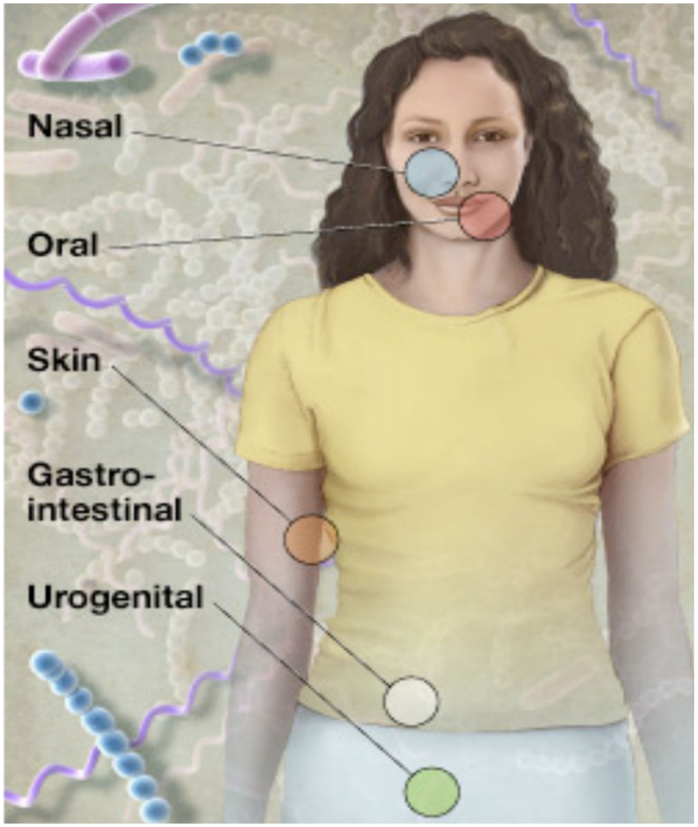
An introduction to metagenomics

Microbes are everywhere

http://npic.orst.edu/envir/soil.html

http://www.rimpro-india.com/articles1/waste-water-treatment-technologies-and-techniques.html



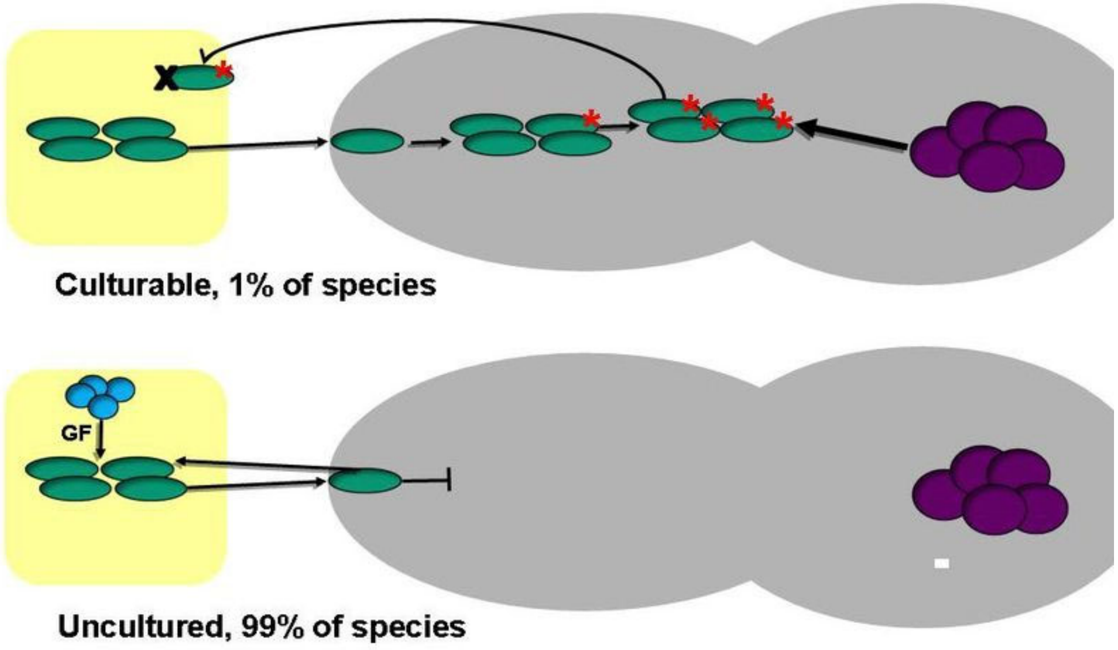
https://www.bcm.edu/departments/molecular-virology-and-microbiology/research/the-human-microbiome-project

http://ocean.nationalgeographic.com/ocean/

https://en.wikipedia.org/wiki/Desert

Culture-independent methods are indispensable

more than 99% of organism genomes in the environment are uncultured



Culturable, 1% of species

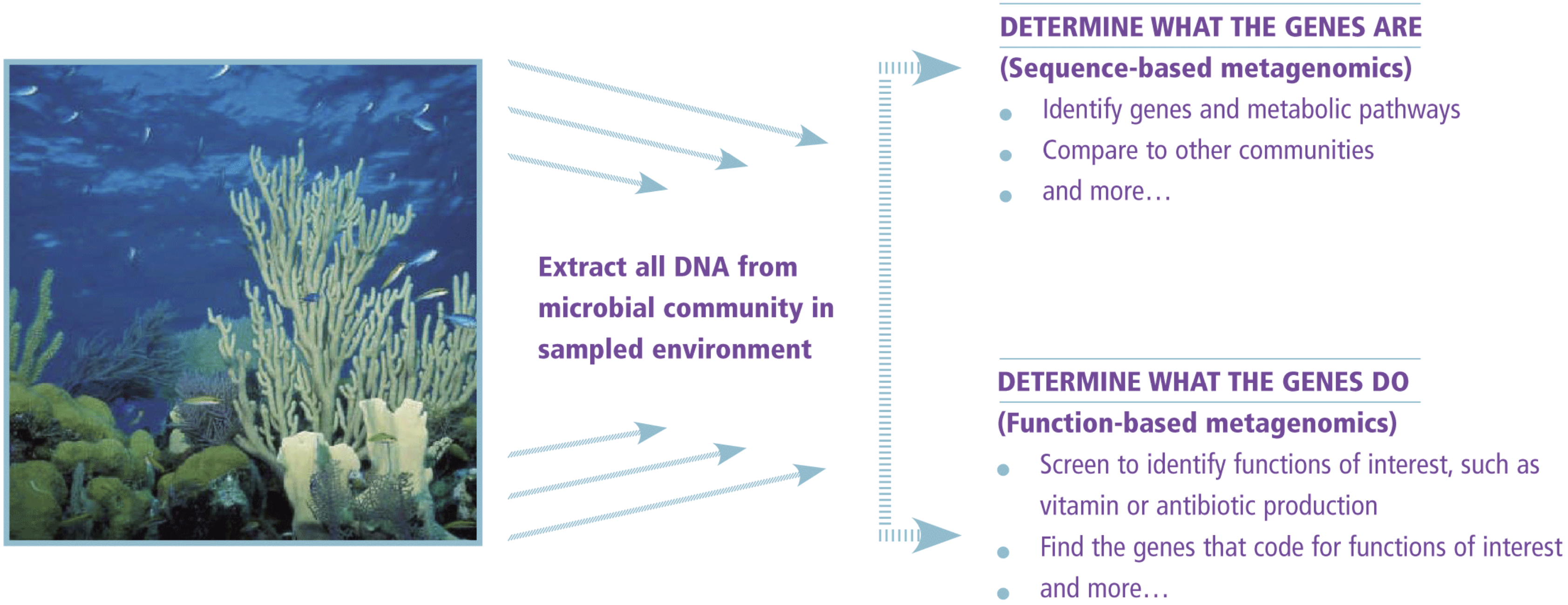
Uncultured, 99% of species

http://schaechter.asmblog.org/schaechter/2010/07/the-uncultured-bacteria.html

Metagenomics, the key to the microbial world

Metagenomics is the study of genetic material recovered directly from environmental samples.

THE METAGENOMICS PROCESS



Extract all DNA from microbial community in sampled environment

DETERMINE WHAT THE GENES ARE

(Sequence-based metagenomics)

- Identify genes and metabolic pathways
- Compare to other communities
- and more…

DETERMINE WHAT THE GENES DO

(Function-based metagenomics)

- Screen to identify functions of interest, such as vitamin or antibiotic production
- Find the genes that code for functions of interest
- and more…

http://blog.openhelix.eu/?p=431

Early history of metagenomics

Norman R Pace propose the idea of cloning DNA directly from environmental samples as early as 1985.

Pace published their study that isolated and cloned bulk DNA from an environmental sample in 1991.

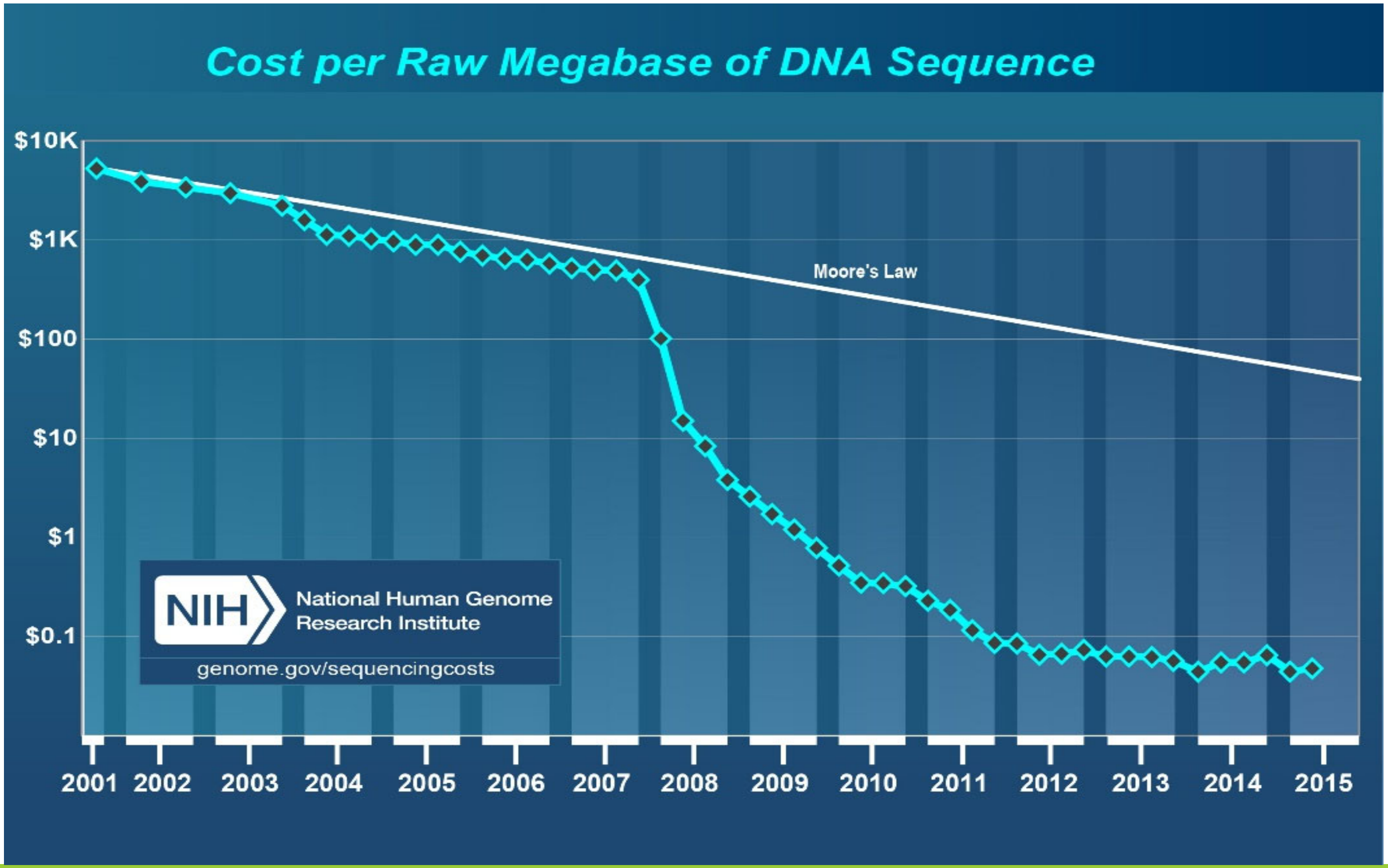
Edward DeLong laid the groundwork for environmental phylogenies based on signature 16S sequences in 1996

In 2002, Mya Breitbart and colleagues used environmental shotgun sequencing to show that 200 liters of seawater contains over 5000 different viruses.

In 2004, Gene Tyson et al acid mine drainage system; Craig Venter et al Global Ocean Sampling Expedition.

In 2005 Stephan C. Schuster et al studied mammoth DNA by 454 sequencer.

