Computational Molecular Biology and Bioinformatics

MetaSUB

Malay Bhattacharyya

Associate Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

November, 2025

What is metagenomics?

Metagenomics is defined as the direct genetic analysis of genomes (microbial communities) contained within a site (environmental, clinical, built-in, etc.).

Metagenomics allows for a detailed study of the diversity of communities, and therefore to clarify the mechanisms of their functioning, to determine the metabolic relationships.

Note: In Greek, meta means "beyond or above the range of normal or physical human experience".

What is microbiome?

Microbiome is the collection of ecological communities of microorganisms (commensal, symbiotic and pathogenic) that reside at a particular site.

What is microbiome?

Microbiome is the collection of ecological communities of microorganisms (commensal, symbiotic and pathogenic) that reside at a particular site.

- **Commensal:** An organism that uses food supplied in the internal or the external environment of the host, without establishing a close association with the host. E.g., Staphylococcus epidermidis found on human skin.

What is microbiome?

Microbiome is the collection of ecological communities of microorganisms (commensal, symbiotic and pathogenic) that reside at a particular site.

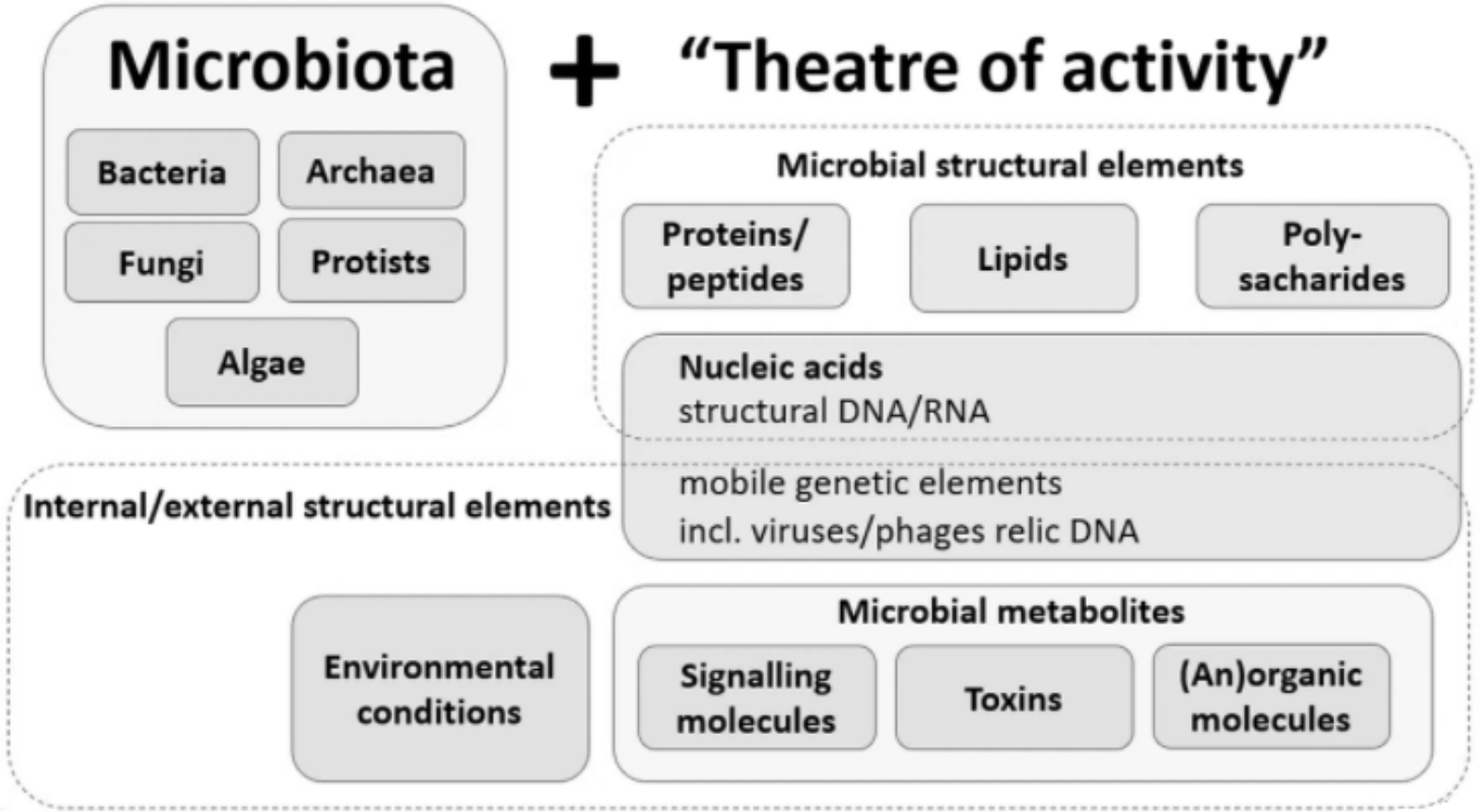
- **Commensal:** An organism that uses food supplied in the internal or the external environment of the host, without establishing a close association with the host. E.g., Staphylococcus epidermidis found on human skin.
- **Symbiotic:** An organism that lives in beneficial association with the host. E.g., Bacteroides thetaiotaomicron found in human intestine.

