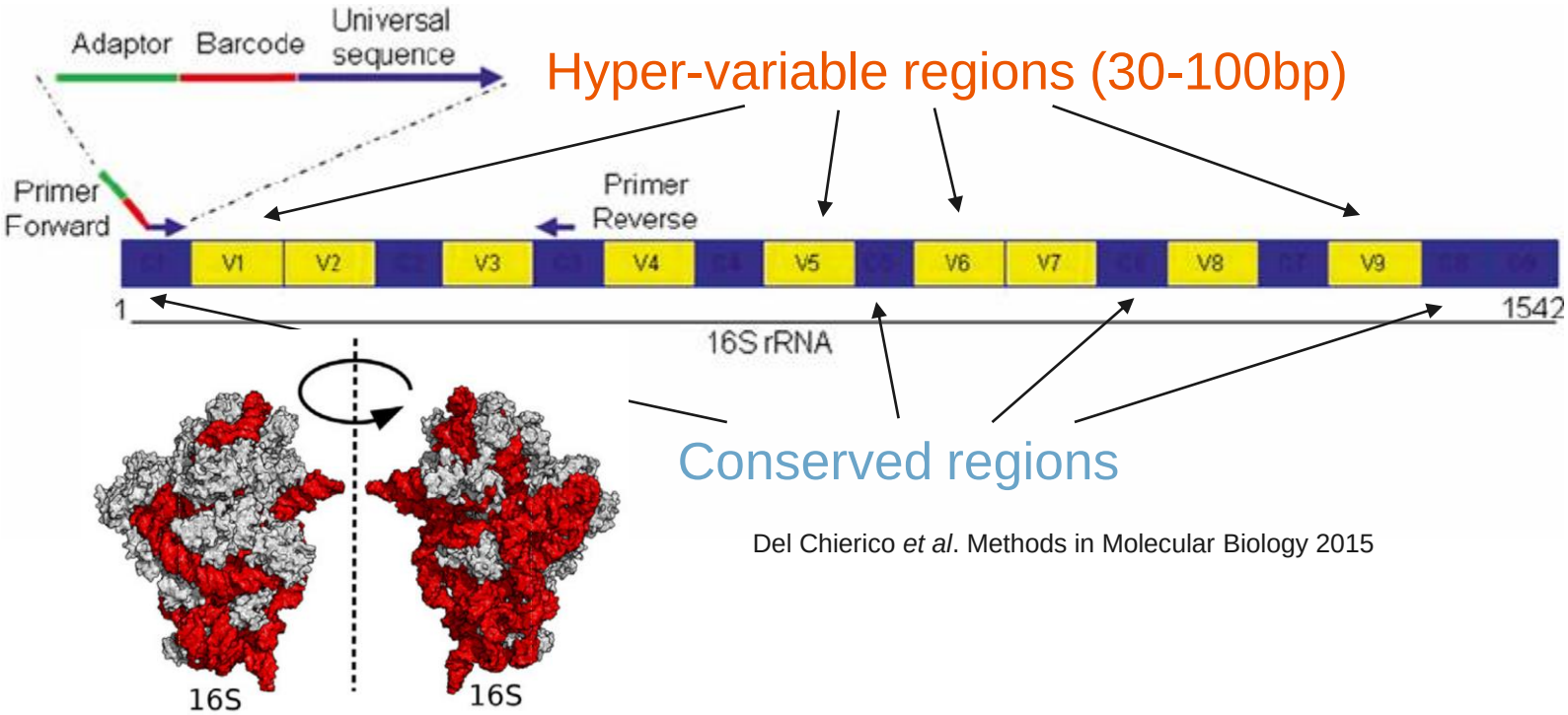
Functional view of taxonomy



Liu *et al*., Marine Life Science & Technology 1: 112-128 (2019)

Taxonomy analysis via rRNA loci

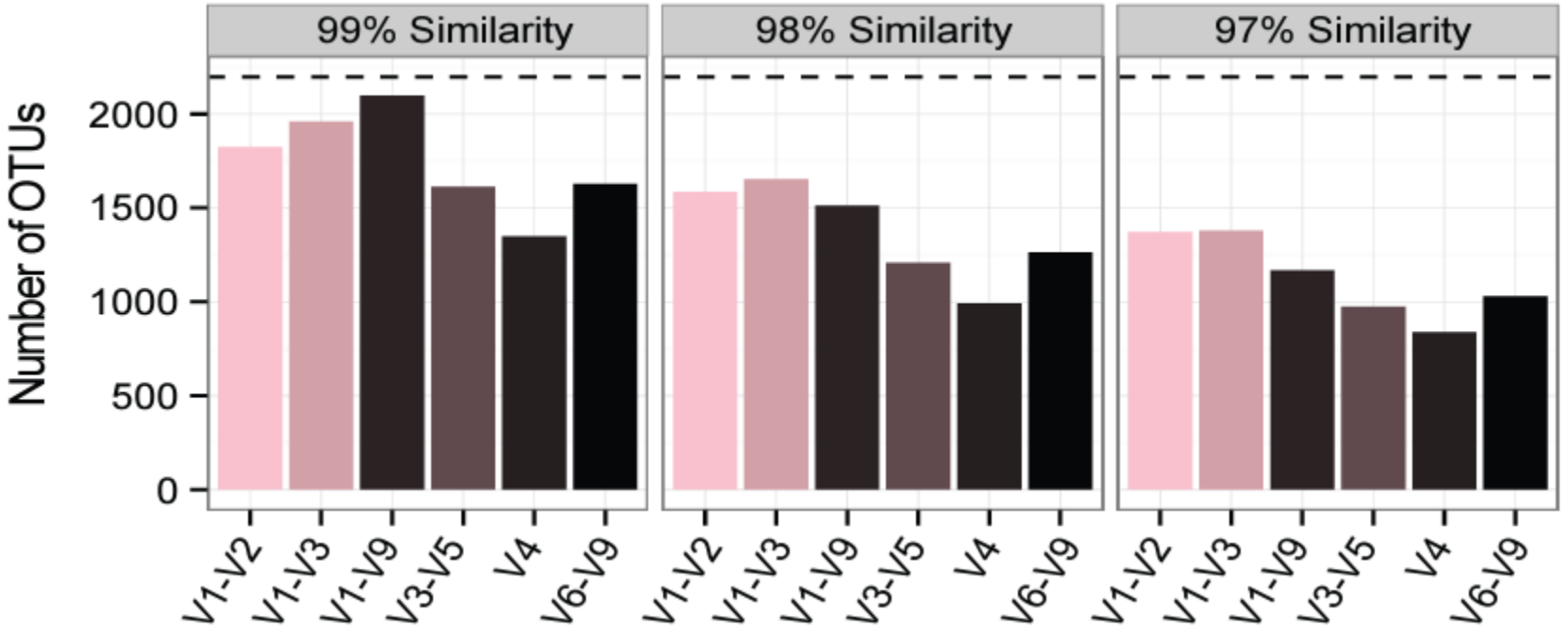
16S rRNA in prokaryotes



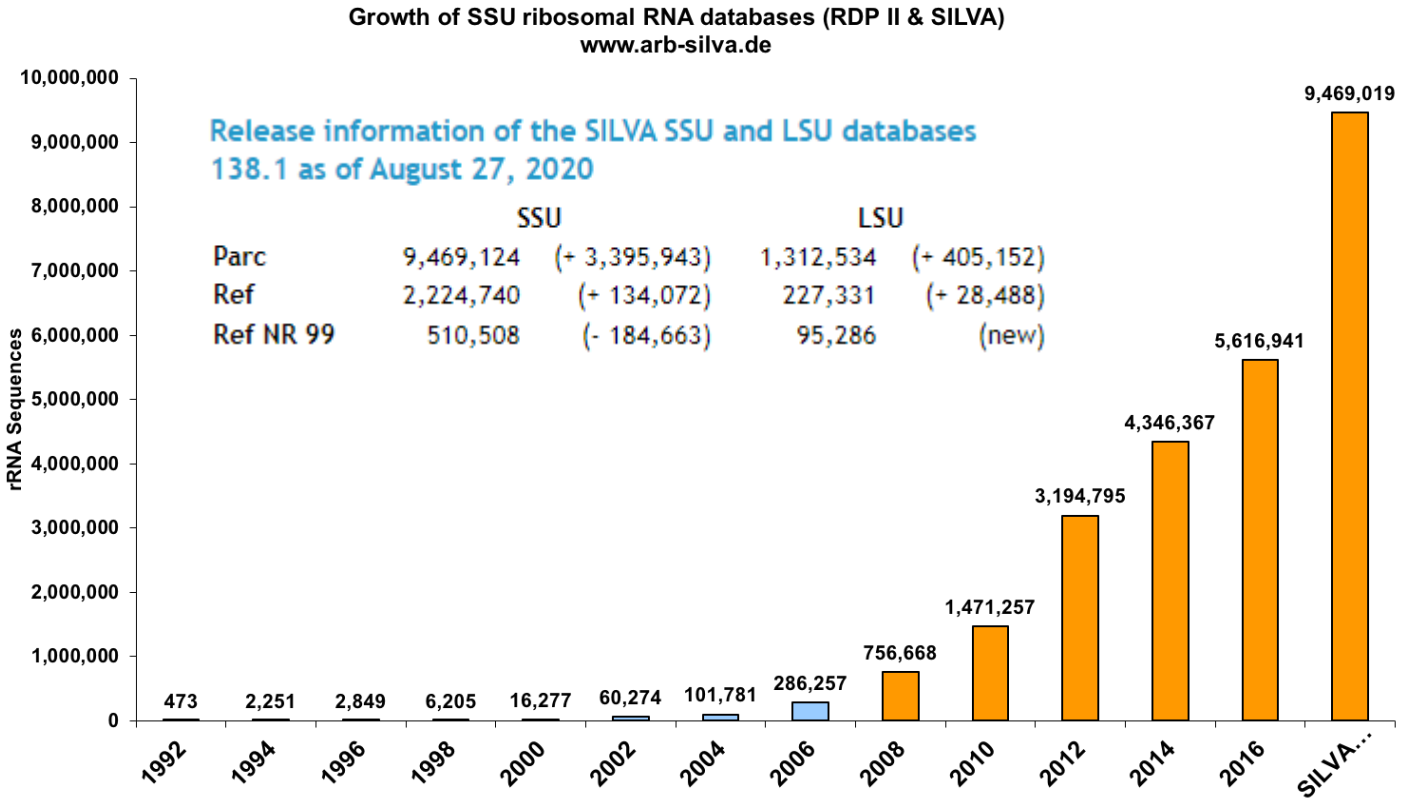
Del Chierico *et al*. Methods in Molecular Biology 2015

Ramazzotti and Bacci, Metagenomics: Perspectives, Methods, and Applications,103-119 (2018)

Impact of hypervariable region choices



Known rRNA sequences



rRNA databases

silva
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

ANNOUNCEMENTS

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***.