Sequencing cost reduced rapidly



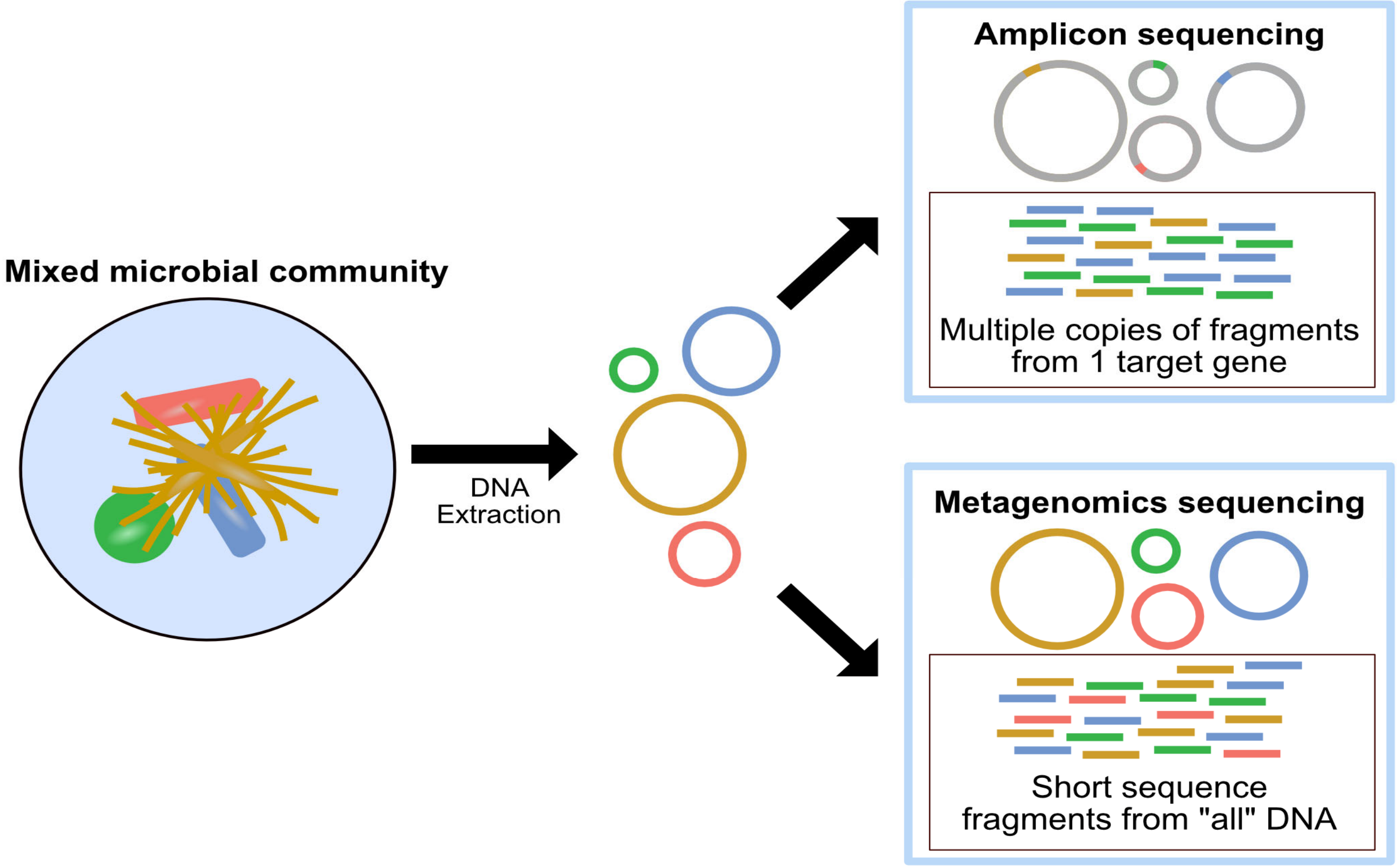
Images from http://www.genome.gov/sequencingcosts/

The almost there future

A human genome sequence

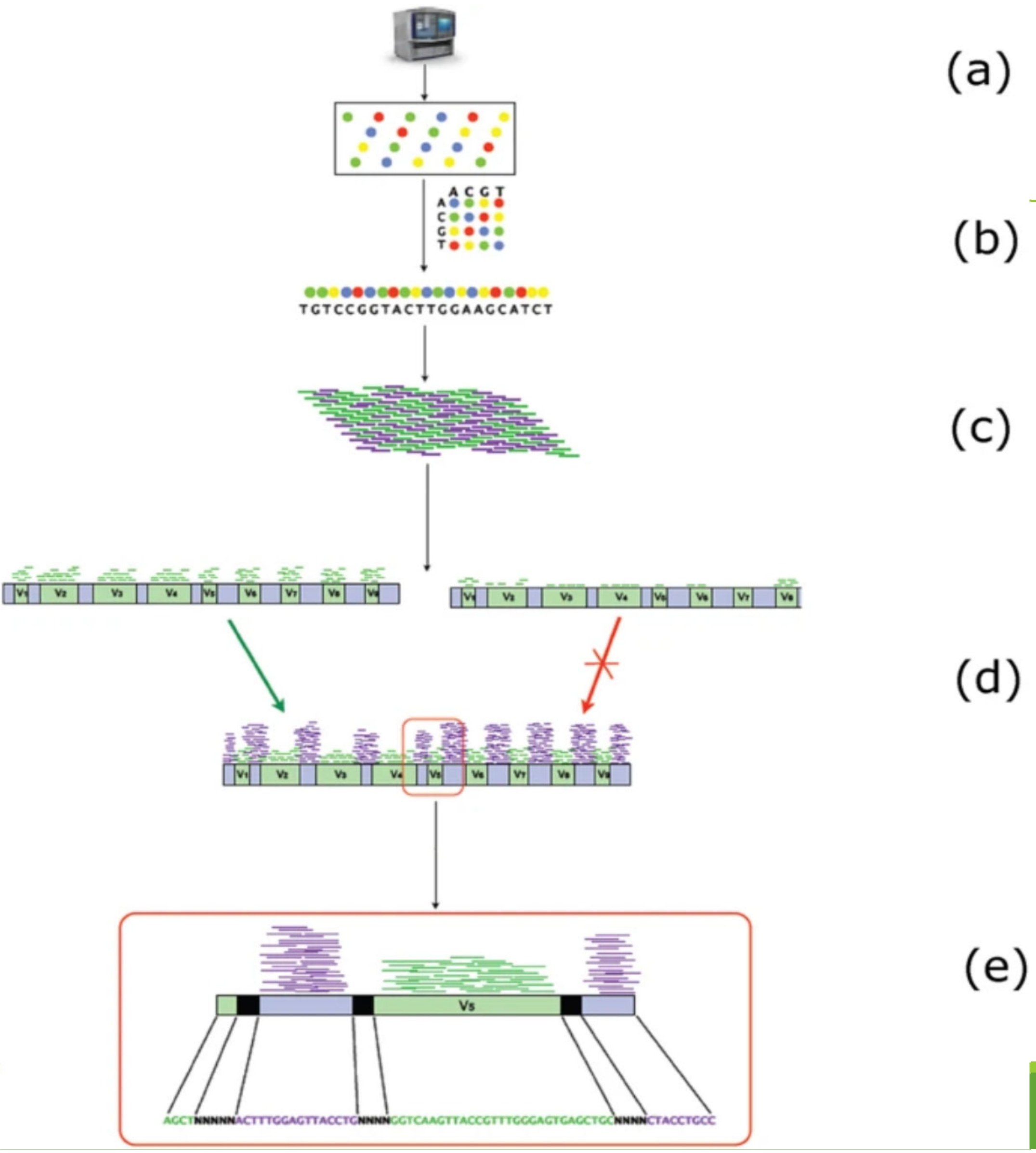
- **2000**
  € 1,000,000,000 in ~10 years
- **2008**
  € 50 - 100,000 in ~4 months
- **2010**
  € 5 - 10,000 in ~2 weeks
- *...2015*
  € 1,000 in ~1 day
- *...2020?*
  € 10 in ~1 hour to minutes

Two types of sequencing



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

Amplicon Sequencing


(a)
(b)
(c)
(d)
(e)
A C G T
A
C
G
T
TGTCCGGTACTTGGAAGCATCT
V1 V2 V3 V4 V5 V6 V7 V8 V9
V5
AGCTNNNNNACTTTGGAGTTACCTGNNNNGGTCAAGTTACCGTTTGGGAGTGAGCTGCNNNNCTACCTGCC

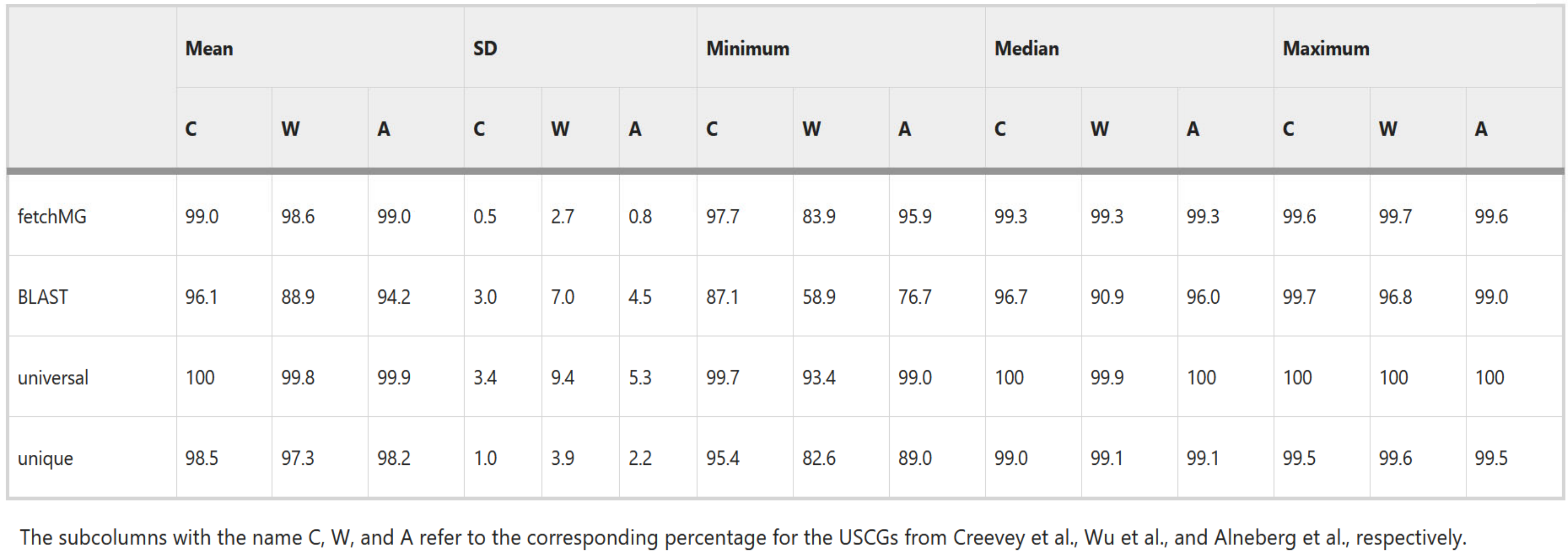
Which marker genes?

Variable regions

https://www.nature.com/articles/srep32165

Neither universal nor unique

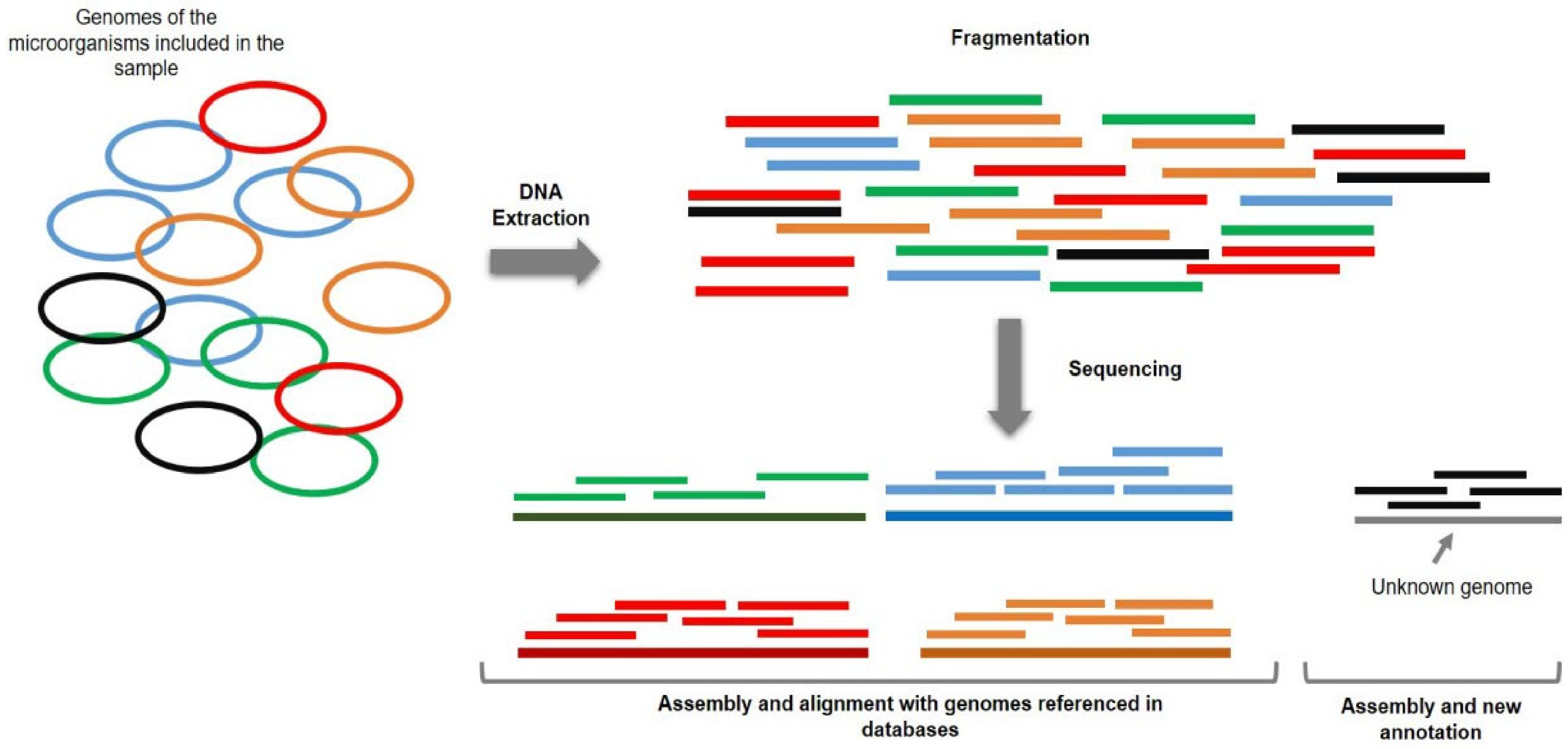
Three groups of widely used marker genes



	Mean			SD			Minimum			Median			Maximum		
	C	W	A	C	W	A	C	W	A	C	W	A	C	W	A
fetchMG	99.0	98.6	99.0	0.5	2.7	0.8	97.7	83.9	95.9	99.3	99.3	99.3	99.6	99.7	99.6
BLAST	96.1	88.9	94.2	3.0	7.0	4.5	87.1	58.9	76.7	96.7	90.9	96.0	99.7	96.8	99.0
universal	100	99.8	99.9	3.4	9.4	5.3	99.7	93.4	99.0	100	99.9	100	100	100	100
unique	98.5	97.3	98.2	1.0	3.9	2.2	95.4	82.6	89.0	99.0	99.1	99.1	99.5	99.6	99.5

The subcolumns with the name C, W, and A refer to the corresponding percentage for the USCGs from Creevey et al., Wu et al., and Alneberg et al., respectively.