What is microbiome?

Microbiome is the collection of ecological communities of microorganisms (commensal, symbiotic and pathogenic) that reside at a particular site.

- **Commensal:** An organism that uses food supplied in the internal or the external environment of the host, without establishing a close association with the host. E.g., Staphylococcus epidermidis found on human skin.
- **Symbiotic:** An organism that lives in beneficial association with the host. E.g., Bacteroides thetaiotaomicron found in human intestine.
- **Pathogenic:** An organism which is capable of causing diseases in a host. E.g., SARS-CoV-2 causing COVID-19 in human.

What is microbiome?

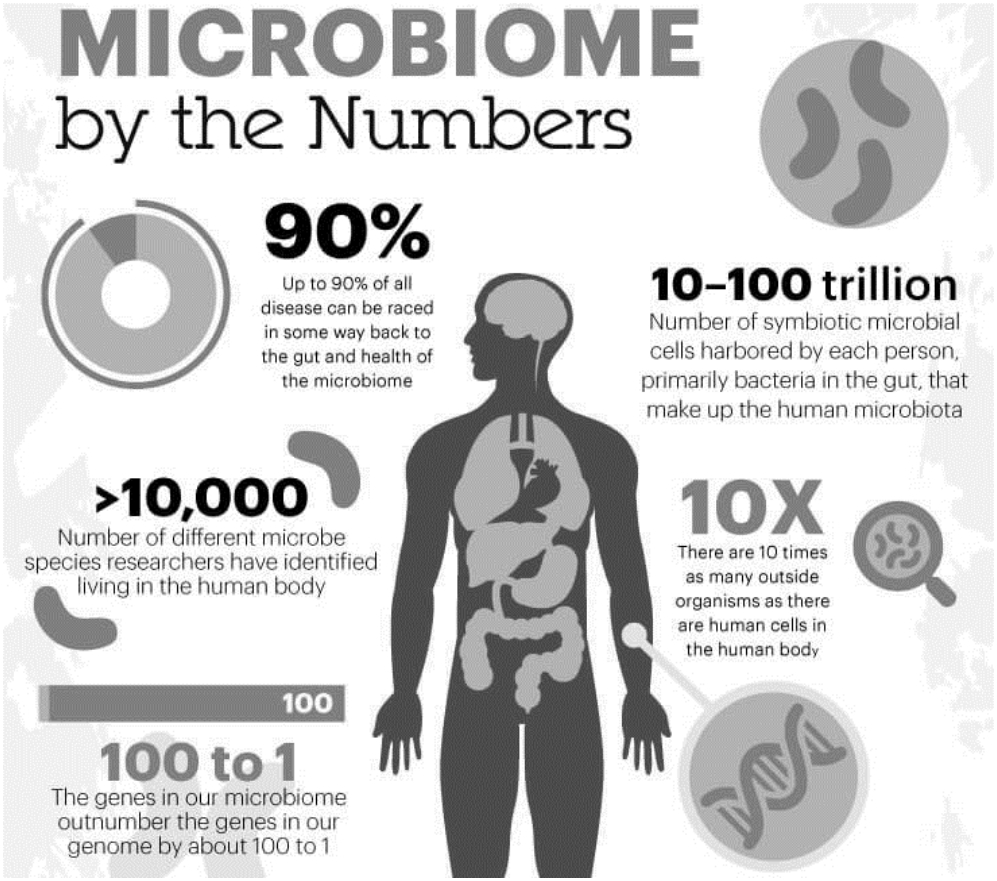


Human microbiome

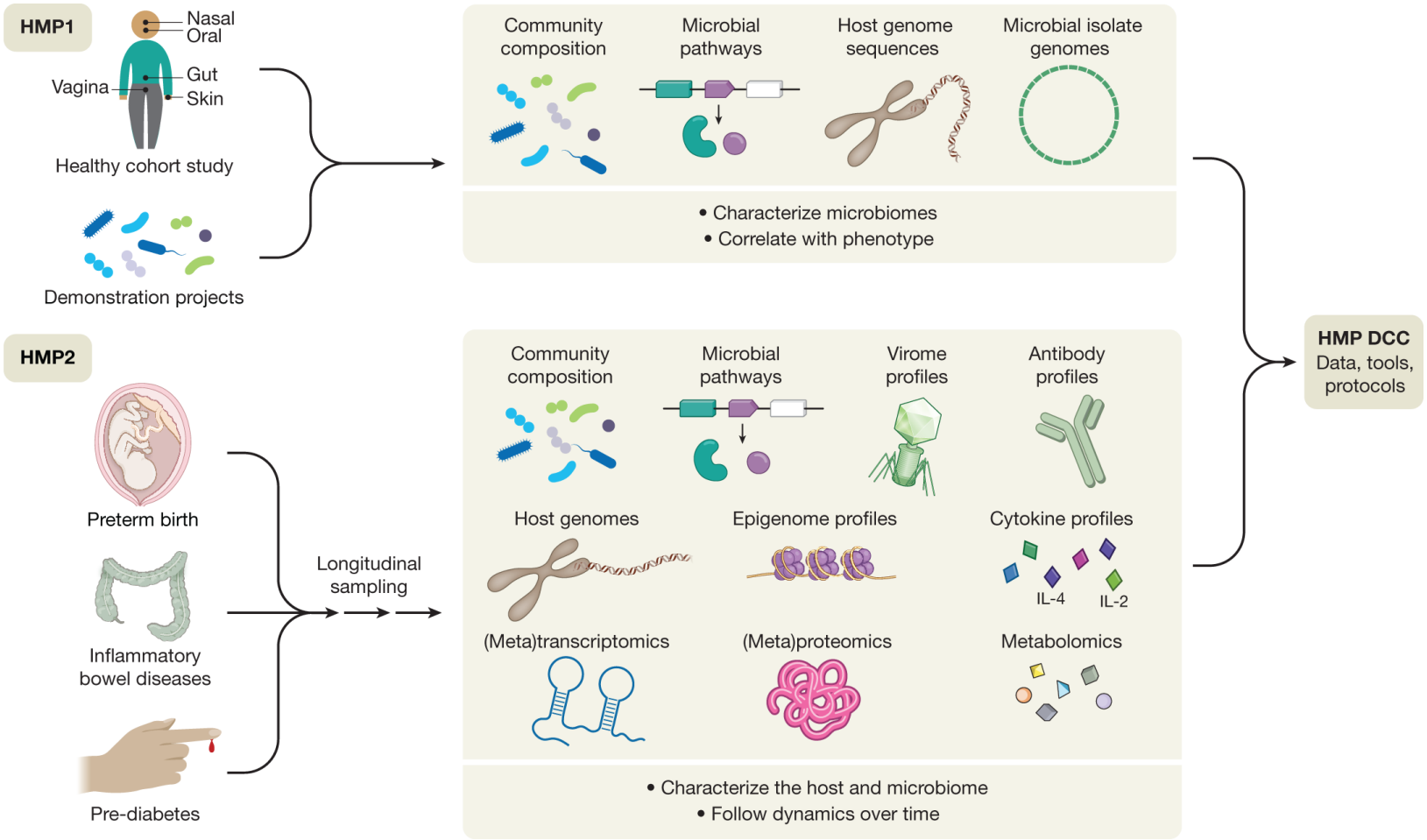
The human microbiome is the collection of microorganisms that reside on or within human tissues and biofluids corresponding to anatomical sites.