GREENGENES

The Greengenes Database

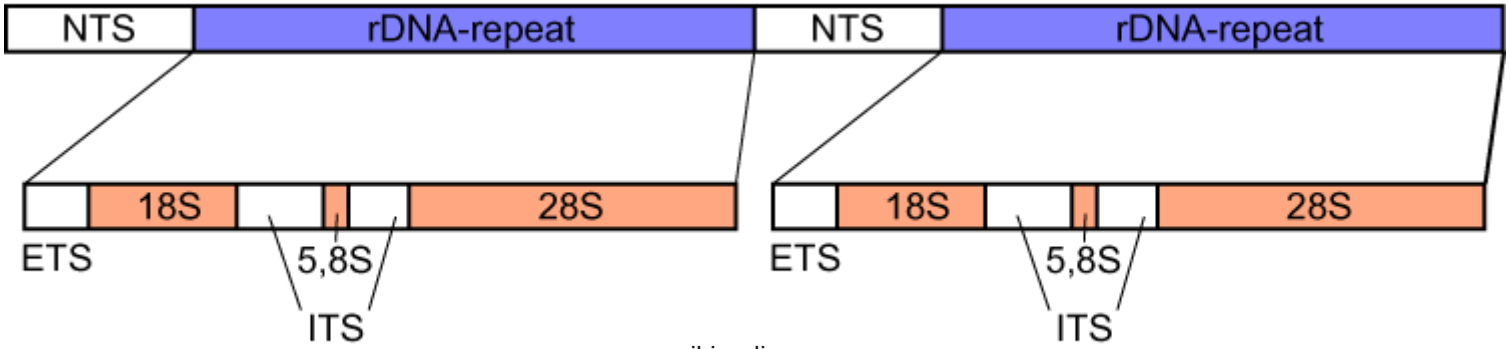
Browse links below to download versions of the Greengenes 16S rRNA gene database or experimental datasets created with the PhyloChip 16S rRNA microarray. Beware that these publicly available versions of the Greengenes database utilize taxonomic terms proposed from phylogenetic methods applied **years 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 Sp

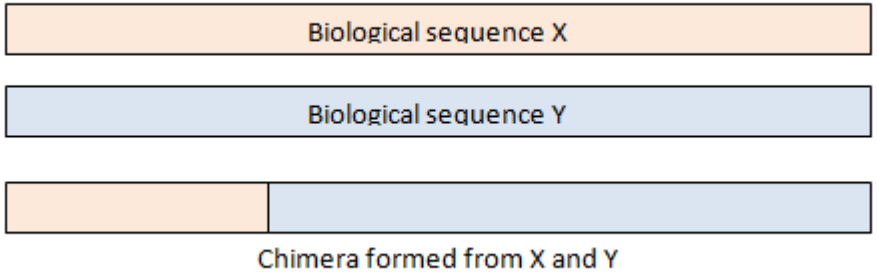
Internal transcribed spacer (ITS)

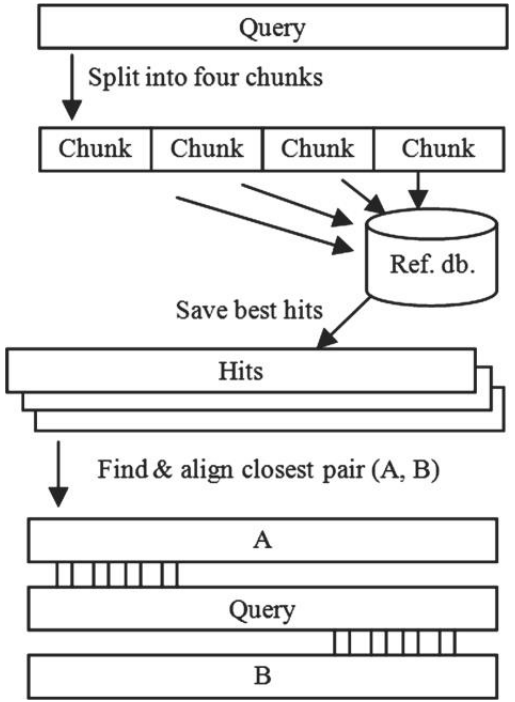


wikipedia.com

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

Chimeric reads in amplicon sequencing

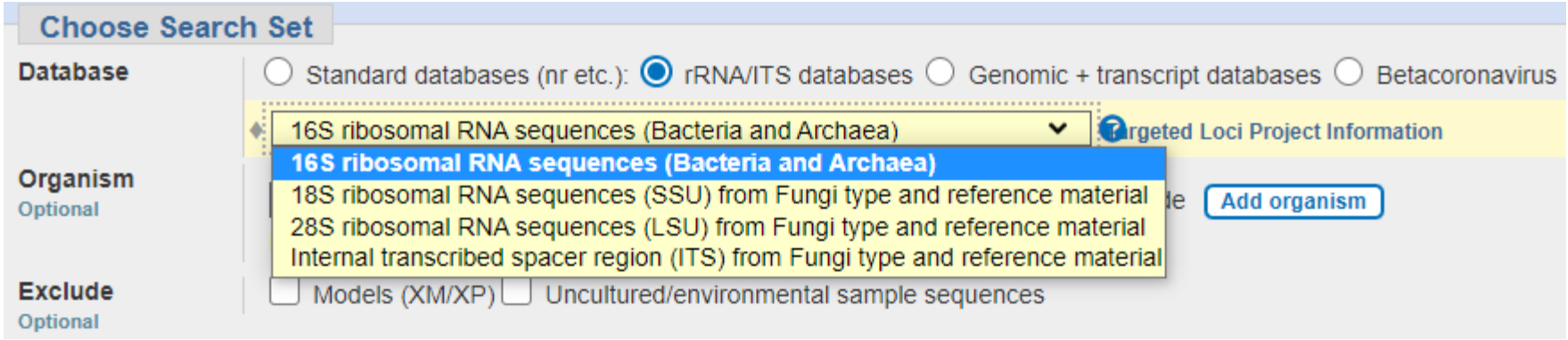




- Produced during PCR amplification
- Detected by alignment different portion of the reads to rRNA databases
- Mismatch of hits = chimeric reads

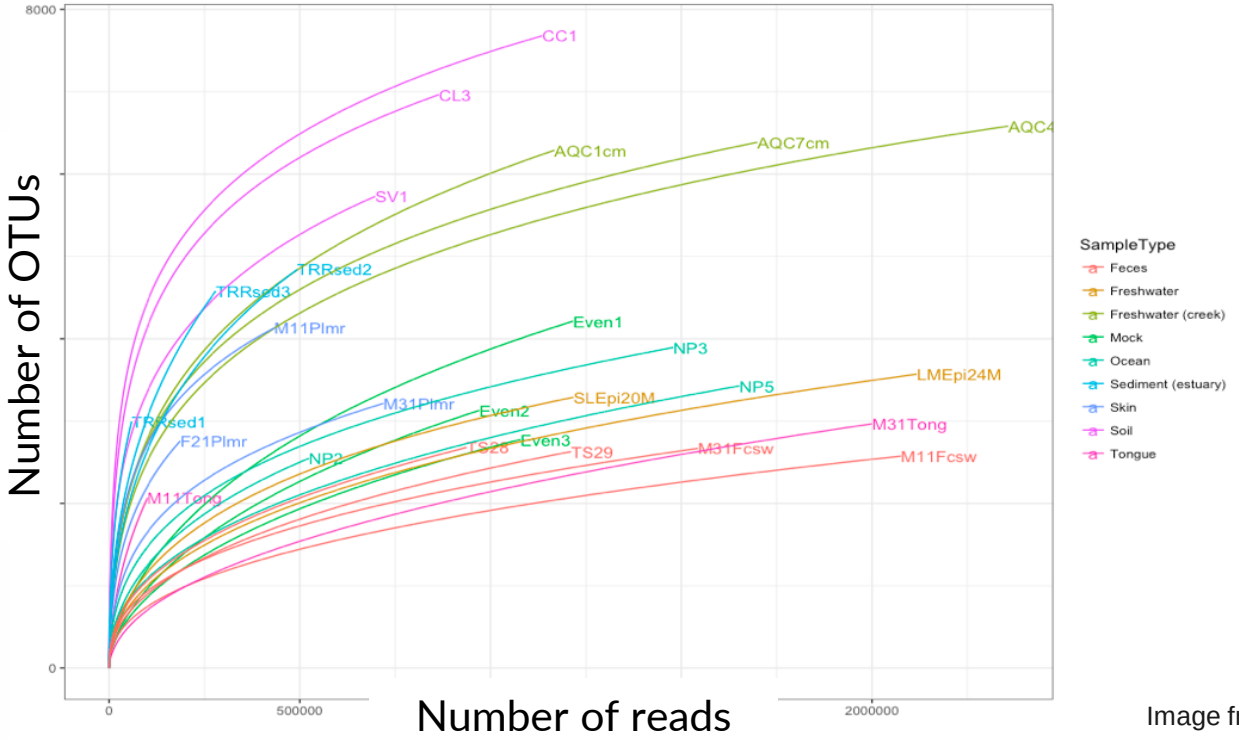
Edgar *et al.* Bioinformatics 27:2194-2200 (2011)