Metagenomic shotgun sequencing



Genomes of the microorganisms included in the sample
DNA Extraction
Fragmentation
Sequencing
Unknown genome
Assembly and alignment with genomes referenced in databases
Assembly and new annotation

https://www.france-genomique.org/technological-expertises/metagenomics/shotgun-metagenomics/?lang=en

Shotgun reads in metagenomics

Sanger sequencing ~1000bp long reads, 99.99

454 pyrosequencing typically produces ~400 bp reads

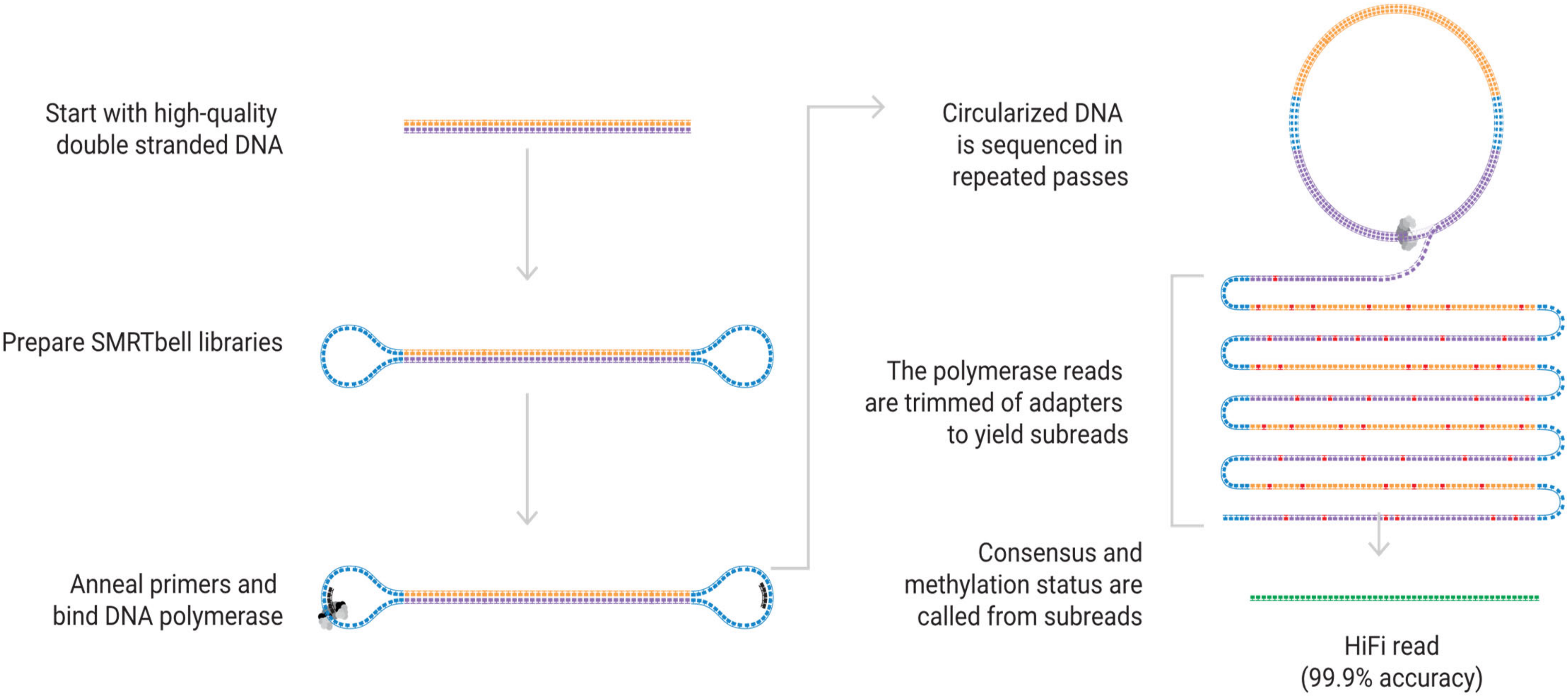
Illumina MiSeq produces 200-600bp reads (depending on whether paired end options are used),

Ion Torrent PGM System ~400 bp reads

Pacific bioscience 10 to 25kb, 99.9% accuracy

Oxford Nanopore minION flow cell R10.4, 10-25kb, 99% accuracy

HiFi reads from PacBio



Start with high-quality double stranded DNA

Prepare SMRTbell libraries

Anneal primers and bind DNA polymerase

Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

Consensus and methylation status are called from subreads

HiFi read (99.9% accuracy)

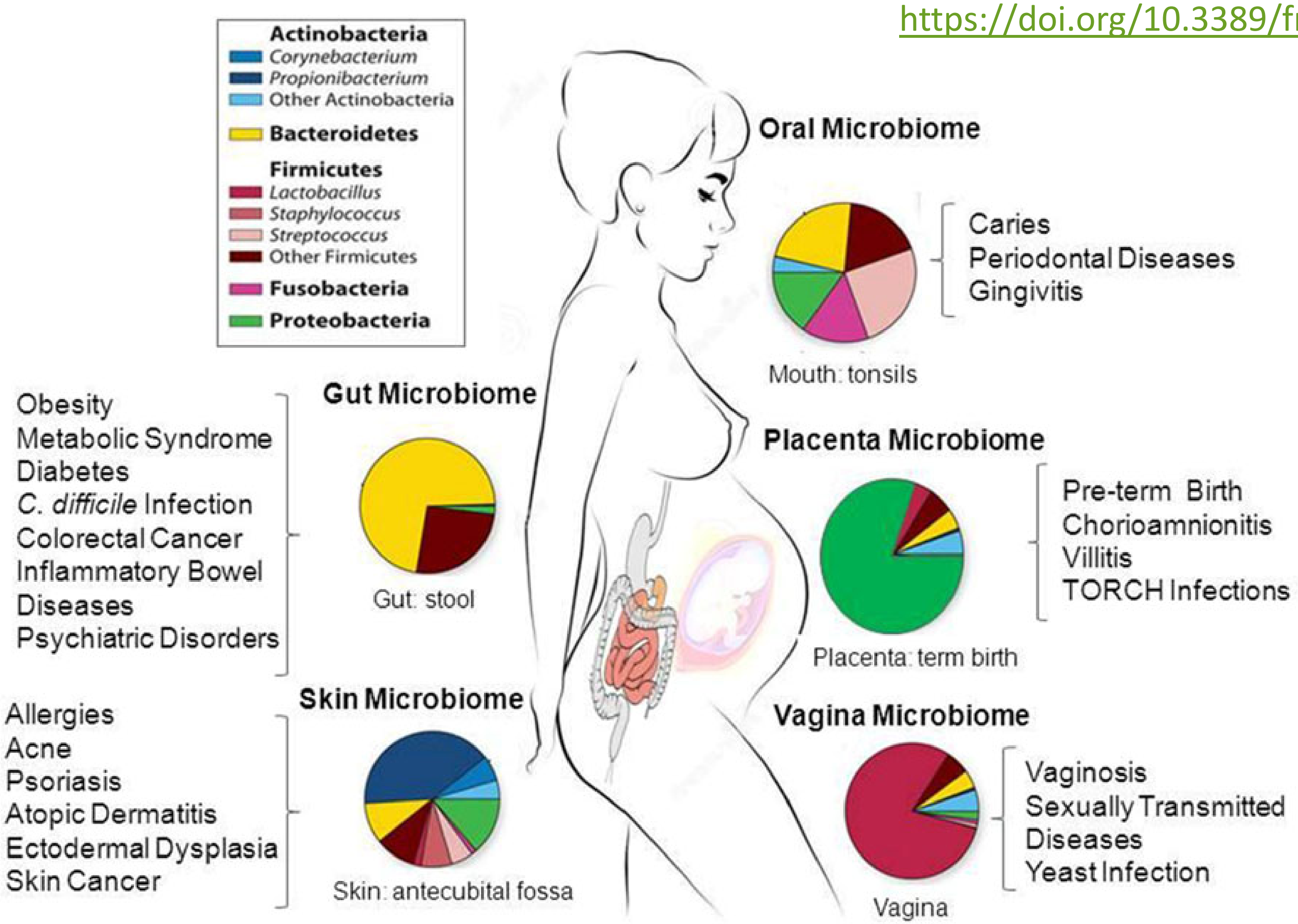
A simple math

How many species can we sequence in a instrumental run of Illumina sequencing (800 Gb)?

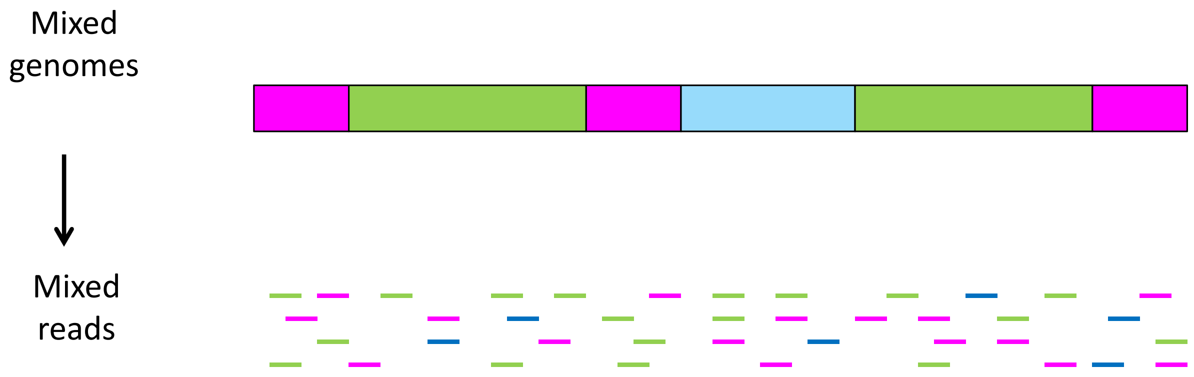
1. Each species is 1Mb long, coverage 1X
2. Each species is 3Mb long, coverage 100X

The Human Microbiome

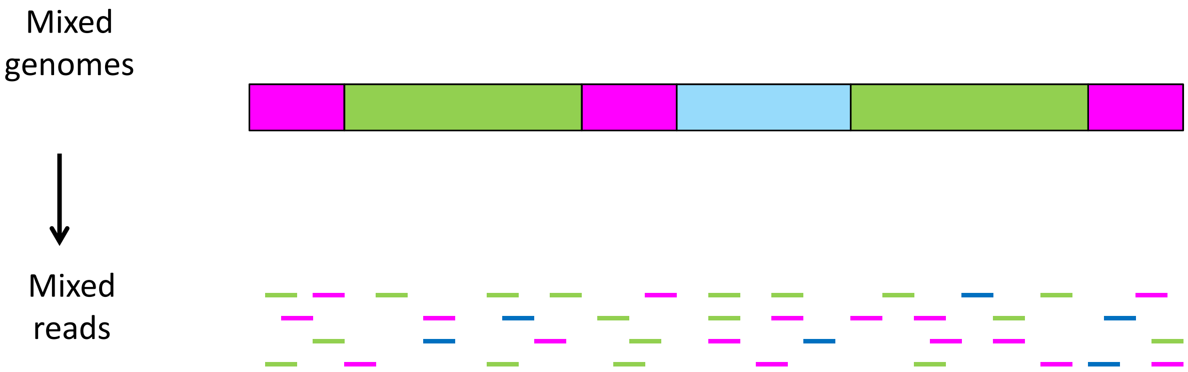
https://doi.org/10.3389/fmicb.2015.01050



What does the data look like



What computational questions can be asked with the data



Mixed genomes

Mixed reads

What computational questions can be asked with the data

Assembly

Read or contig binning

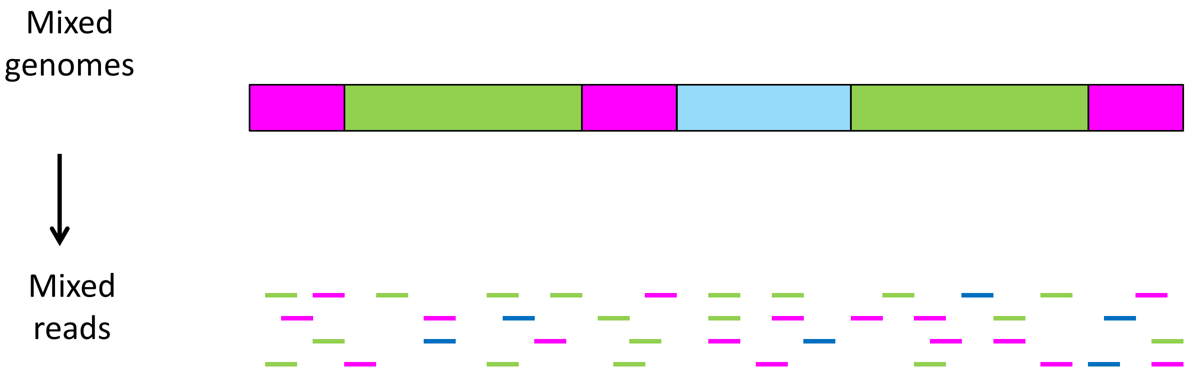
Strain genome reconstruction

Gene prediction

Taxonomic analyses

Functional analyses

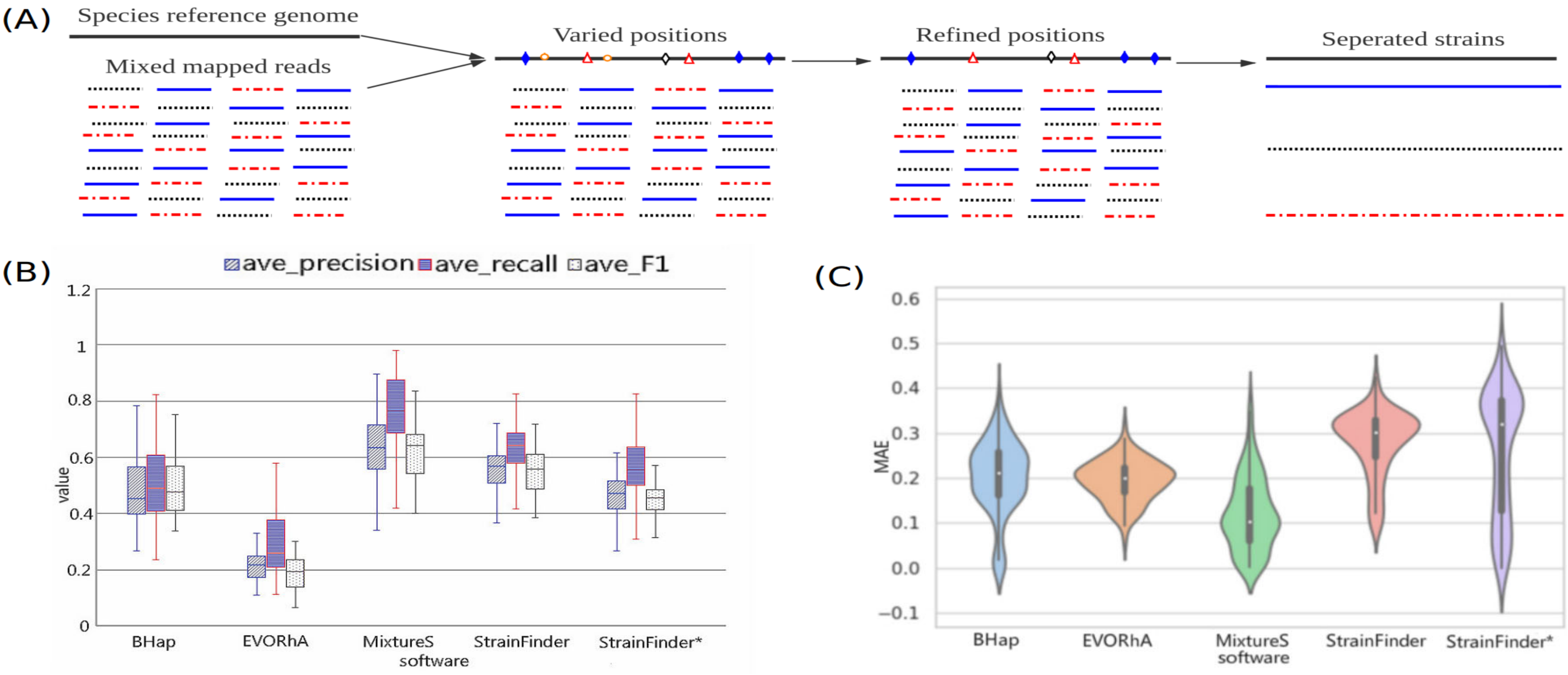
Comparative metagenomics



Mixed genomes

Mixed reads

Strain genome reconstruction



https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa728/5893551

Other questions can be asked ?