Human microbiome

The human microbiome is the collection of microorganisms that reside on or within human tissues and biofluids corresponding to anatomical sites.



The integrative Human Microbiome Project (iHMP)

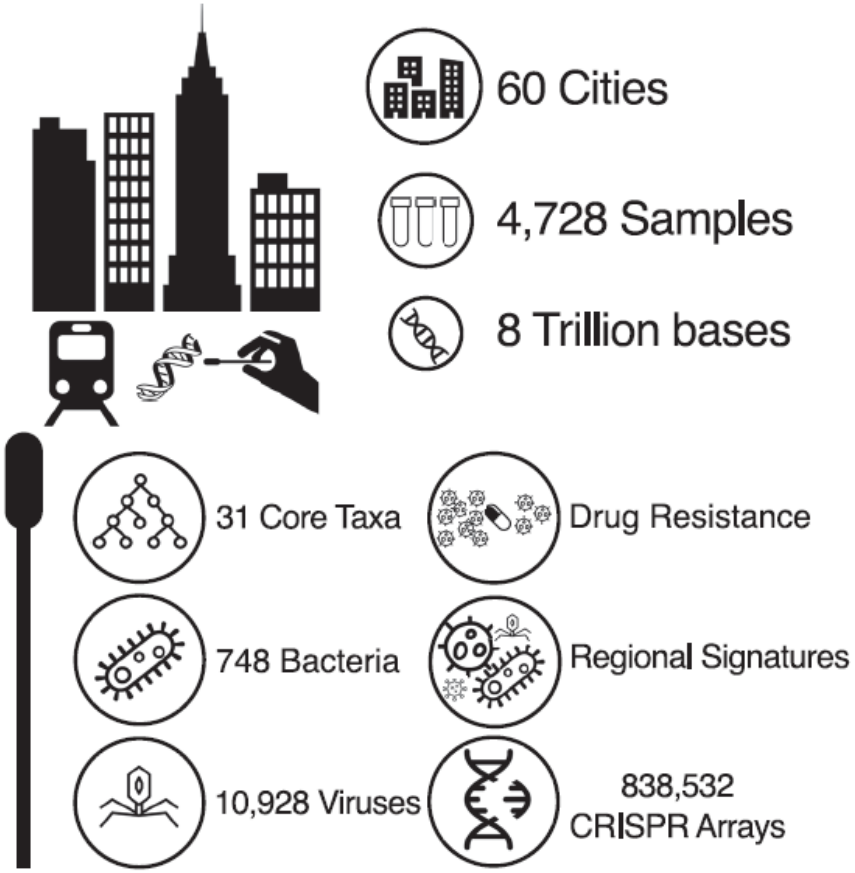


Source: Nature, 569:641-648, 2019.

Microbiome of the built environment

Microbiome of the built environment comprises the communities of microorganisms that reside in human constructed environments.

Microbiome in mass-transit systems



Source: Cell, 184(13):3376-3393, 2021.