rRNA BLAST



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

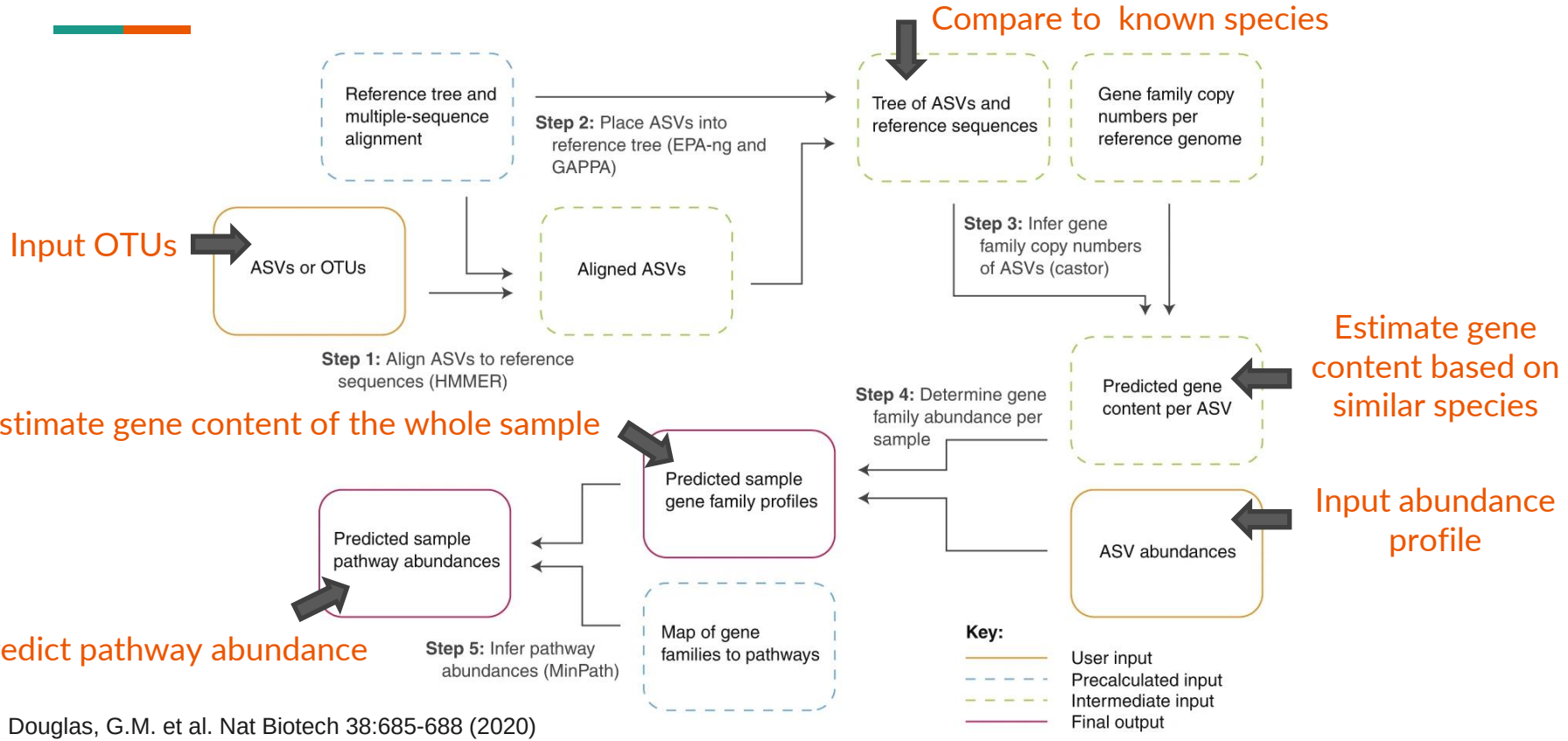
Rarefaction curve for evaluating depth of sequencing



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

Image from PhyloSeq GitHub

From taxonomy to functional pathway



Douglas, G.M. et al. Nat Biotech 38:685-688 (2020)

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

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

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 $H_{min} = 0$

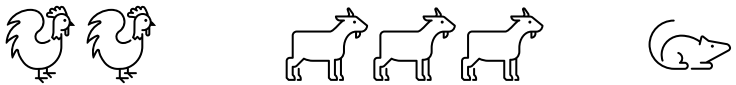
Behavior of Simpson's index

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

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

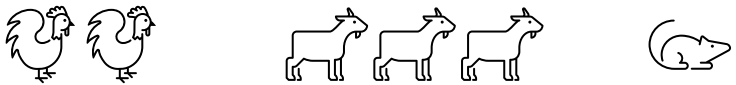
Sample 1 

Sample 2 

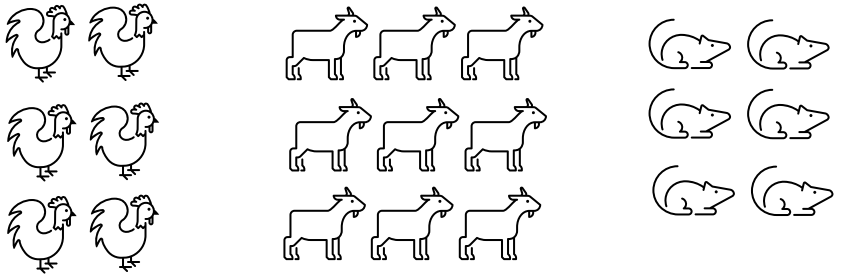
- S_1 = # of individuals in sample 1 = 6
- S_2 = # of individuals in sample 2 = 5
- Overlap = 1 + 2 + 1 = 4
- Bray-Curtis dissimilarity = $1 - \frac{2 \times \text{Overlap}}{S_1 + S_2} = 1 - \frac{8}{11} = \frac{3}{11}$