Assembly

Read or contig binning

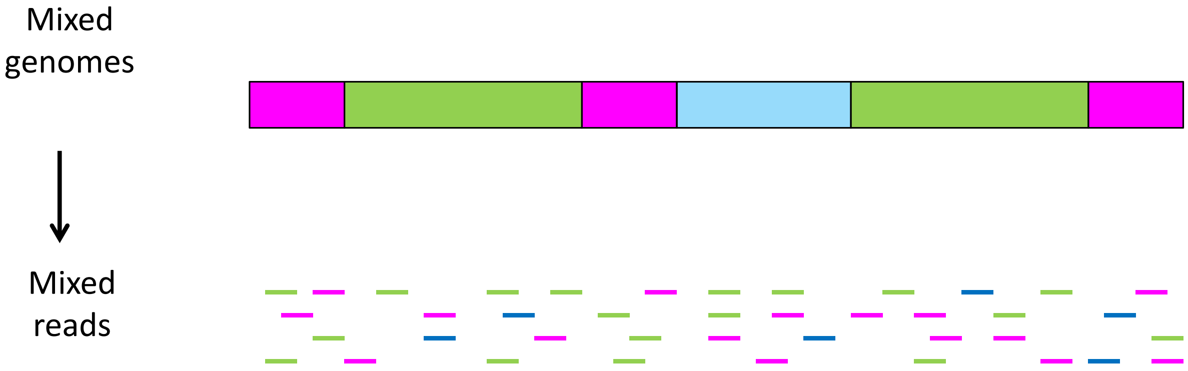
Strain genome reconstruction

Gene prediction

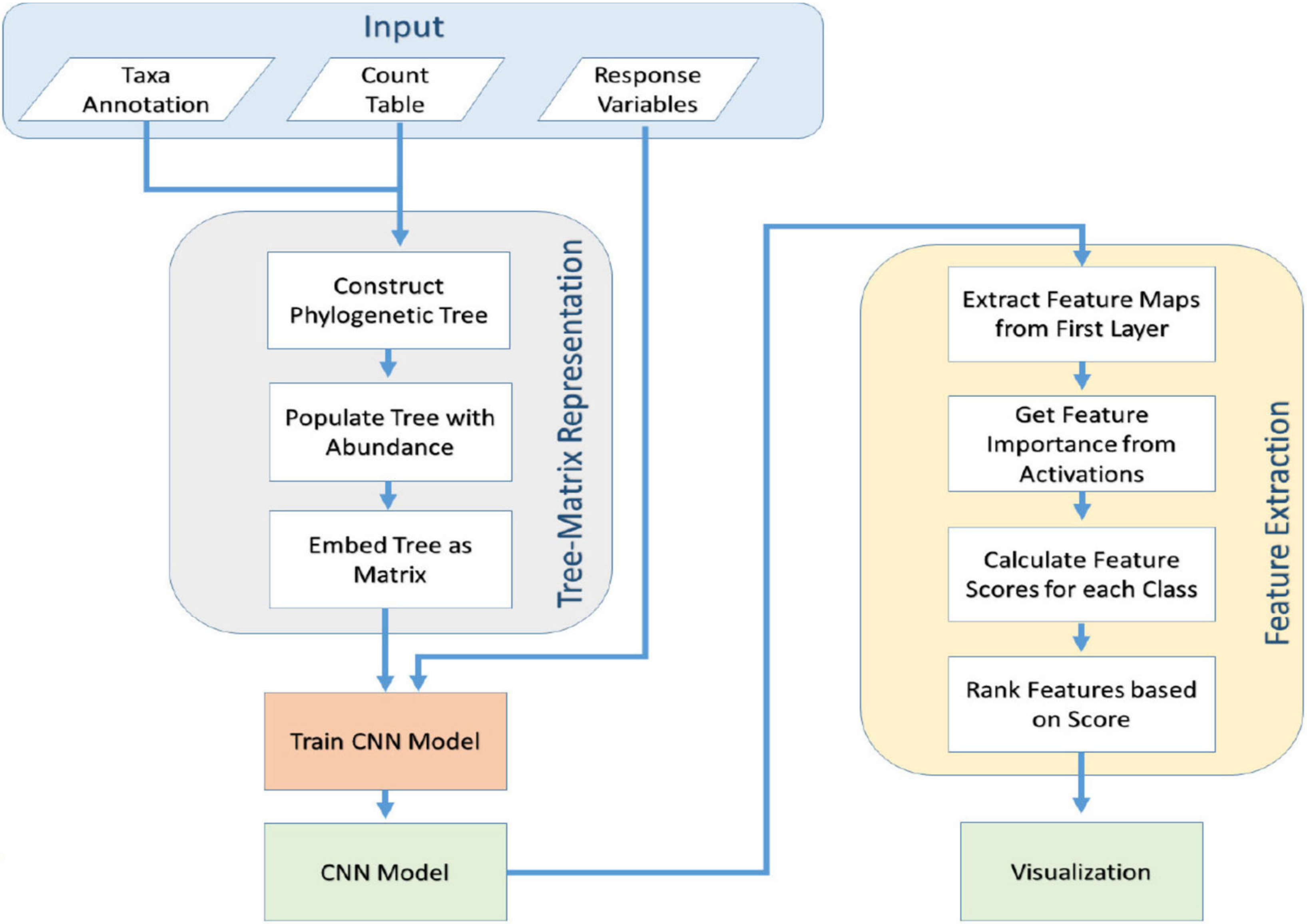
Taxonomic analyses

Functional analyses

Comparative metagenomics



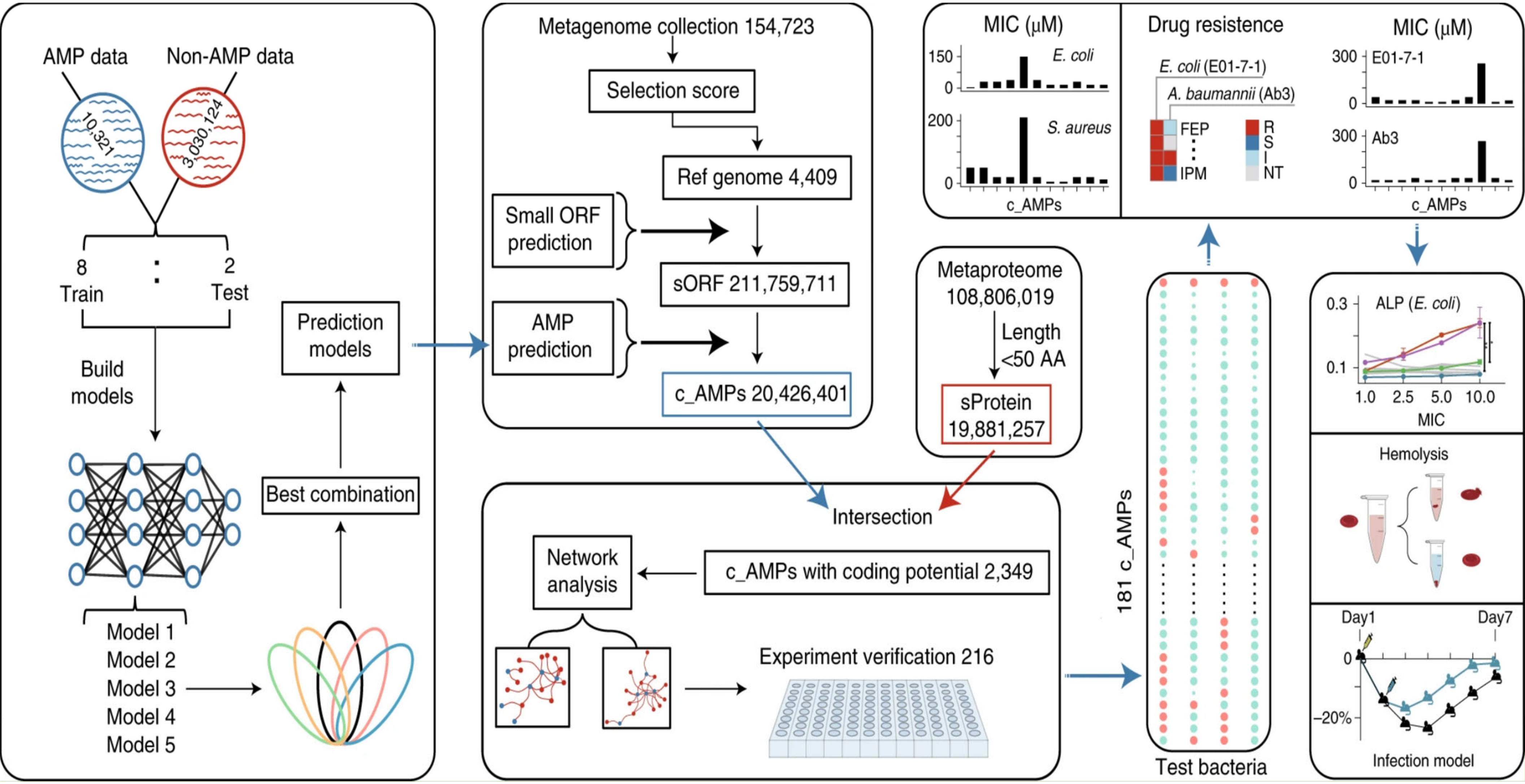
Disease status prediction


Input
Taxa Annotation
Count Table
Response Variables
Tree-Matrix Representation
Construct Phylogenetic Tree
Populate Tree with Abundance
Embed Tree as Matrix
Train CNN Model
CNN Model
Feature Extraction
Extract Feature Maps from First Layer
Get Feature Importance from Activations
Calculate Feature Scores for each Class
Rank Features based on Score
Visualization

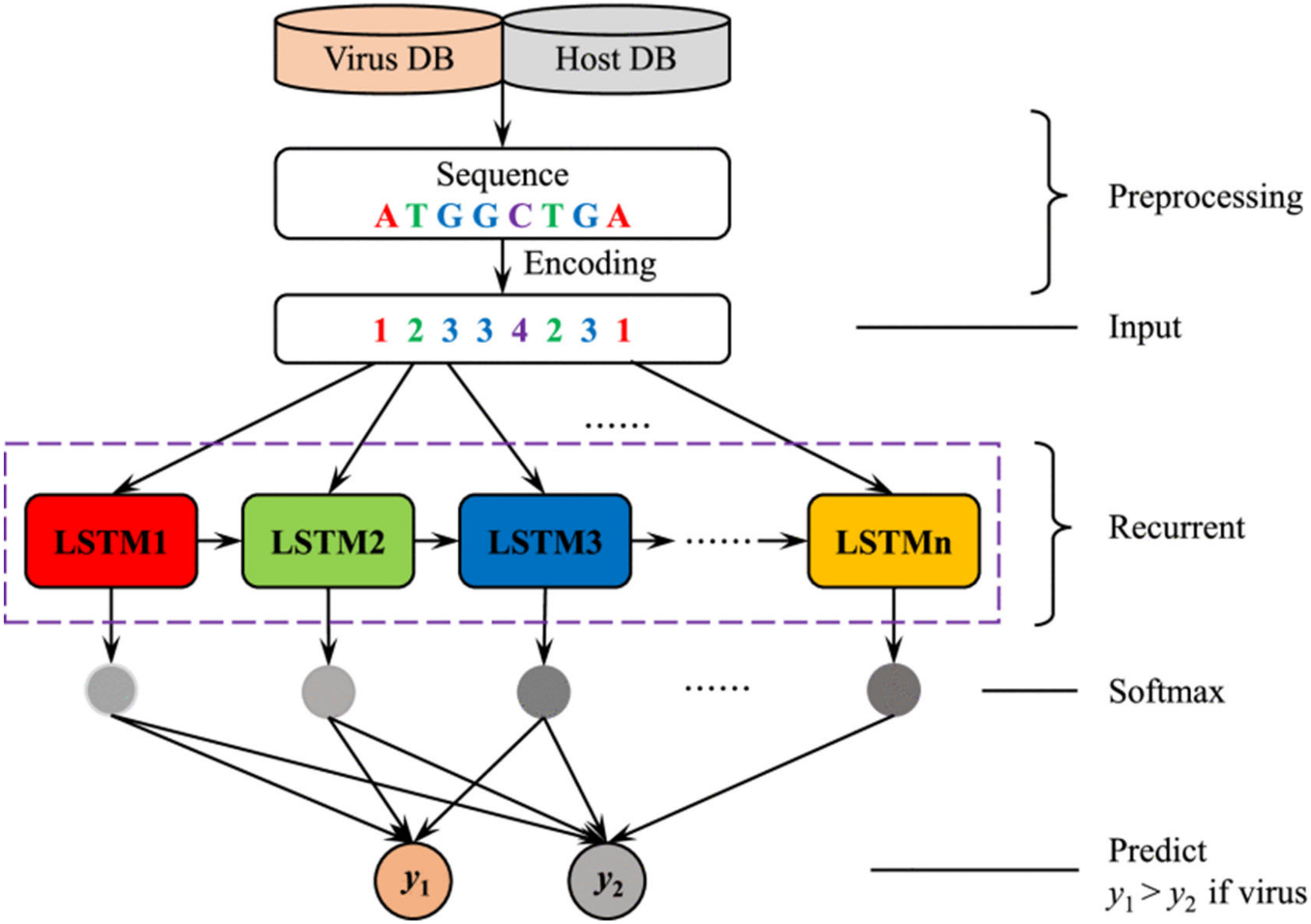
https://www.biorxiv.org/content/10.1101/257931v1.full