Availability of data

MetaSUB sequencing data is available at:
`https://pngb.io/metasub-2021`

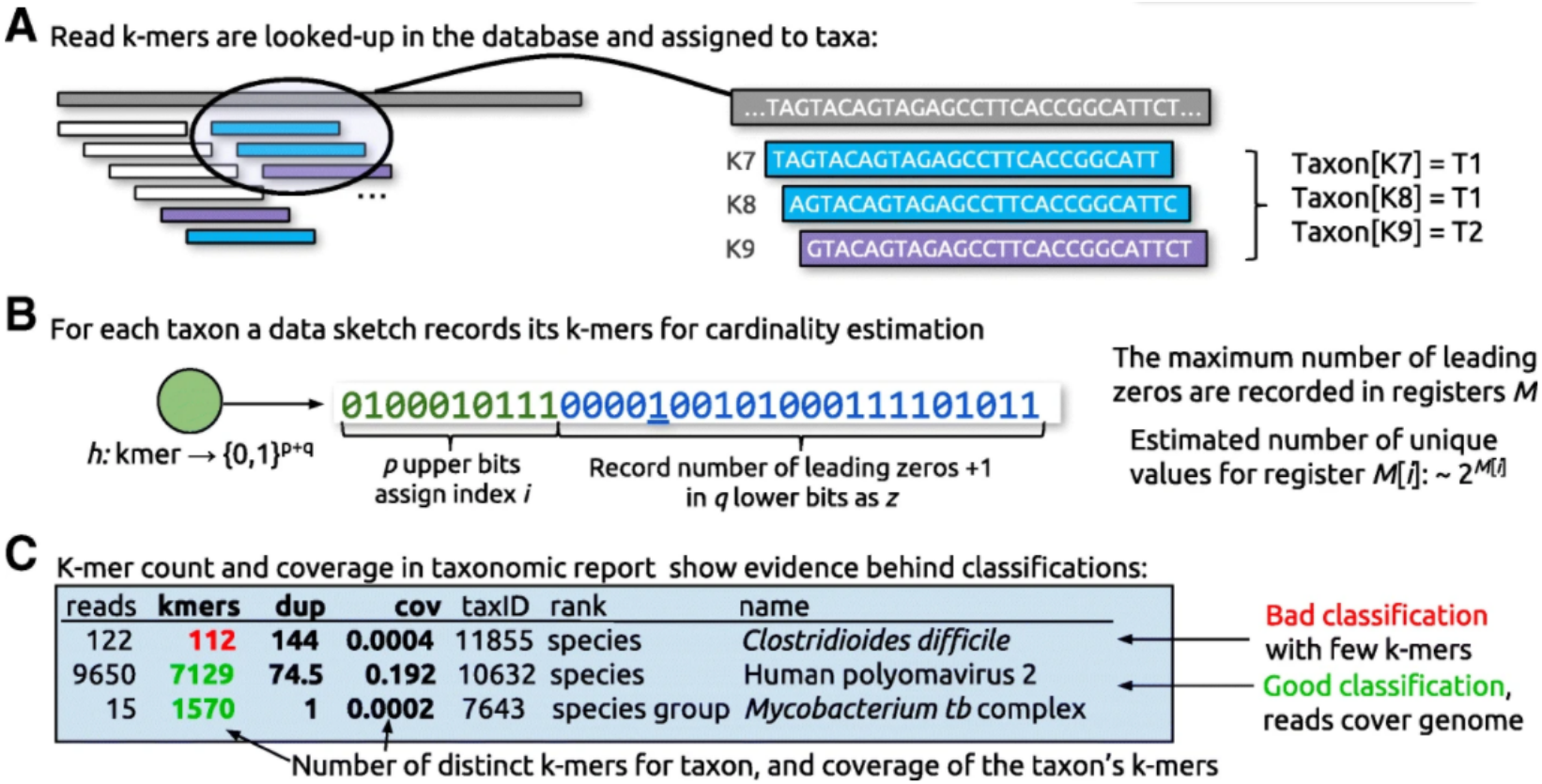
Generating taxonomic profiles for samples

KrakenUniq (formerly KrakenHLL) is a novel metagenomics classifier that combines the fast k-mer-based classification of Kraken with an efficient algorithm for assessing the coverage of unique k-mers found in each species in a dataset.

On various test datasets, KrakenUniq gives better recall and precision than other methods and effectively classifies and distinguishes pathogens with low abundance from false positives in infectious disease samples.