Impact of sequencing depth

Sample 1

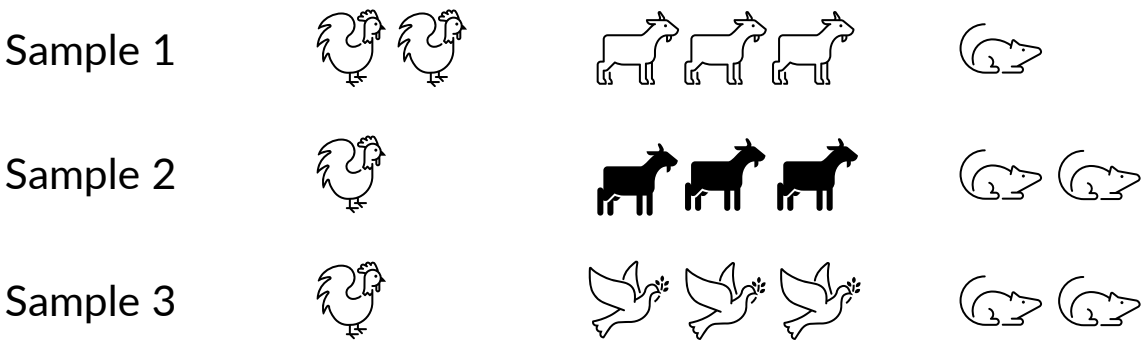


Sample 2



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

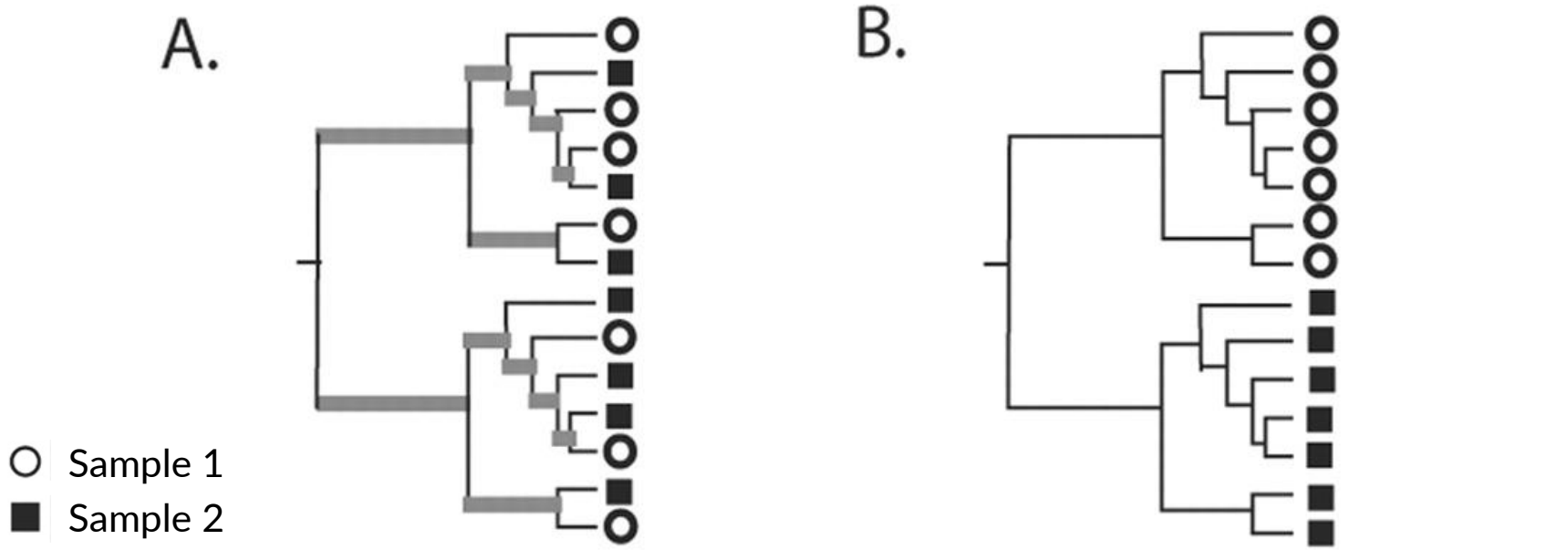
Impact of taxonomic similarity



- Bray-Curtis does not take into account taxonomic similarity

UniFrac distance

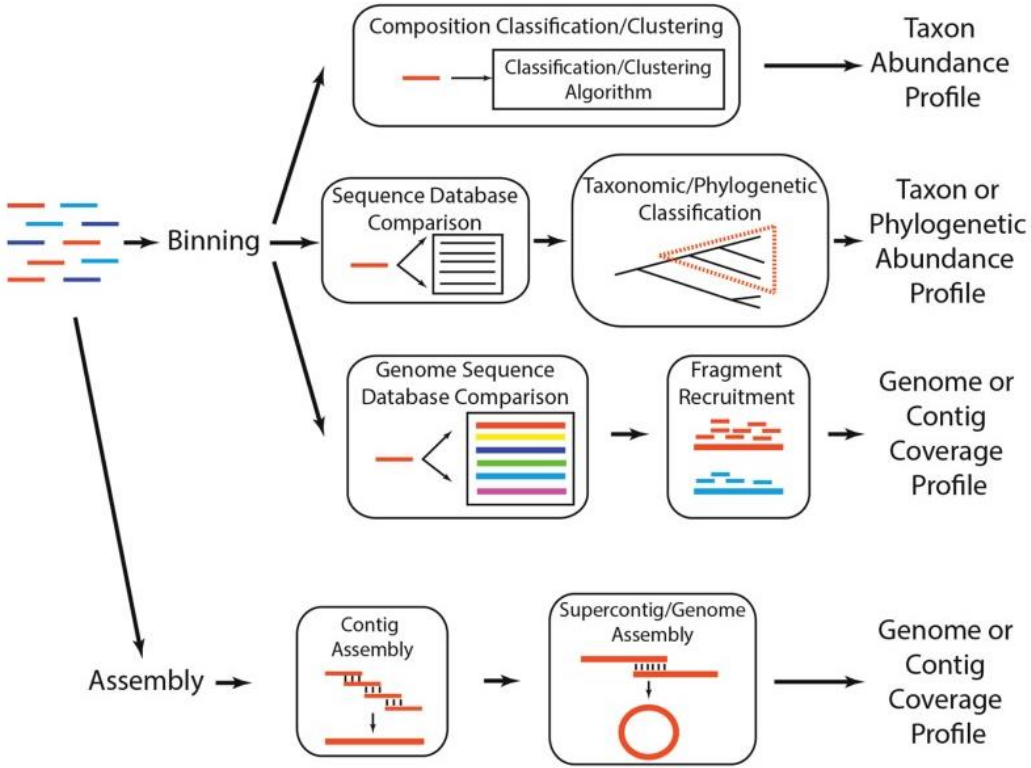
Lozupone, C. and Knight, R. Applied and Environmental Microbiology 71 (2005)



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

Shotgun metagenomics

Key steps in shotgun metagenomics



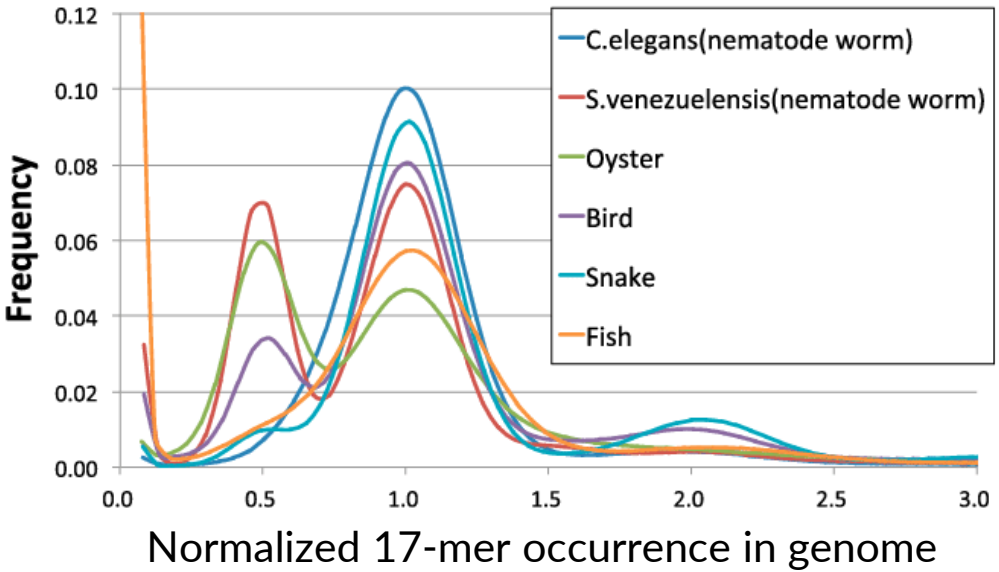
- Dealing with contamination
  - Host DNA
- Binning = grouping DNA/RNA from the same host organisms together
- Direct assembly is possible for abundant species

Sharpton *et al*. Front. Plant Sci. 5:209 (2014)

DNA *k*-mer can serve as species signature

***k*-mers for GTAGAGCTGT**

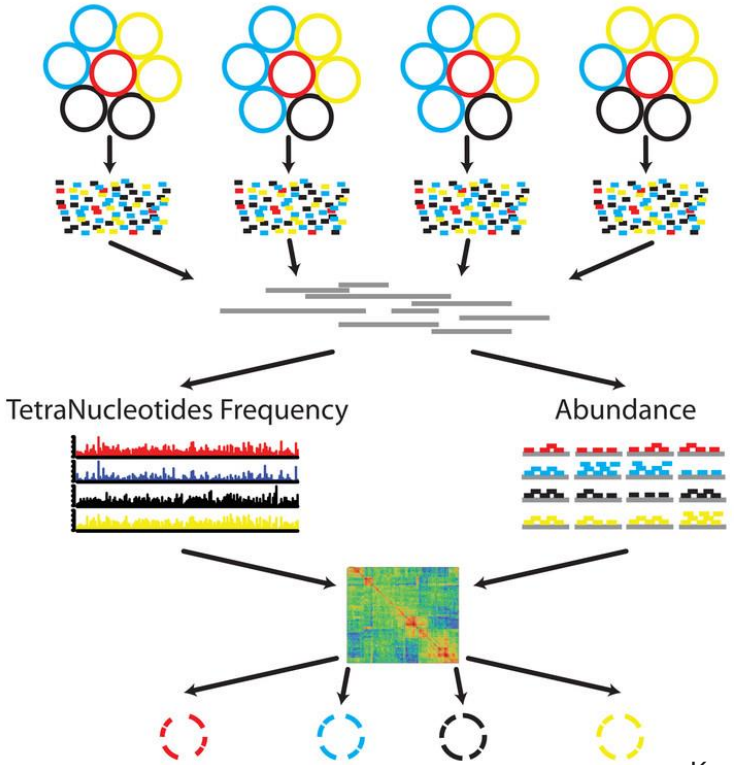
k	*k*-mers
1	G, T, A, G, A, G, C, T, G, T
2	GT, TA, AG, GA, AG, GC, CT, TG, GT
3	GTA, TAG, AGA, GAG, AGC, GCT, CTG, TGT
4	GTAG, TAGA, AGAG, GAGC, AGCT, GCTG, CTGT
5	GTAGA, TAGAG, AGAGC, GAGCT, AGCTG, GCTGT
6	GTAGAG, TAGAGC, AGAGCT, GAGCTG, AGCTGT
7	GTAGAGC, TAGAGCT, AGAGCTG, GAGCTGT
8	GTAGAGCT, TAGAGCTG,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



https://en.wikipedia.org/wiki/K-mer

Kajitani, R. *et al*. Genome Research 24:1384-1395 (2014)