antimicrobial peptides

https://www.nature.com/articles/s41587-022-01226-0

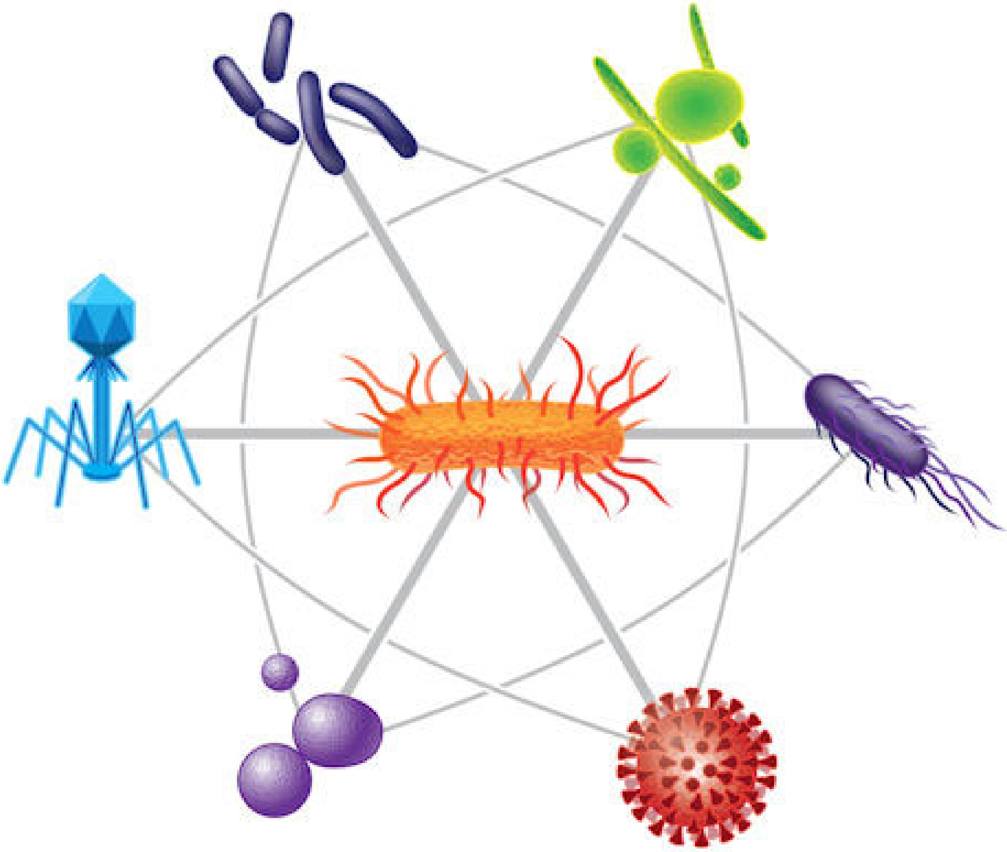


virus

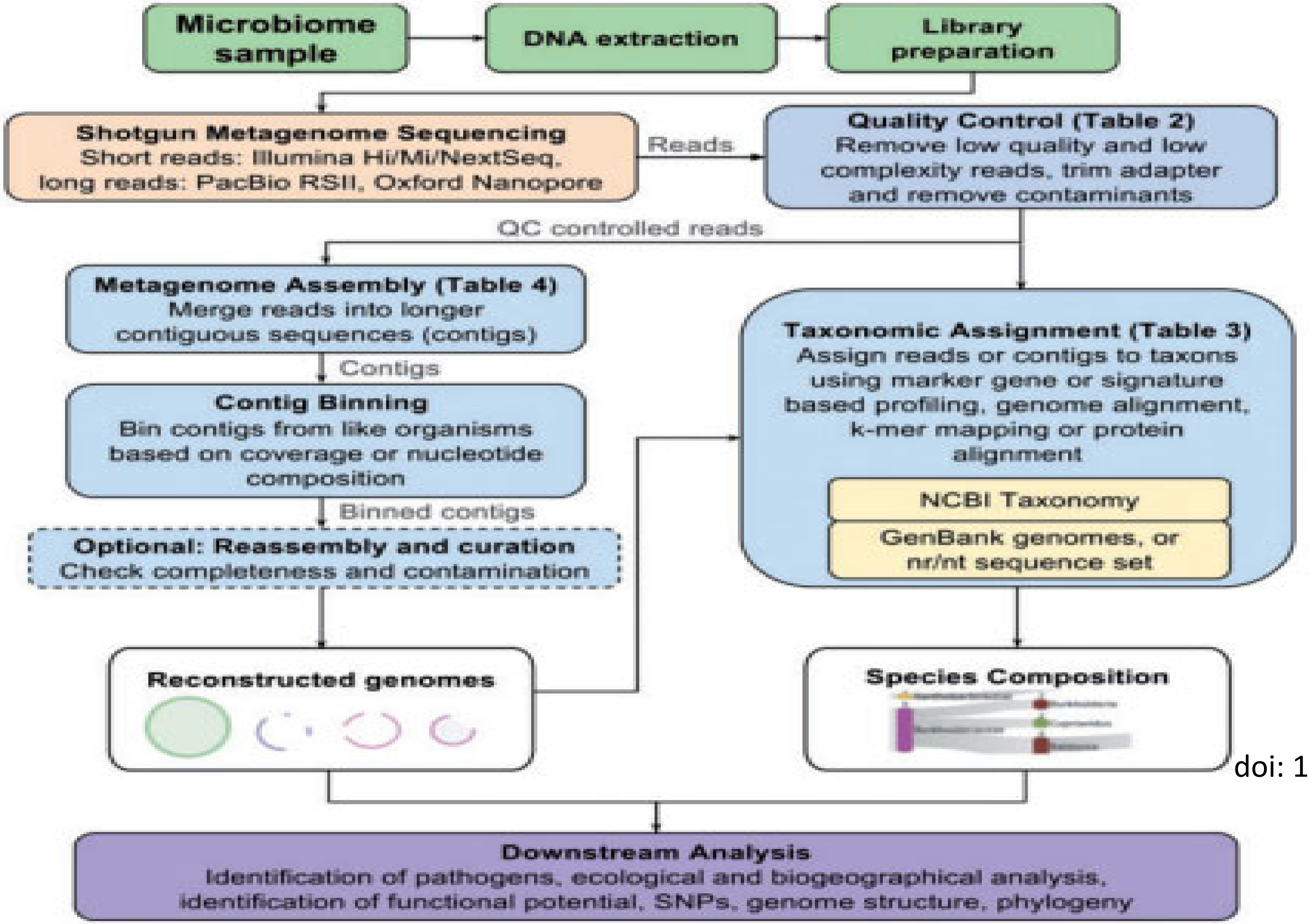


https://ieeexplore.ieee.org/document/9293366

Microbial Interaction



https://asm.org/magazine/2023/spring/friends-foes-microbial-interactions-infections



Microbiome sample

DNA extraction

Library preparation

Shotgun Metagenome Sequencing
Short reads: Illumina Hi/Mi/NextSeq,
long reads: PacBio RSII, Oxford Nanopore

Reads

Quality Control (Table 2)
Remove low quality and low complexity reads, trim adapter and remove contaminants

QC controlled reads

Metagenome Assembly (Table 4)
Merge reads into longer contiguous sequences (contigs)

Contigs

Contig Binning
Bin contigs from like organisms based on coverage or nucleotide composition

Binned contigs

Optional: Reassembly and curation
Check completeness and contamination

Taxonomic Assignment (Table 3)
Assign reads or contigs to taxons using marker gene or signature based profiling, genome alignment, k-mer mapping or protein alignment

NCBI Taxonomy

GenBank genomes, or nr/nt sequence set

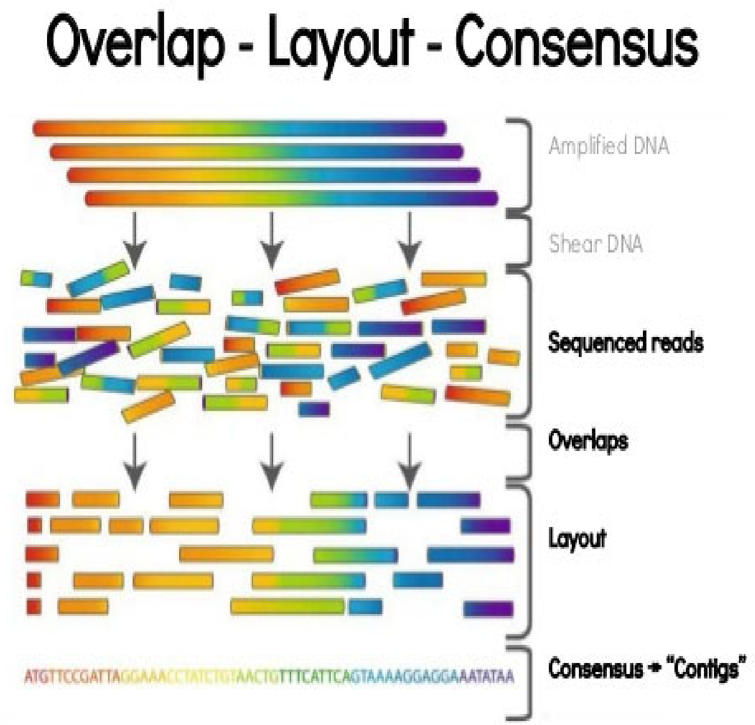
Reconstructed genomes

Species Composition

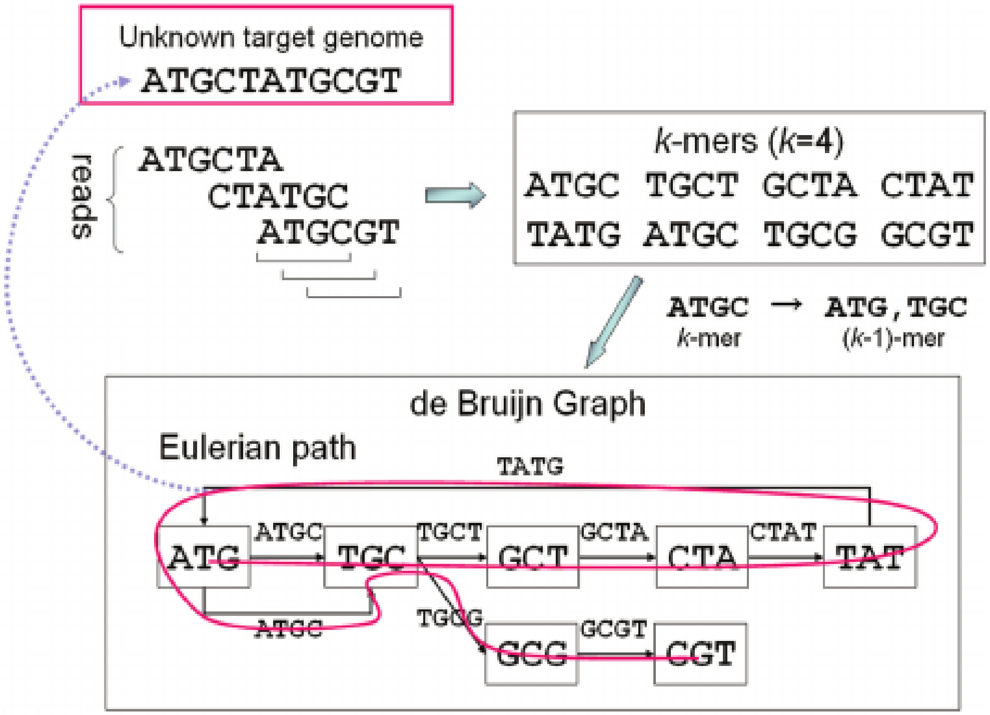
Downstream Analysis
Identification of pathogens, ecological and biogeographical analysis, identification of functional potential, SNPs, genome structure, phylogeny

doi: 10.1093/bib/bbx120

Genome Assembly

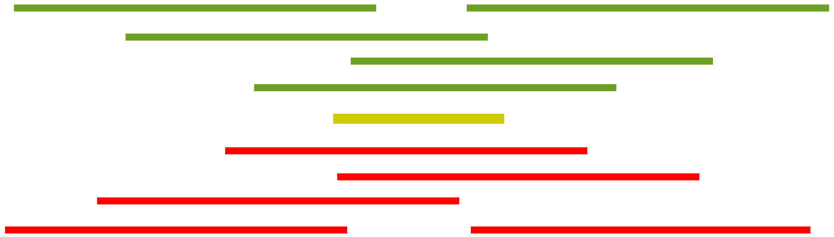

Overlap - Layout - Consensus
Amplified DNA
Shear DNA
Sequenced reads
Overlaps
Layout
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
Consensus → "Contigs"

de Bruijn Graph

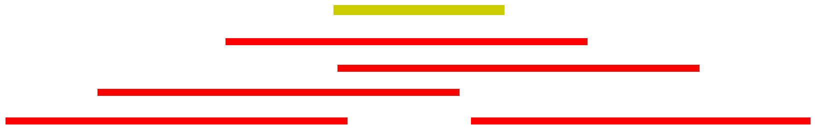


http://www.chromnet.net/Genome%20De%20Novo%20Sequencing%20and%20Assembly.aspx
https://www.researchgate.net/figure/Illustration-of-de-Bruijn-graph-based-assembly_fig1_229437536

Troublesome Repeats



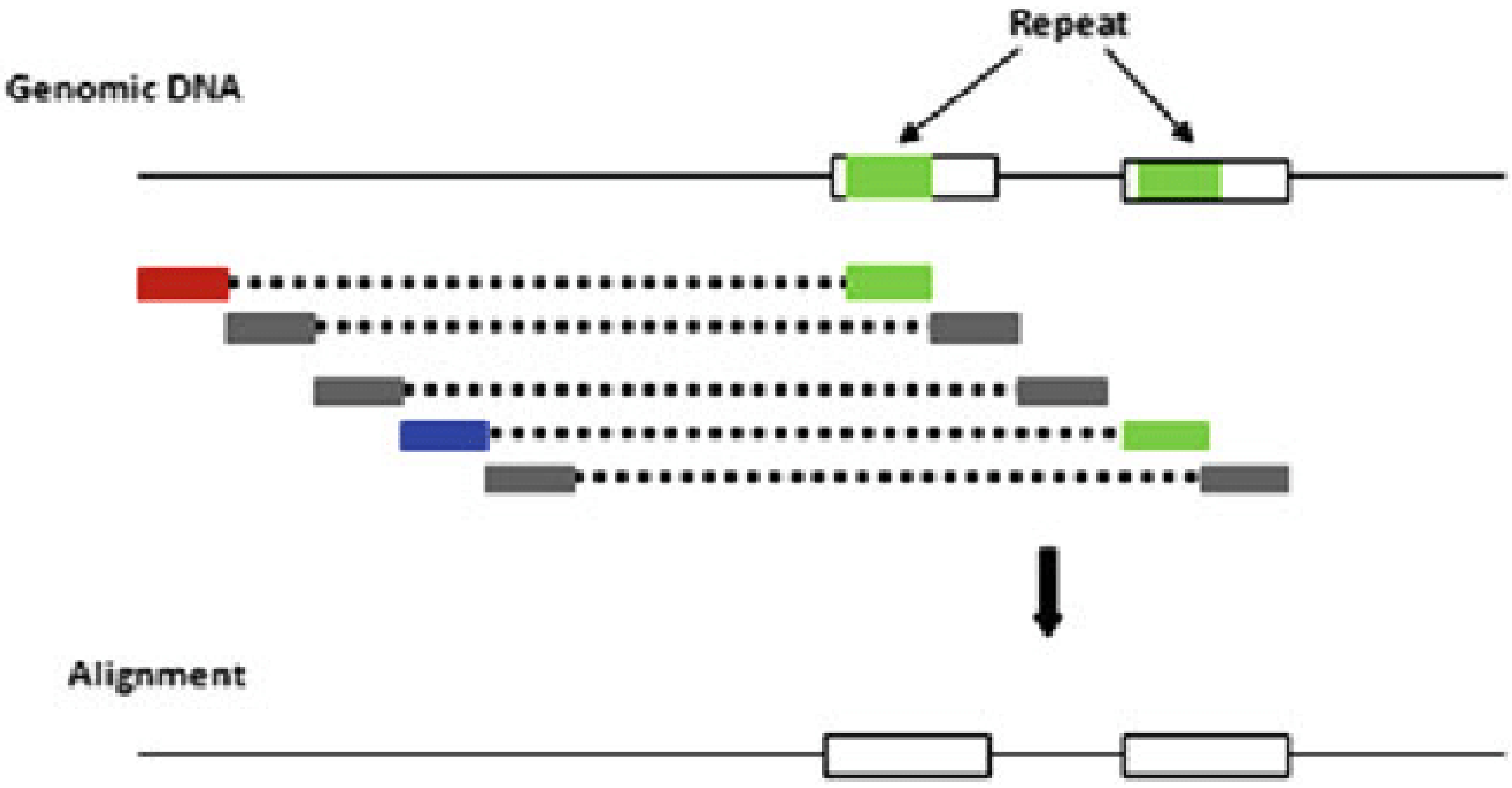
?????