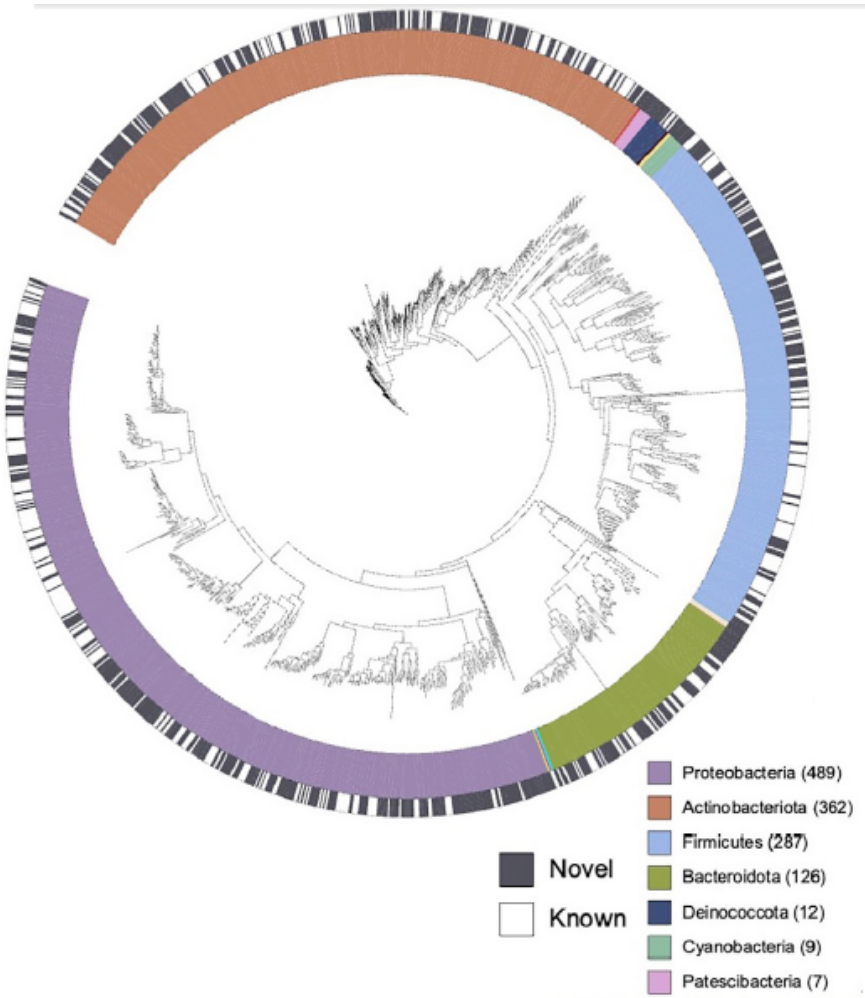
Note: KrakenUniq requires a huge amount of primary memory (ideally 128-512 GB). For performing more memory efficient classification, consider using Centrifuge (ideally requiring 4-12 GB).

Overview of the KrakenUniq algorithm



Source: Genome Biology, 19:198, 2018.

Taxonomic tree for metagenome-assembled genomes (MAGs)



Dimensionality Reduction (UMAP vs t-SNE vs PCA)

UMAP	**t-SNE**	**PCA**
1. Uniform Manifold Approximation and Projection	1. t-distributed Stochastic Neighbourhood Embedding	1. Principal Component Analysis
2. Unsupervised	2. Unsupervised	2. Unsupervised
3. Non-linear	3. Non-linear	3. Linear
4. Not deterministic	4. Not deterministic	4. Deterministic
5. Captures the global structure of data	5. Captures the local structure of data	5. Captures the global structure of data
6. Relatively faster	6. Relatively slower	6. Relatively faster

Analysis with UMAP

UMAP is a dimension reduction technique that can be used for visualisation of data in a lower dimension. The algorithm is founded on three assumptions about the data:

1. The data is uniformly distributed on Riemannian manifold
2. The Riemannian metric is locally constant (or can be approximated as such)
3. The manifold is locally connected