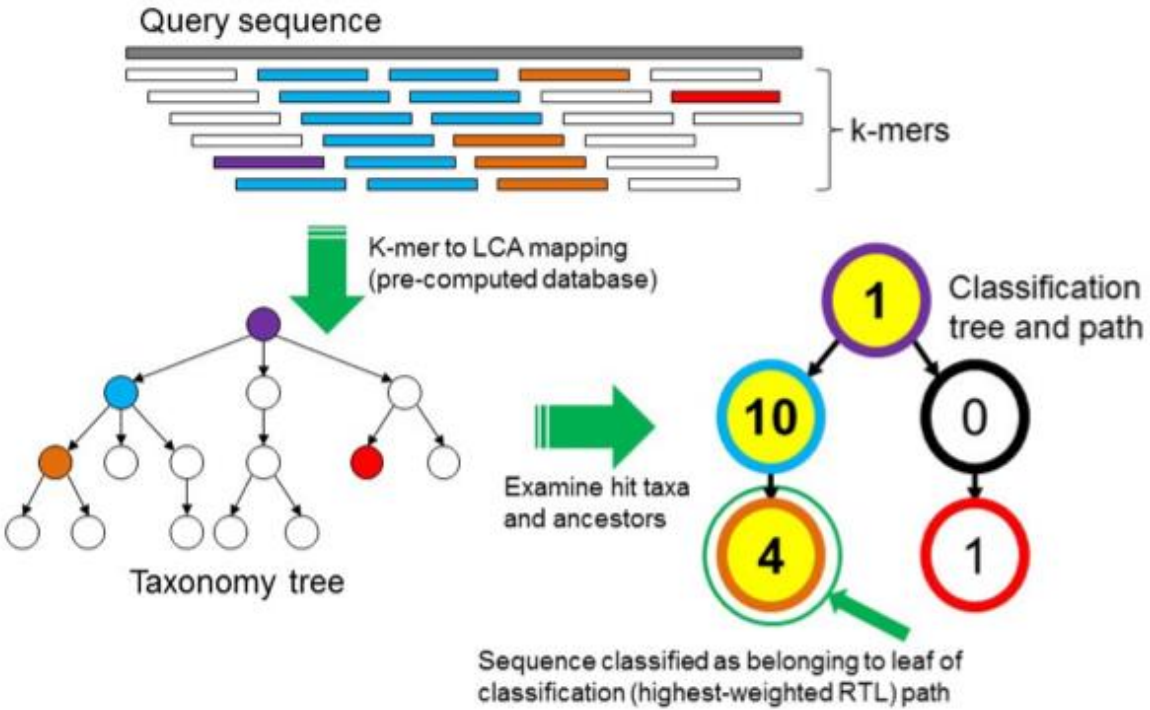
Read binning strategies



- 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)

Kang *et al.* PeerJ (2015)

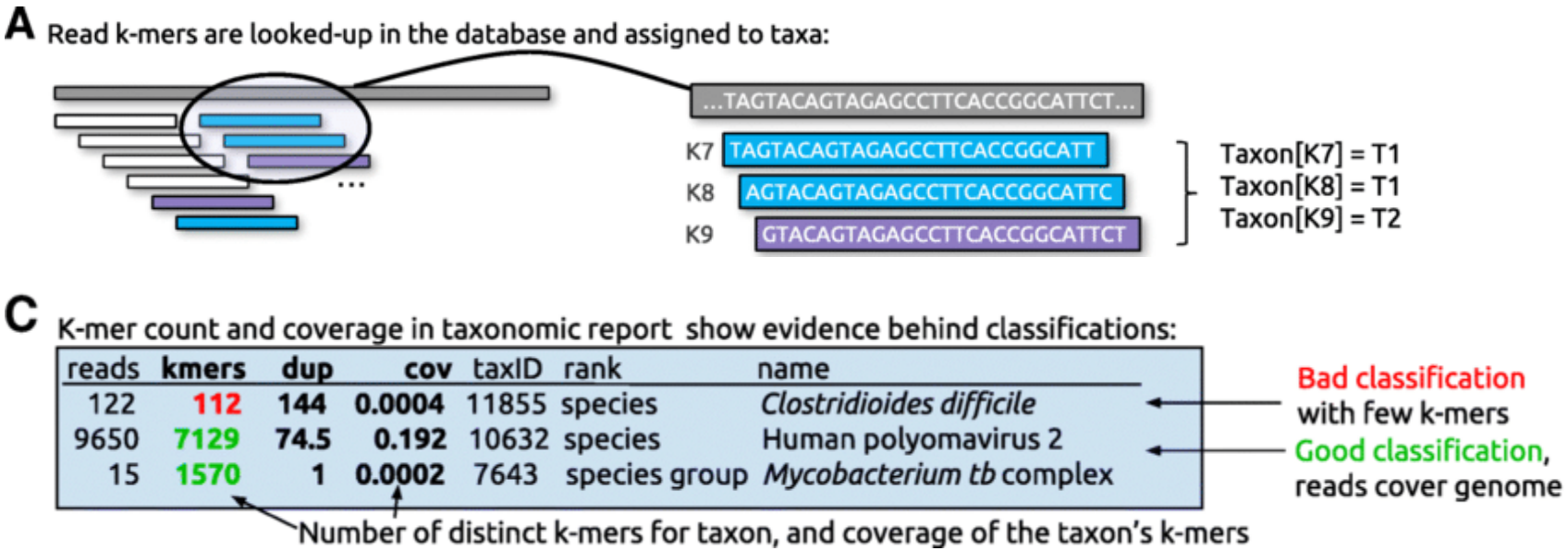
k-mer matching to predict taxonomy



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

Wood and Salzberg, Genome Biology (2014)

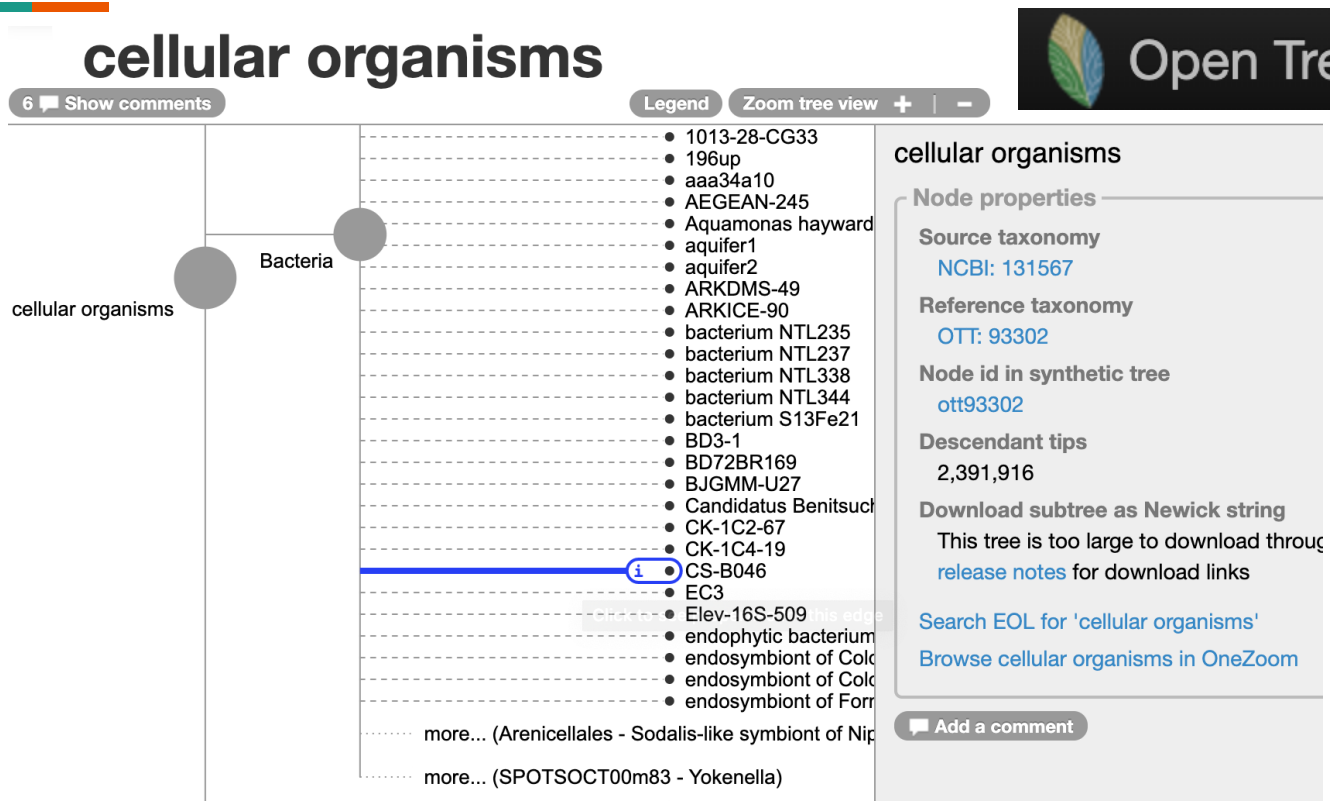
k-mer matching to predict taxonomy



Breitwieser, F. et al. Genome Biology 19:198 (2018)

Resources for metagenomics

Pre-built phylogenetics



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 | Downloac

Choose Columns | Page 1 of 5,327 | 50

#	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/Melair	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

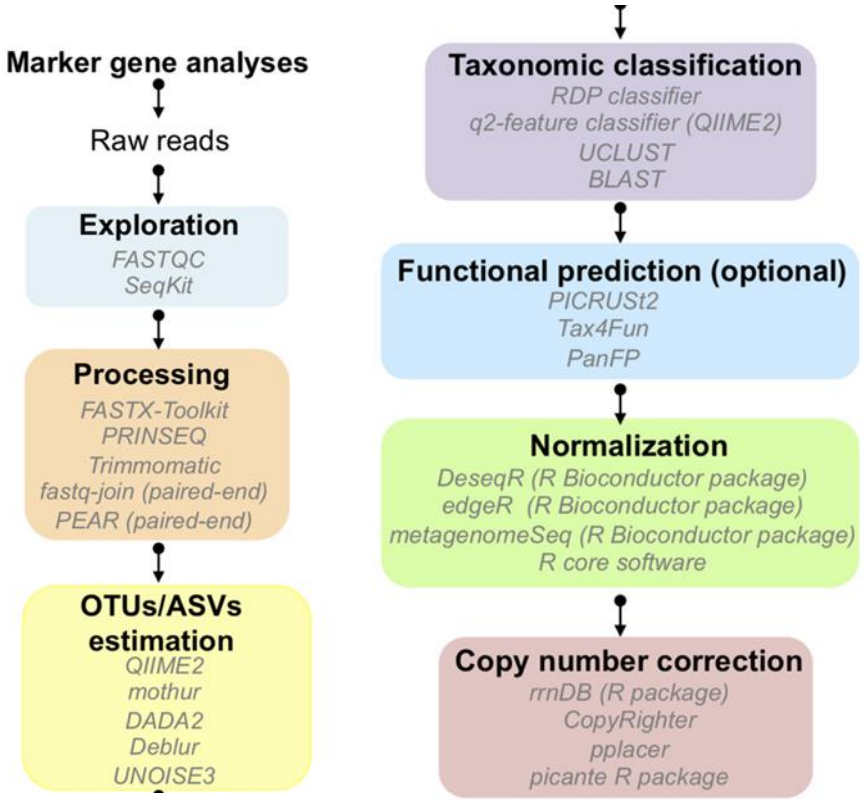
Pérez-Cobas *et al.*, *Microbial Genomics* 2020;6