Unambiguous

- Insert non-maximal reads whenever unambiguous

Paired-end reads and Repeats



Challenges in Metagenomic Assembly

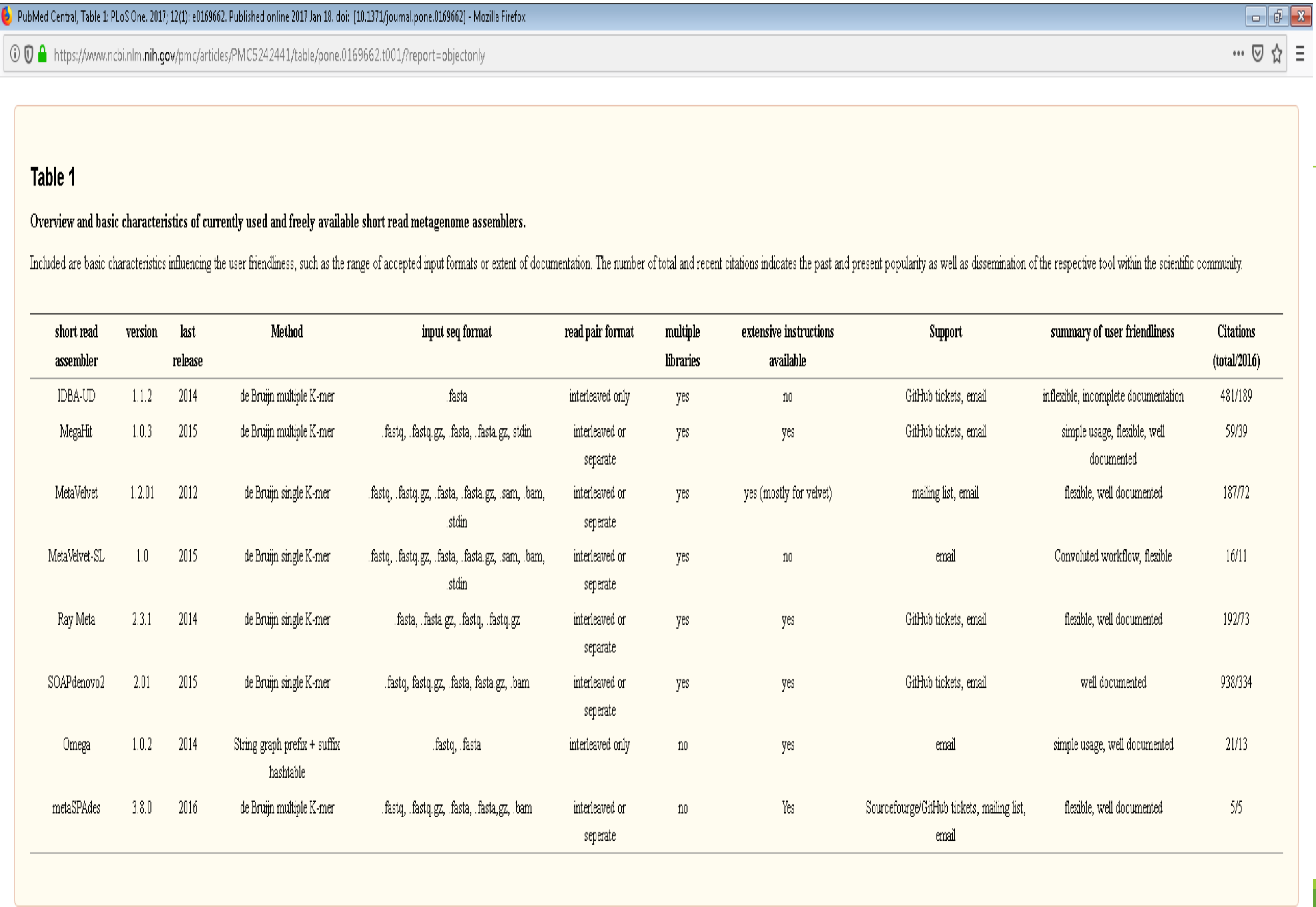
Mixed reads

Horizontal gene transfer

Highly conserved genes

Uneven sequencing coverage

Assembly



PubMed Central, Table 1: PLoS One. 2017; 12(1): e0169662. Published online 2017 Jan 18. doi: [10.1371/journal.pone.0169662] - Mozilla Firefox

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5242441/table/pone.0169662.t001/?report=objectonly

Table 1

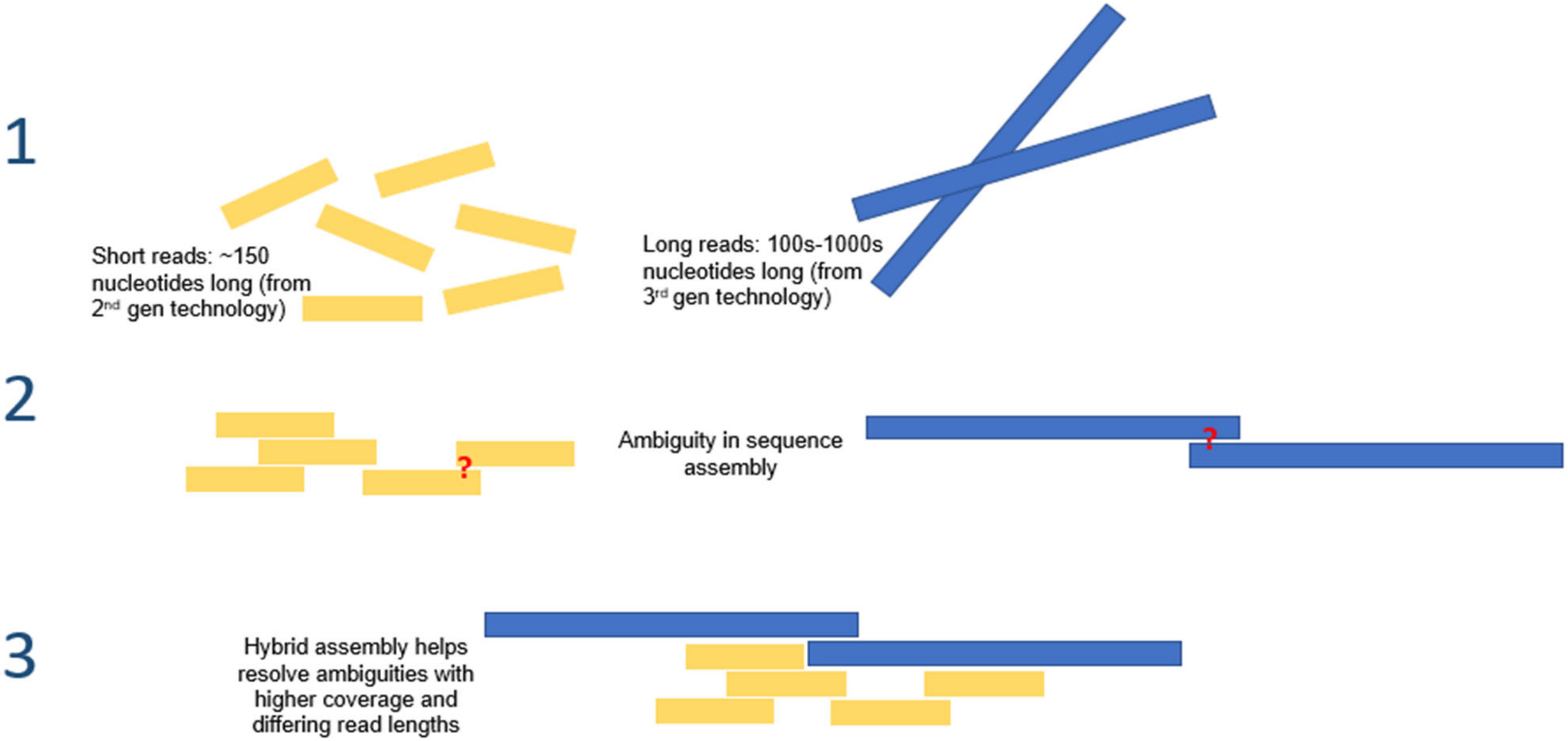
Overview and basic characteristics of currently used and freely available short read metagenome assemblers.

Included are basic characteristics influencing the user friendliness, such as the range of accepted input formats or extent of documentation. The number of total and recent citations indicates the past and present popularity as well as dissemination of the respective tool within the scientific community.

short read assembler	version	last release	Method	input seq format	read pair format	multiple libraries	extensive instructions available	Support	summary of user friendliness	Citations (total/2016)
IDBA-UD	1.1.2	2014	de Bruijn multiple K-mer	.fasta	interleaved only	yes	no	GitHub tickets, email	inflexible, incomplete documentation	481/189
MegaHit	1.0.3	2015	de Bruijn multiple K-mer	.fastq, .fastq.gz, .fasta, .fasta.gz, stdin	interleaved or separate	yes	yes	GitHub tickets, email	simple usage, flexible, well documented	59/39
MetaVelvet	1.2.01	2012	de Bruijn single K-mer	.fastq, .fastq.gz, .fasta, .fasta.gz, .sam, .bam, .stdin	interleaved or seperate	yes	yes (mostly for velvet)	mailing list, email	flexible, well documented	187/72
MetaVelvet-SL	1.0	2015	de Bruijn single K-mer	.fastq, .fastq.gz, .fasta, .fasta.gz, .sam, .bam, .stdin	interleaved or seperate	yes	no	email	Convoluted workflow, flexible	16/11
Ray Meta	2.3.1	2014	de Bruijn single K-mer	.fasta, .fasta.gz, .fastq, .fastq.gz	interleaved or separate	yes	yes	GitHub tickets, email	flexible, well documented	192/73
SOAPdenovo2	2.01	2015	de Bruijn single K-mer	.fastq, fastq.gz, .fasta, fasta.gz, .bam	interleaved or seperate	yes	yes	GitHub tickets, email	well documented	938/334
Omega	1.0.2	2014	String graph prefix + suffix hashtable	.fastq, .fasta	interleaved only	no	yes	email	simple usage, well documented	21/13
metaSPAdes	3.8.0	2016	de Bruijn multiple K-mer	.fastq, .fastq.gz, .fasta, .fasta,gz, .bam	interleaved or seperate	no	Yes	Sourcefourge/GitHub tickets, mailing list, email	flexible, well documented	5/5

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5242441/

Long and short read hybrid assembly



1

Short reads: ~150 nucleotides long (from 2nd gen technology)

Long reads: 100s-1000s nucleotides long (from 3rd gen technology)

2

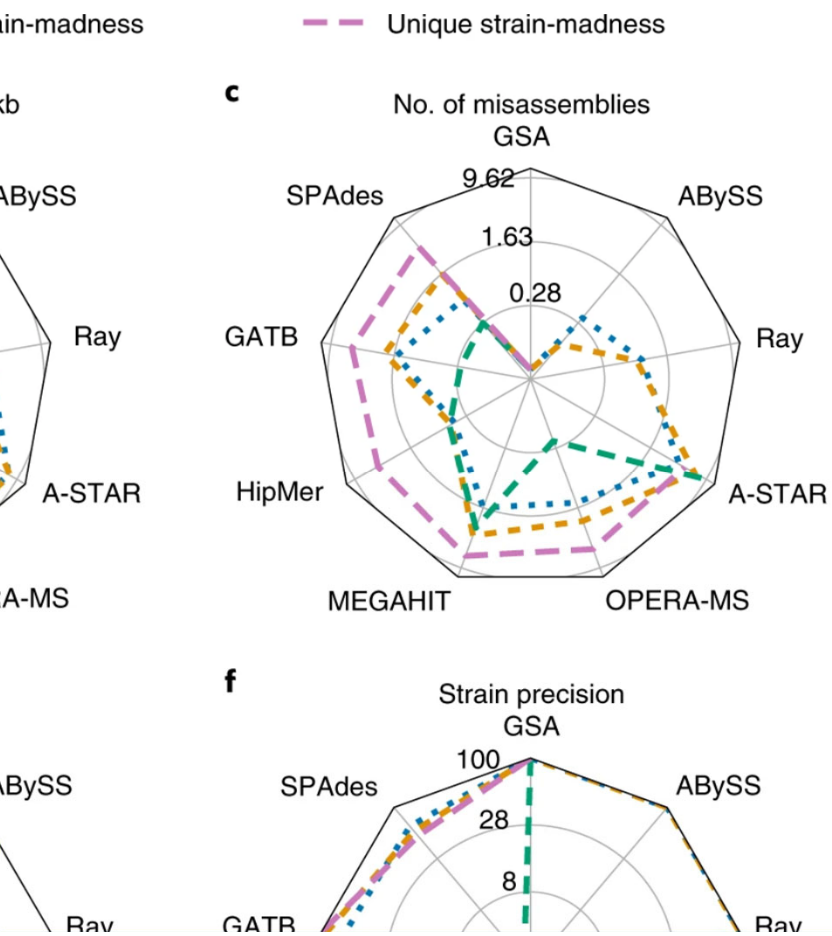
Ambiguity in sequence assembly

3

Hybrid assembly helps resolve ambiguities with higher coverage and differing read lengths

https://en.wikipedia.org/wiki/Hybrid_genome_assembly

Current Status

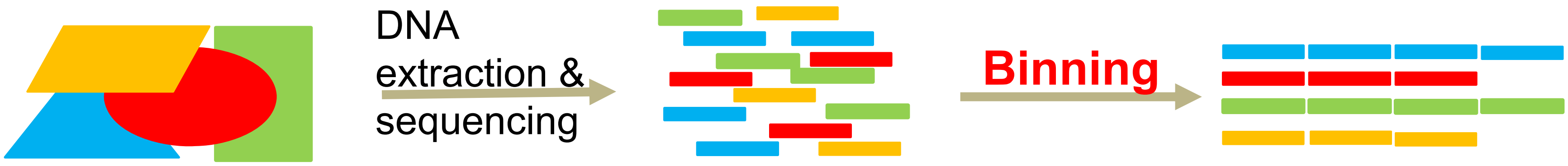


For difficult regions (repeats or highly conserved regions)