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)



k-mer, machine learning, or alignment

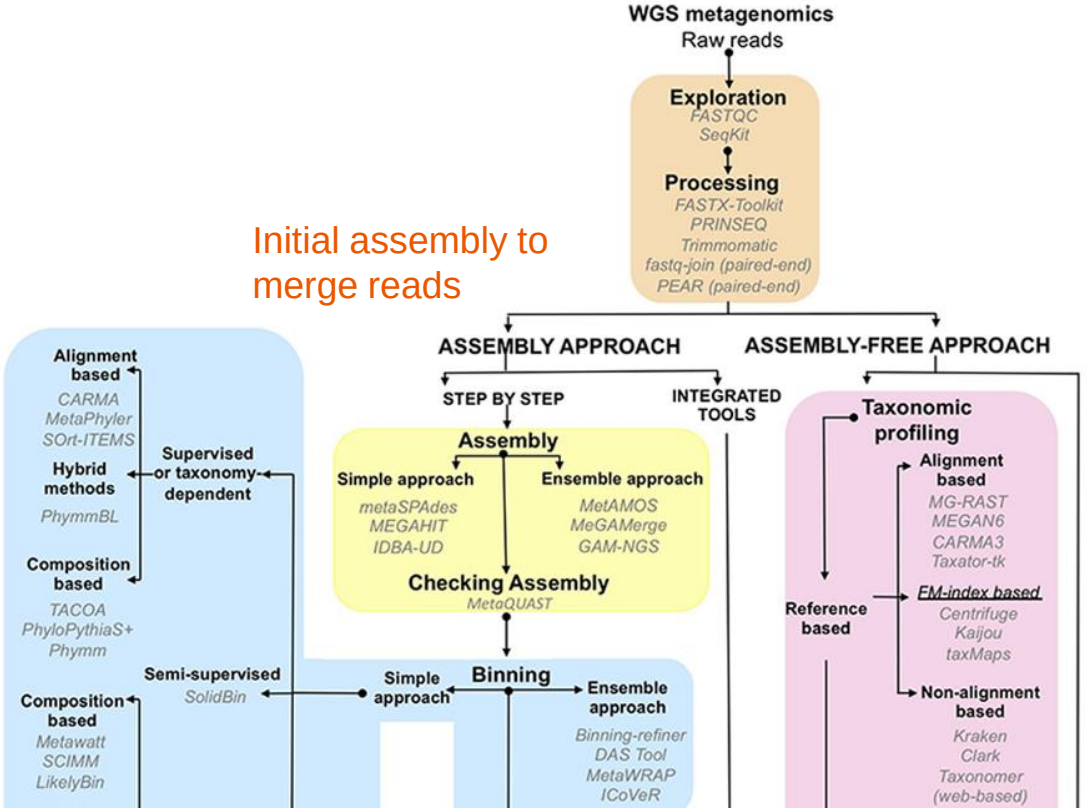
Remove noises, reads with high error, and chimera

Shotgun analysis pipeline

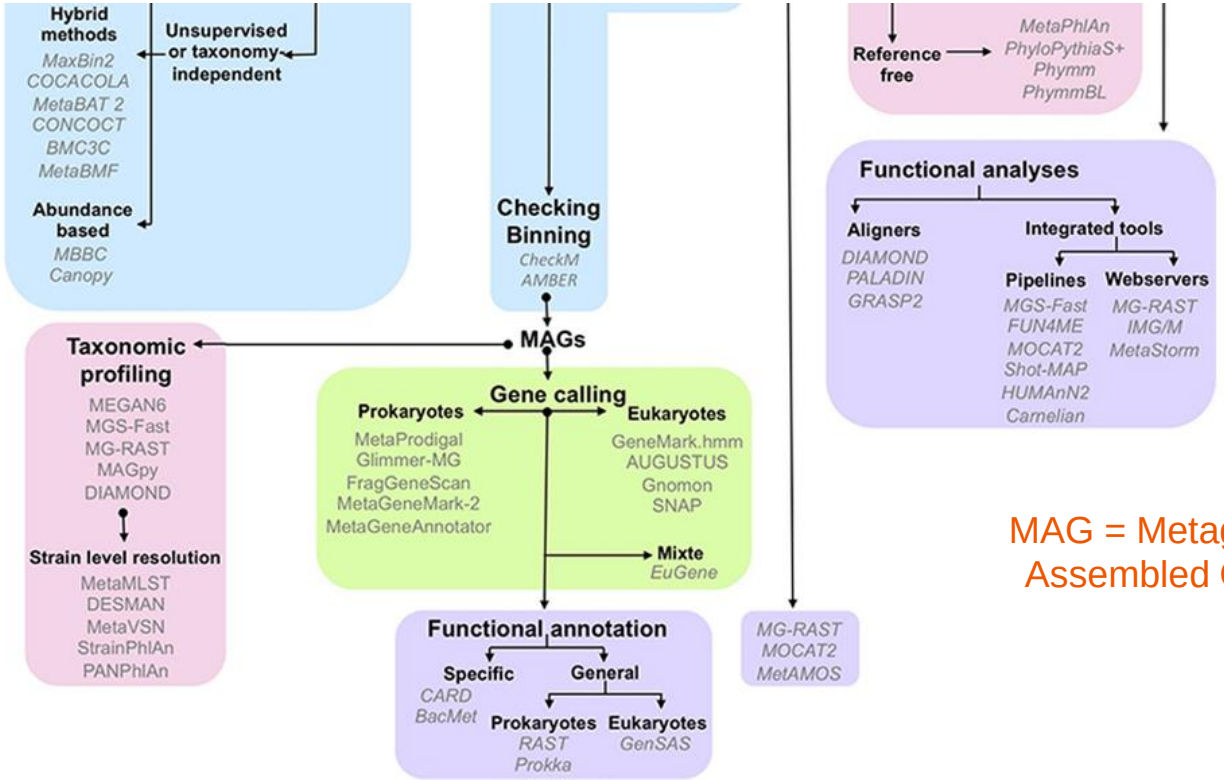
Initial assembly to merge reads

Direct taxonomic annotation

Read binning



Shotgun analysis pipeline

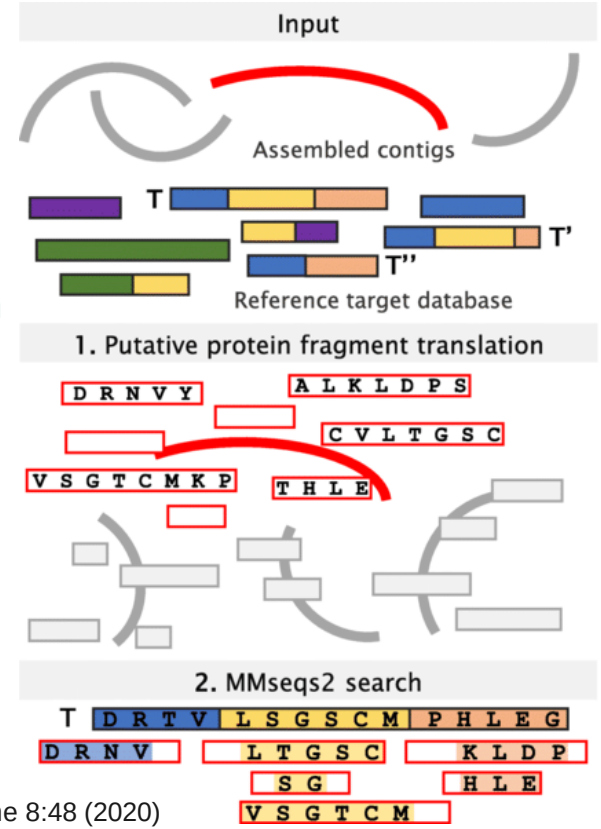


MAG = Metagenomics Assembled Genome

Gene prediction

Six-frame translation

Align to known proteins



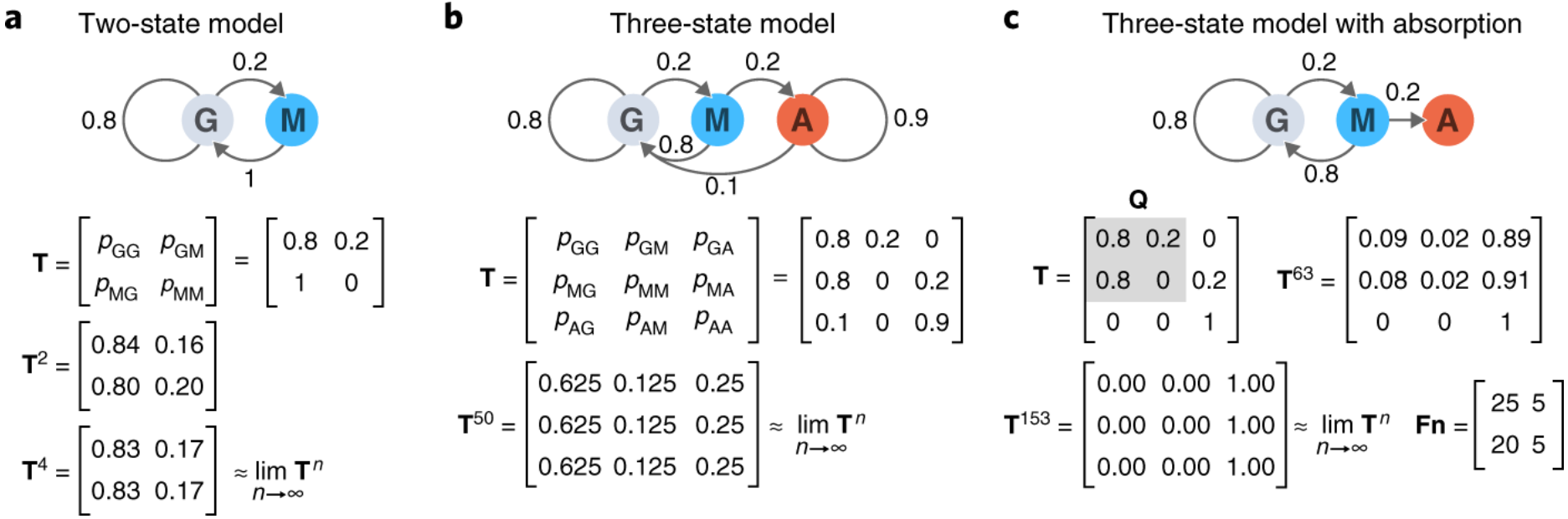
Order hits

Remove duplicated genes

Remove contradicting predictions

Karin, E.L. et al. Microbiome 8:48 (2020)

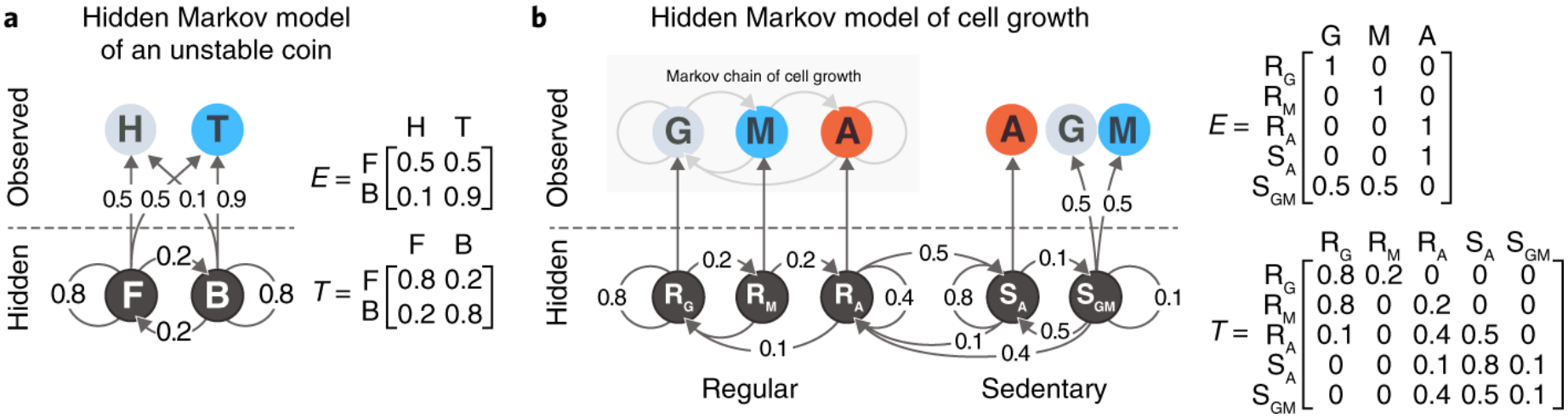
Markov Model



Grewal, J. et al. Nature Methods, 16:663-664 (2019)

- State with state-transition probabilities

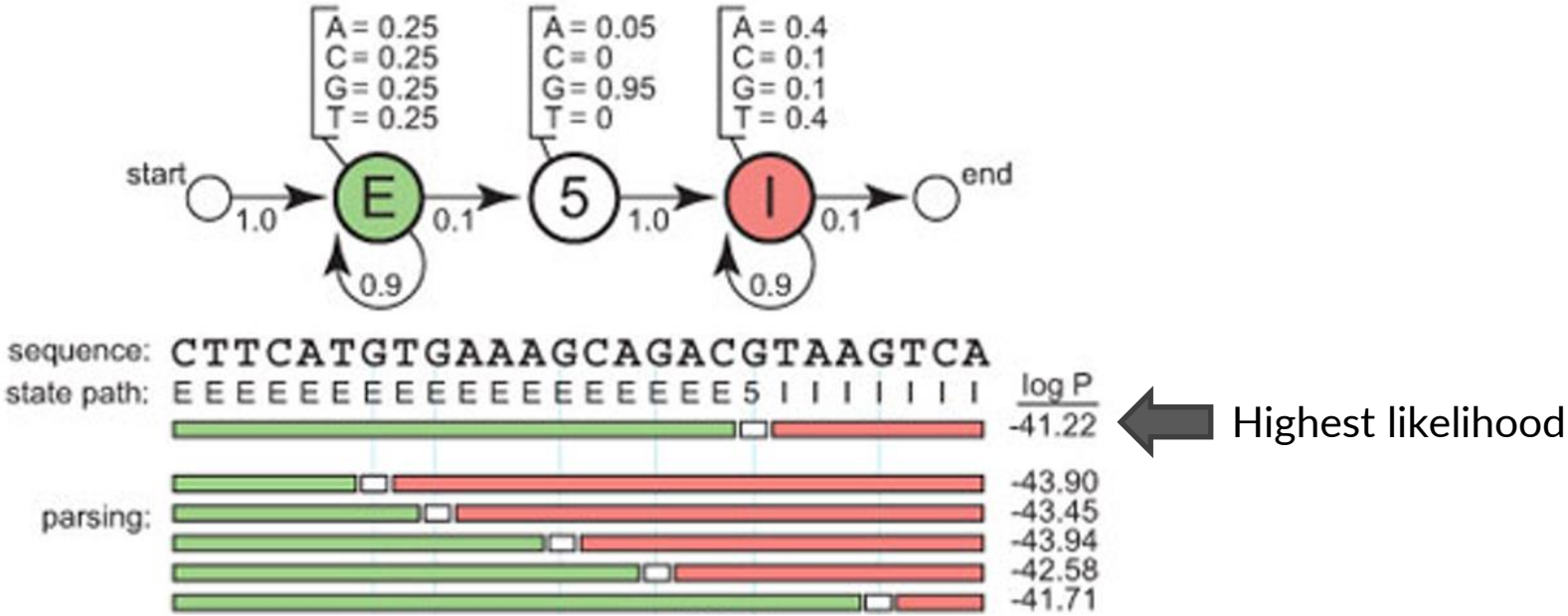
Hidden Markov Model



Grewal, J. et al. Nature Methods, 16:795-796 (2019)

- (Underlying) states with state-transition probabilities
- Each state emits outputs with certain probabilities
- We only observe the outputs → Infer states

HMM for gene prediction

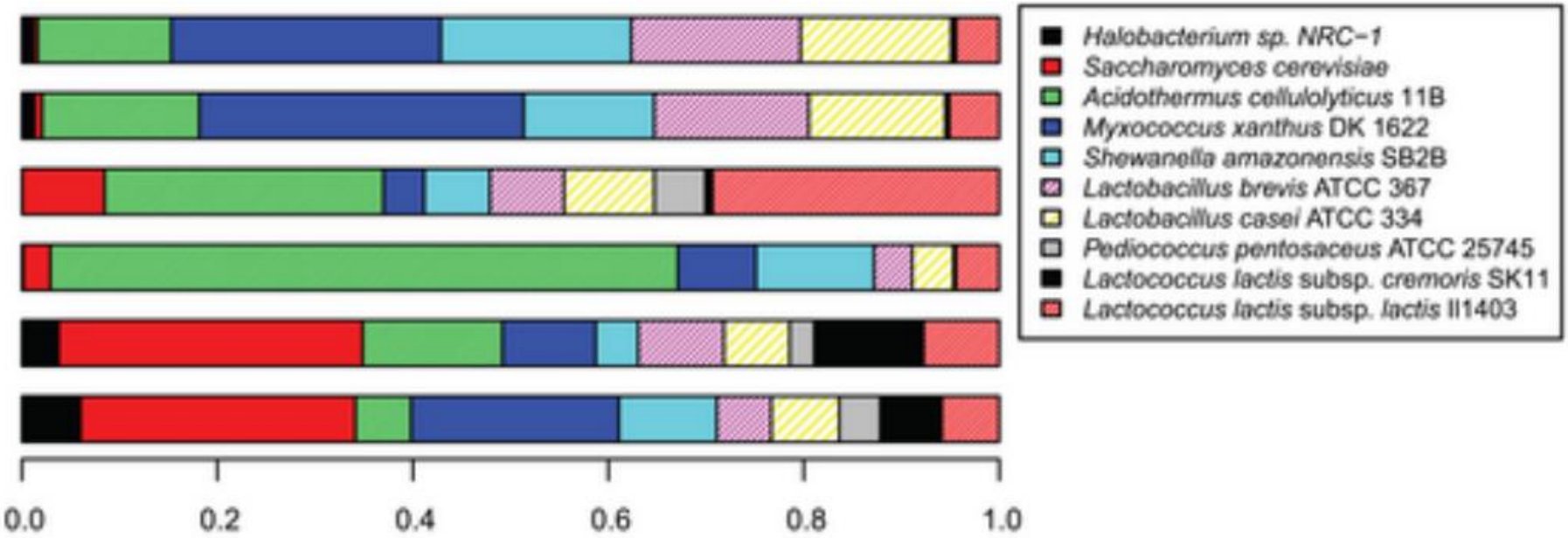


- States = exon, intron, start, stop
- Outputs = nucleotides (codons)