-STAR partially recovered 102 (78%) of 131 16S sequences. The hybrid assemblers GATB (mean completeness 60.1%) and OPERA-MS (mean 47.1%) recovered the most complete 16S sequences. Mean completeness for short-read assemblies ranged from 29.6% (HipMer) to 36.9% (MEGAHIT)

https://www.nature.com/articles/s41592-022-01431-4/figures/1

Binning



DNA extraction & sequencing

Binning

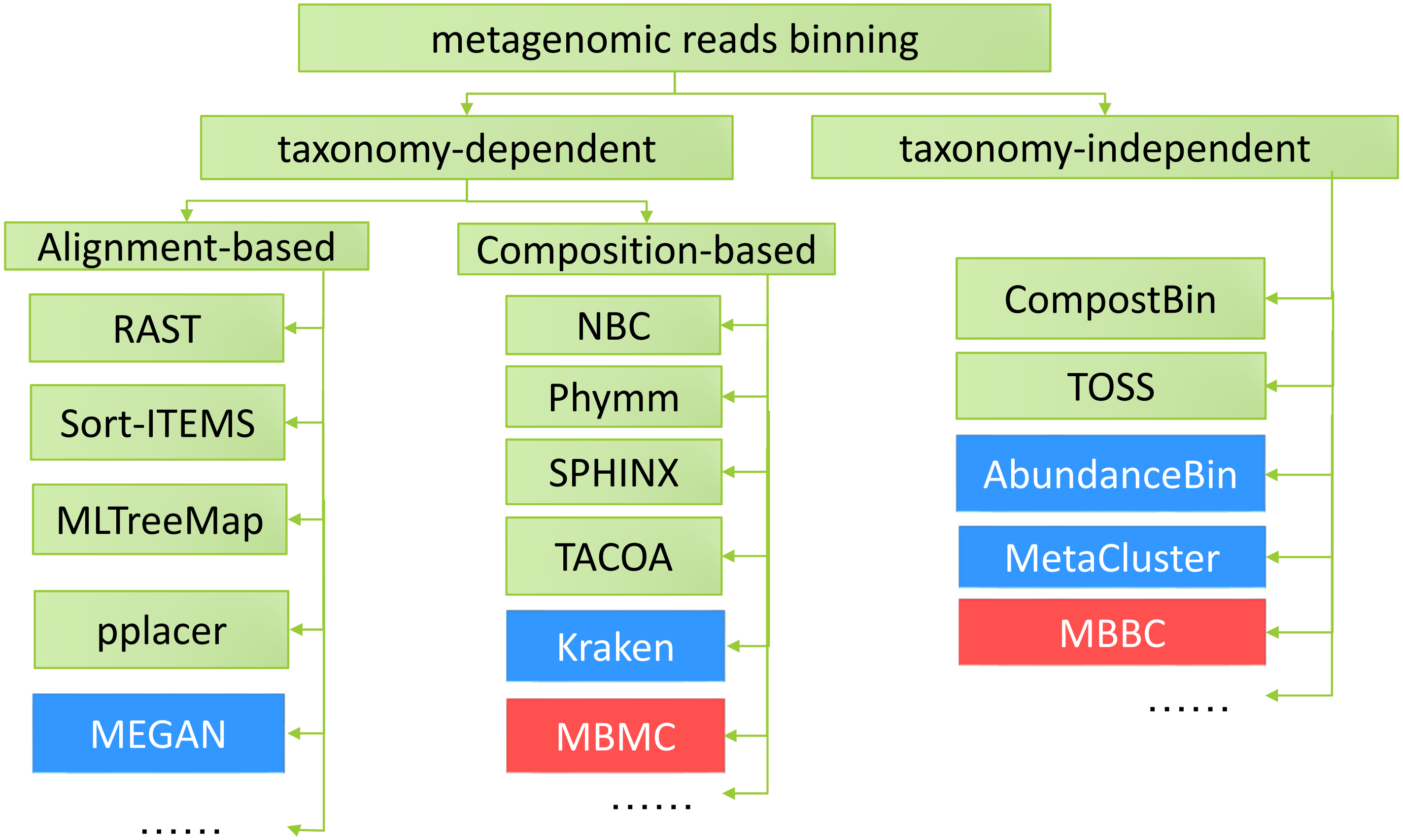
environmental sample

reads

Binned reads

- Short reads length
- Sequencing errors
- high-throughput data

Current approaches



Taxonomy-dependent methods:

Alignment-based binning (**MEGAN**)

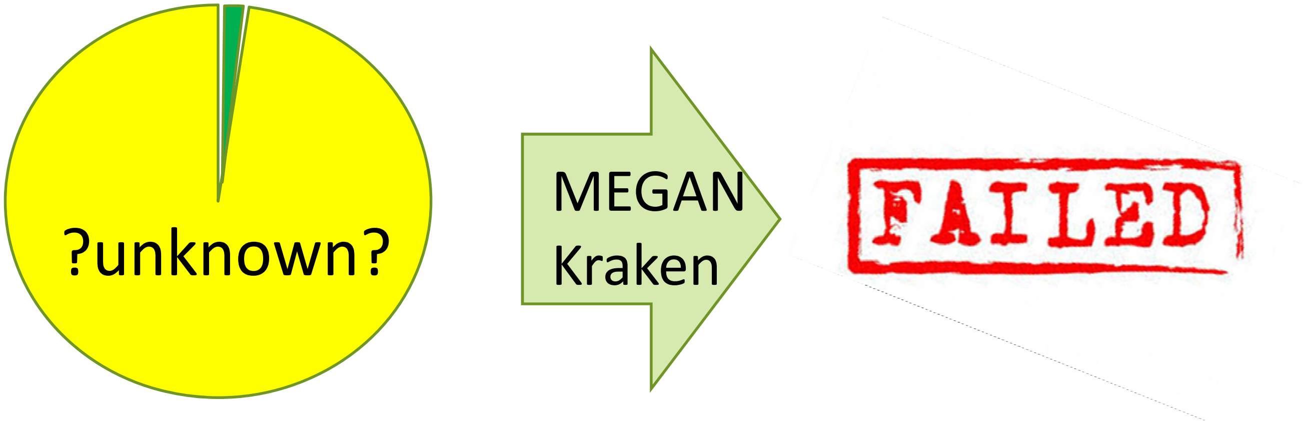
http://ab.inf.uni-tuebingen.de/software/welcome.html/megan5

Composition-based binning (**Kraken**)

Kraken
Taxonomic Sequence Classification System

https://ccb.jhu.edu/software/kraken/

Problems in taxonomy-dependent methods



MBMC works better for datasets that contain unknown species

Taxonomy-independent methods

AbundanceBin, **MetaCluster**

the difference of k-mer (short sequence with length k) frequencies of different microbes in the environmental samples

Observations:

- k-mer frequency ⟷ genome's coverage
- long k-mers are usually unique in each genome
- k-mer frequency distribution from the same genome are similar

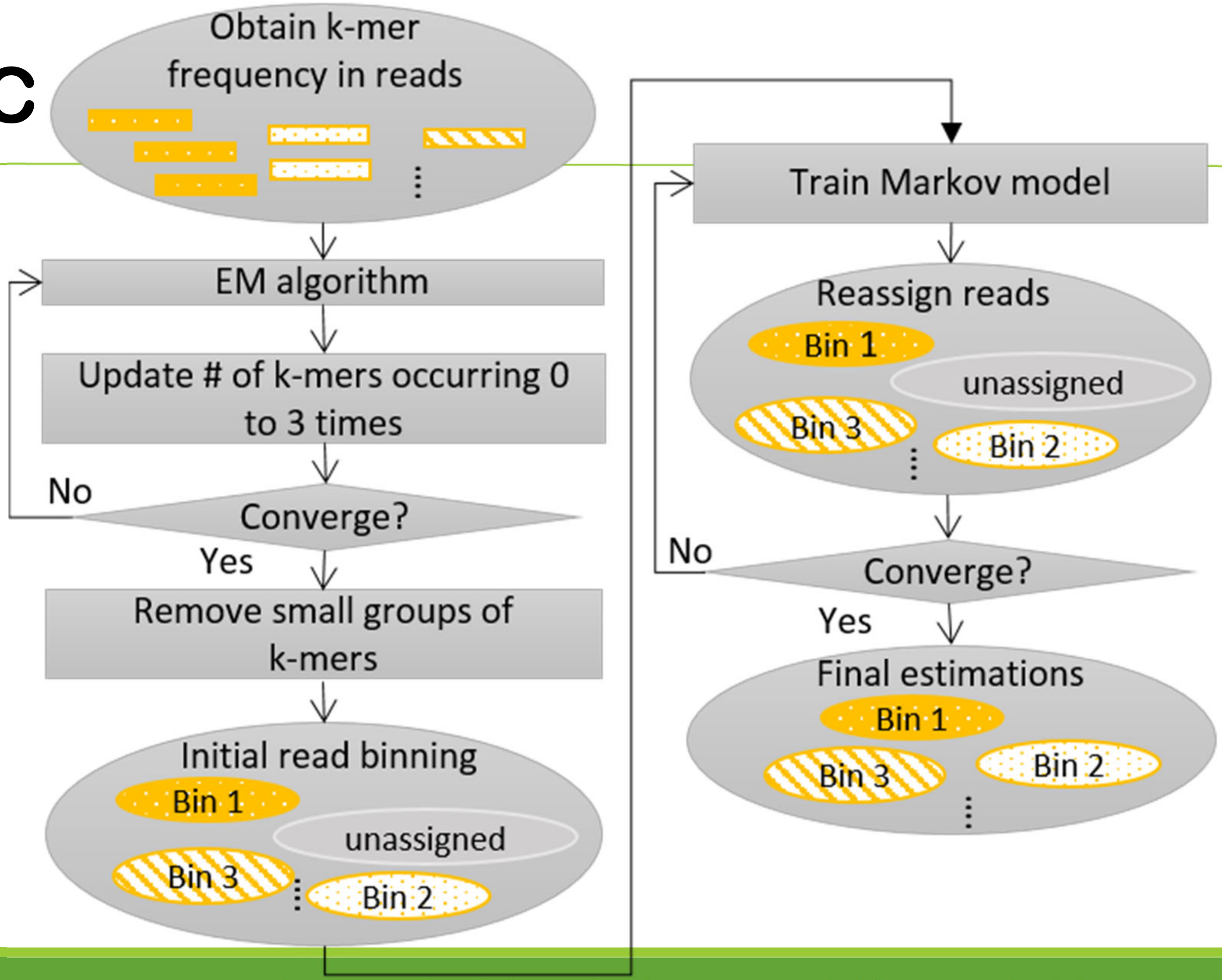
Problems in taxonomy-independent methods

➢AbundanceBin/MetaCluster: only utilize the k-mers frequency.

→MBBC: improve the method to utilize the k-mer frequency;

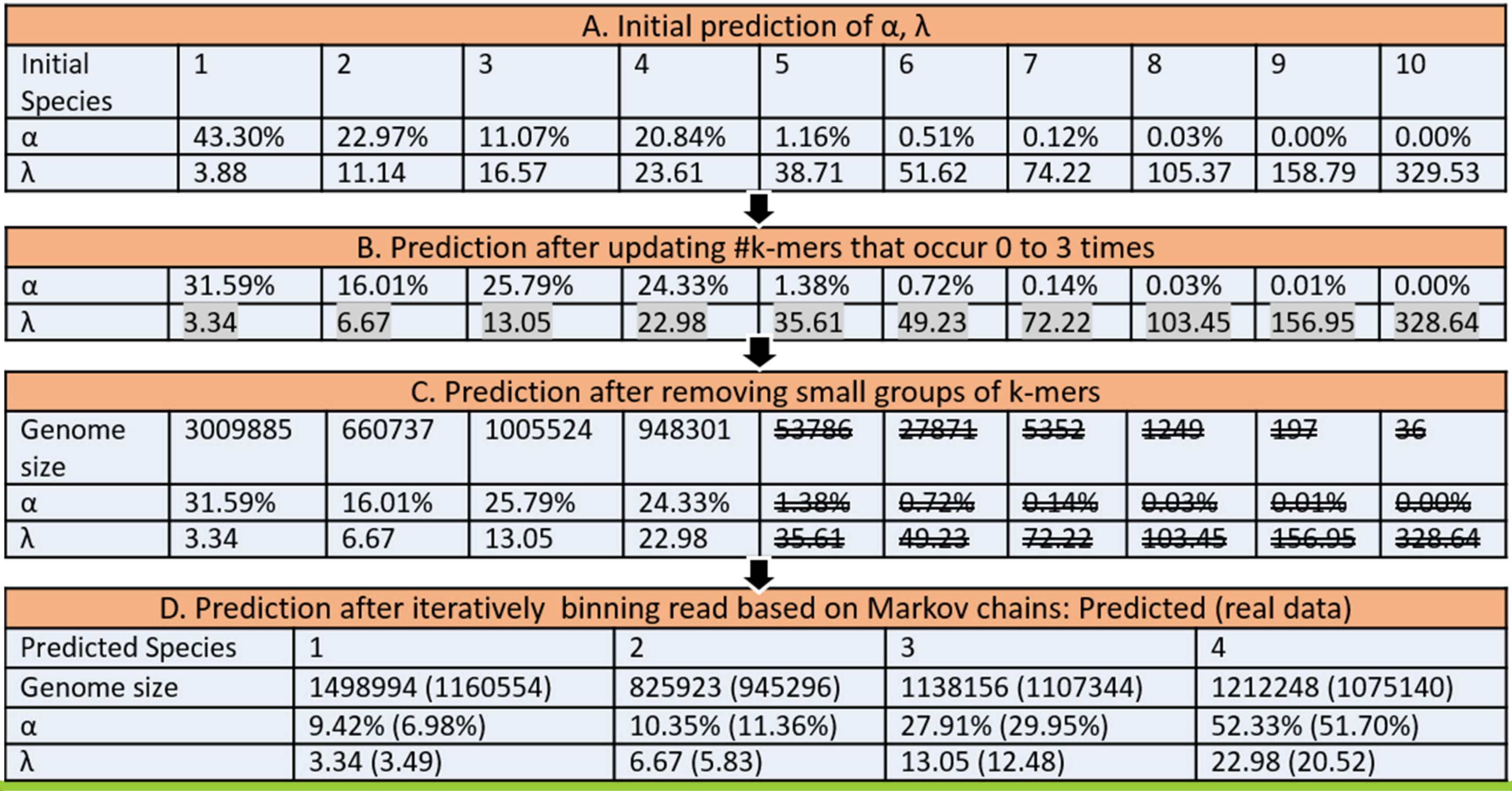
utilized Markov properties shared by a group of reads

MBBC



The procedure of read clustering in MBBC. The output on the right from each of the main steps on the left is connected with the corresponding steps.
Wang *et al. BMC Bioinformatics* 2015 **16**:36 doi:10.1186/s12859-015-0473-8

Results1 MBBC reliably estimates the species number, genome sizes, relative species abundances, and k-mer coverage



A. Initial prediction of α, λ

Initial Species	1	2	3	4	5	6	7	8	9	10
α	43.30%	22.97%	11.07%	20.84%	1.16%	0.51%	0.12%	0.03%	0.00%	0.00%
λ	3.88	11.14	16.57	23.61	38.71	51.62	74.22	105.37	158.79	329.53

B. Prediction after updating #k-mers that occur 0 to 3 times

α	31.59%	16.01%	25.79%	24.33%	1.38%	0.72%	0.14%	0.03%	0.01%	0.00%
λ	3.34	6.67	13.05	22.98	35.61	49.23	72.22	103.45	156.95	328.64

C. Prediction after removing small groups of k-mers

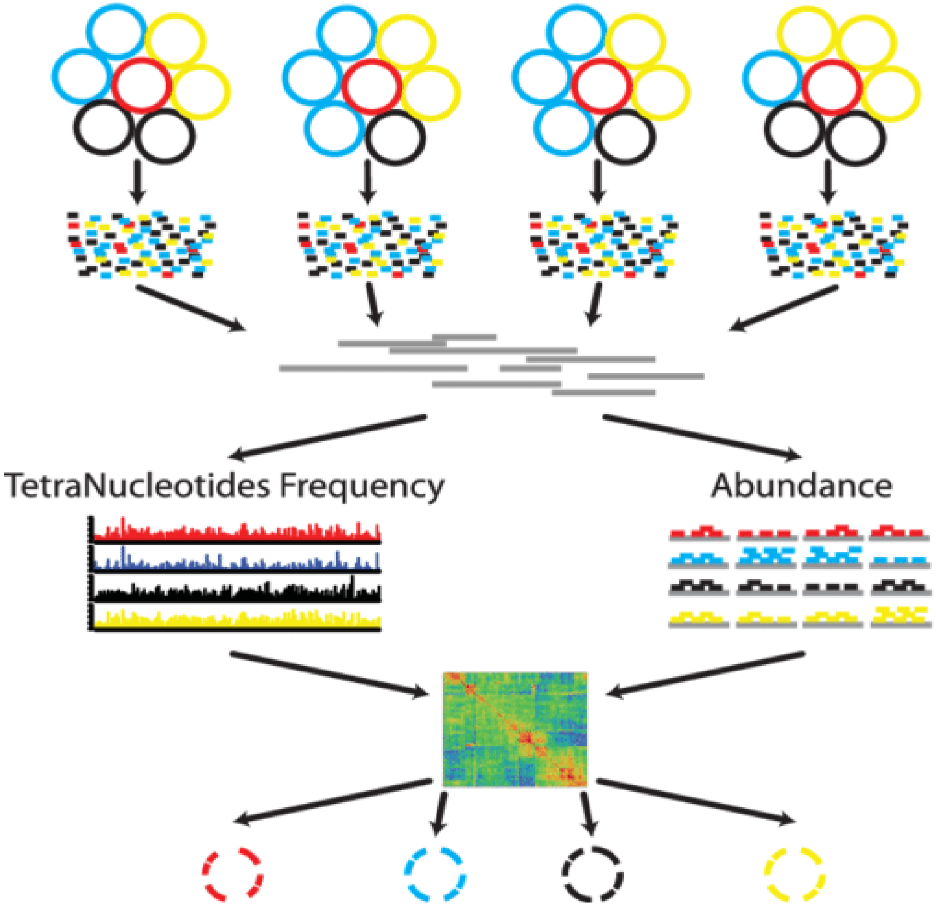
Genome size	3009885	660737	1005524	948301	~~53786~~	~~27871~~	~~5352~~	~~1249~~	~~197~~	~~36~~
α	31.59%	16.01%	25.79%	24.33%	~~1.38%~~	~~0.72%~~	~~0.14%~~	~~0.03%~~	~~0.01%~~	~~0.00%~~
λ	3.34	6.67	13.05	22.98	~~35.61~~	~~49.23~~	~~72.22~~	~~103.45~~	~~156.95~~	~~328.64~~

D. Prediction after iteratively binning read based on Markov chains: Predicted (real data)

Predicted Species	1	2	3	4
Genome size	1498994 (1160554)	825923 (945296)	1138156 (1107344)	1212248 (1075140)
α	9.42% (6.98%)	10.35% (11.36%)	27.91% (29.95%)	52.33% (51.70%)
λ	3.34 (3.49)	6.67 (5.83)	13.05 (12.48)	22.98 (20.52)

MBBC: an efficient approach for metagenomic binning based on clustering
BMC Bioinformatics. 2015;16(1):36.

Contig Binning



Preprocessing

1 Samples from multiple sites or times

2 Metagenome libraries

3 Initial de-novo assembly using the combined library

MetaBAT

4 Calculate TNF for each contig

5 Calculate Abundance per library for each contig

6 Calculate the pairwise distance matrix using pre-trained probabilistic models

7 Forming genome bins iteratively

https://pubmed.ncbi.nlm.nih.gov/26336640/

Concoct

Tetramer frequency (V= 136 tetramers)

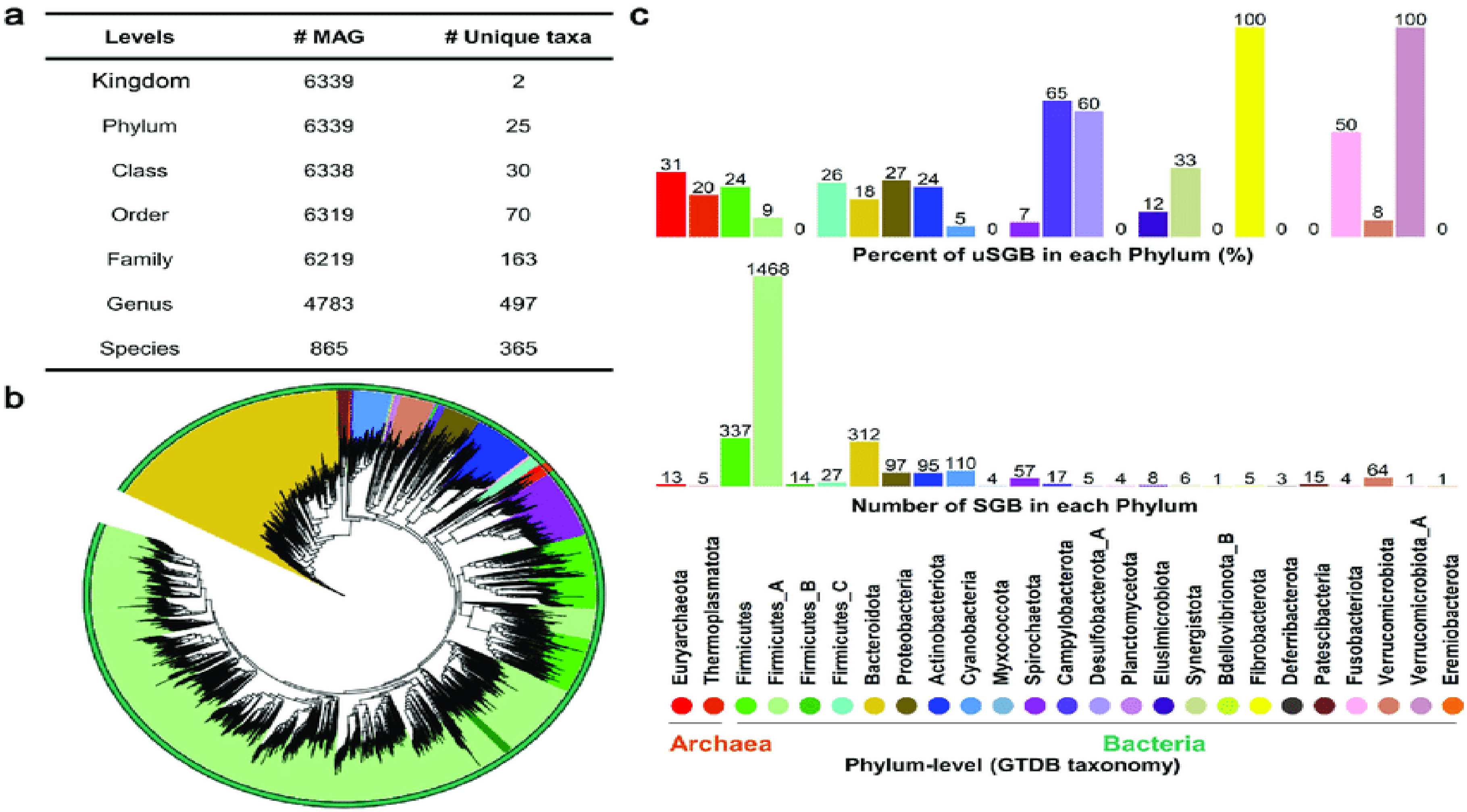
Normalized Coverage of a contig across M samples

Total coverage of a contig across M samples

Dimension reduction by SVD on all N contigs' Log M+1+V vectors to account for >90% variation of the vectors, then apply Gaussian mixture decomposition to obtain different MAGs.

https://pubmed.ncbi.nlm.nih.gov/26336640/

Taxonomic Analysis

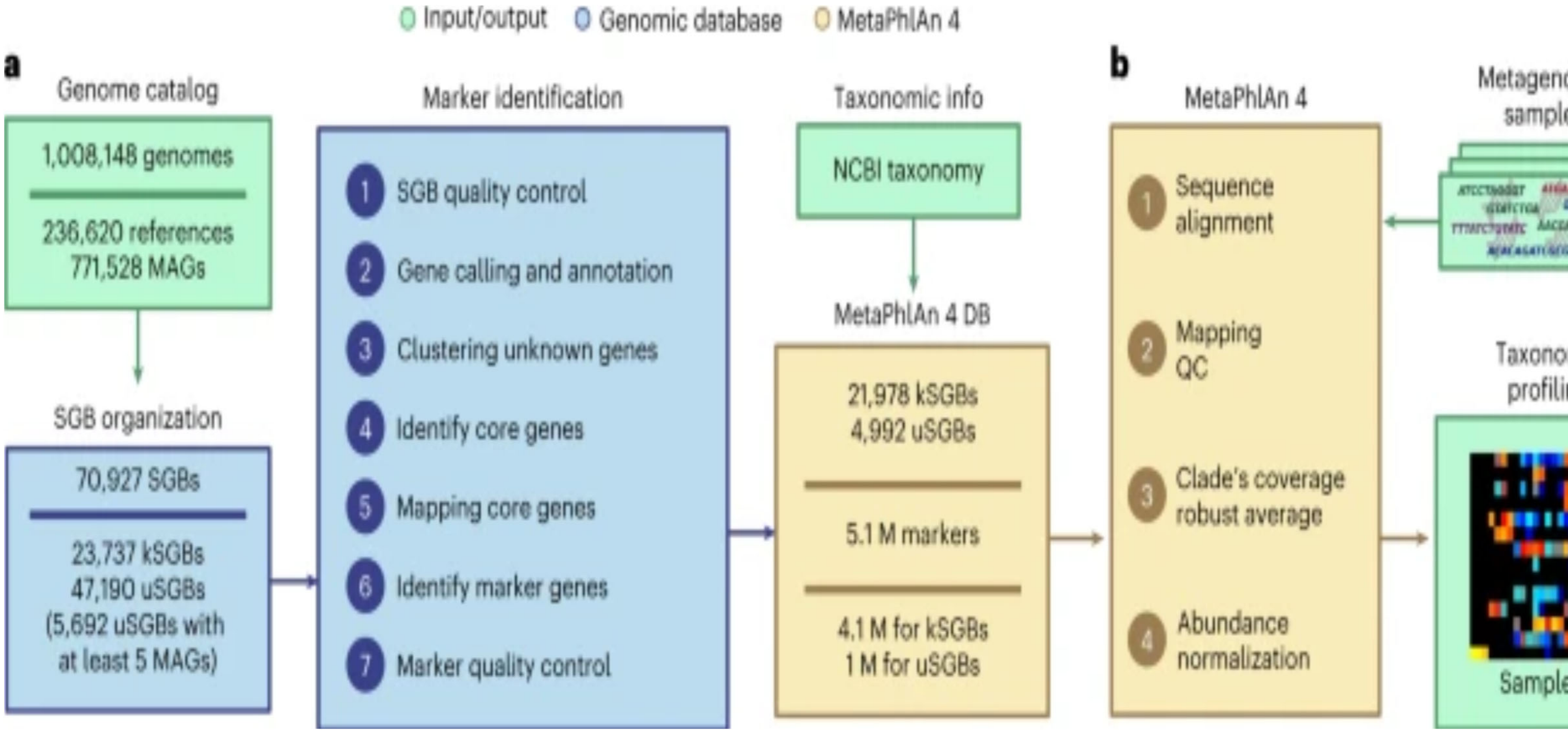


a

Levels	# MAG	# Unique taxa
Kingdom	6339	2
Phylum	6339	25
Class	6338	30
Order	6319	70
Family	6219	163
Genus	4783	497
Species	865	365

https://www.nature.com/articles/s41467-021-21295-0

MetaPhlAn



https://www.nature.com/articles/s41587-023-01688-w

Other metagenomic data

metatranscirptomics

metaproteomics

Single cell metagenomics

Hi-C