Brandon Malone, HMM and Gene Prediction. Jan 30, 2014

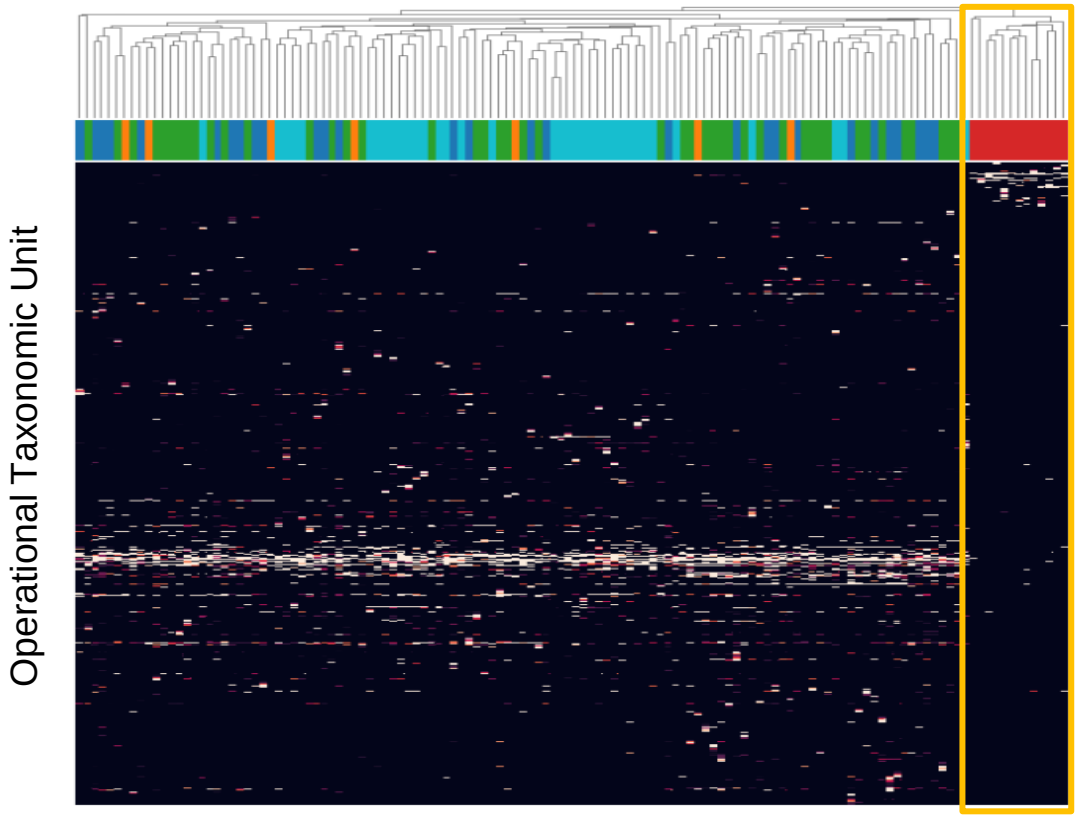
Variability of microbiome analysis

Same sample, different profiles



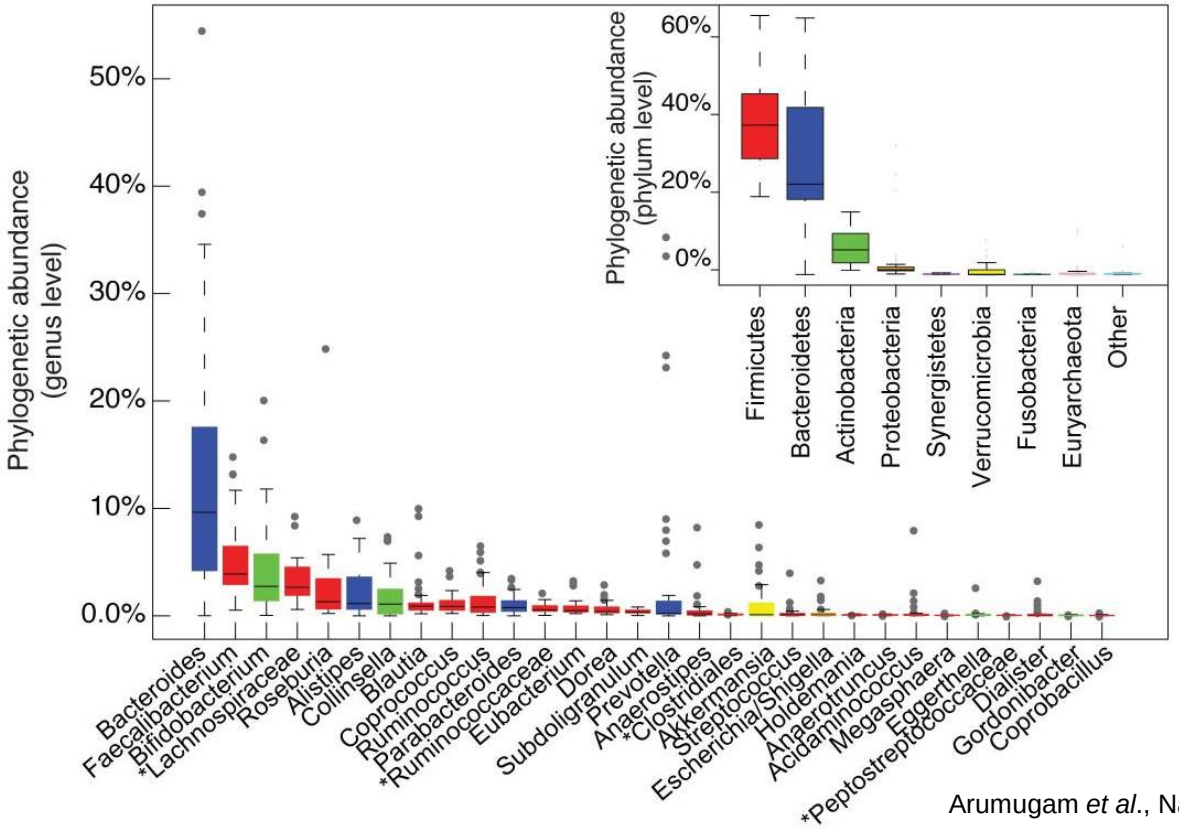
Source: http://www.cbs.dtu.dk/courses/27626

Batch effects



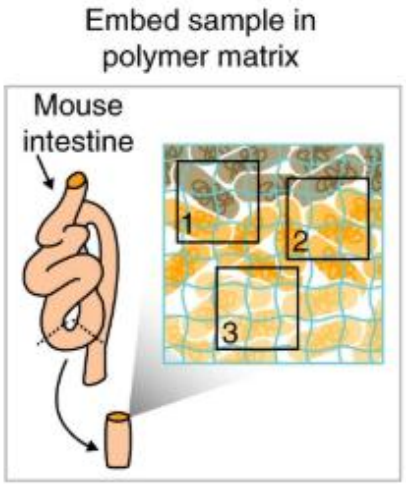
Red = same batch

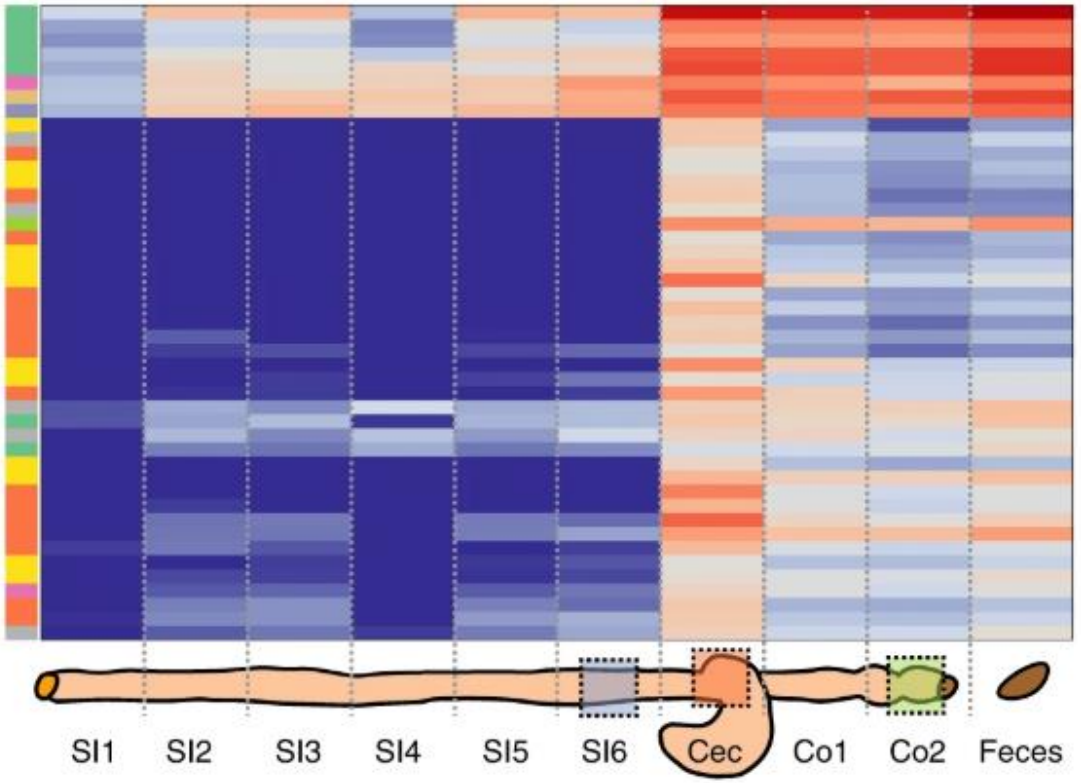
Variability of human gut microbiomes



Arumugam *et al*., Nature 473: 174-80 (2011)

Spatial variability





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

Any question?

- See you on September 14