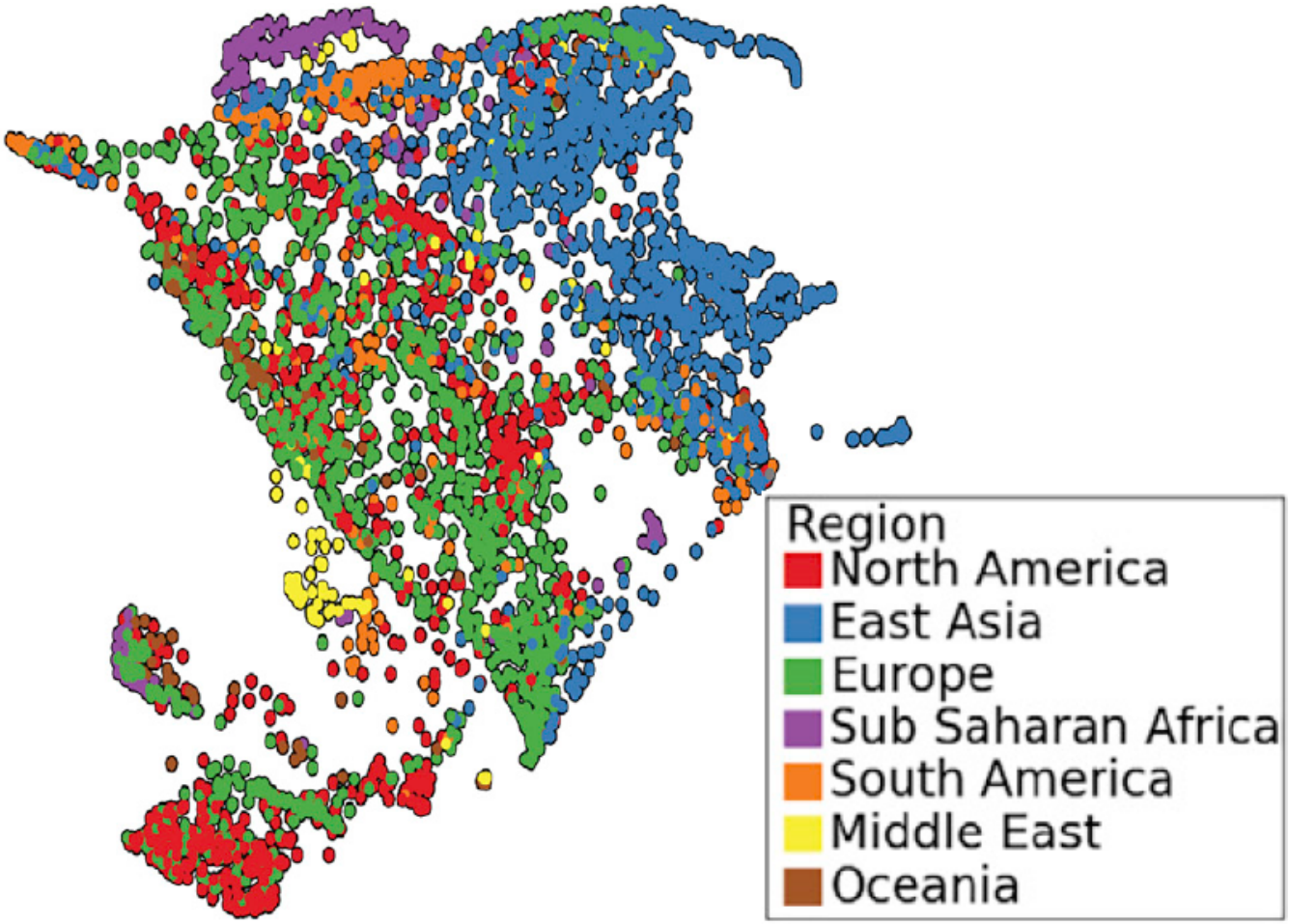
Analysis with UMAP

UMAP is a dimension reduction technique that can be used for visualisation of data in a lower dimension. The algorithm is founded on three assumptions about the data:

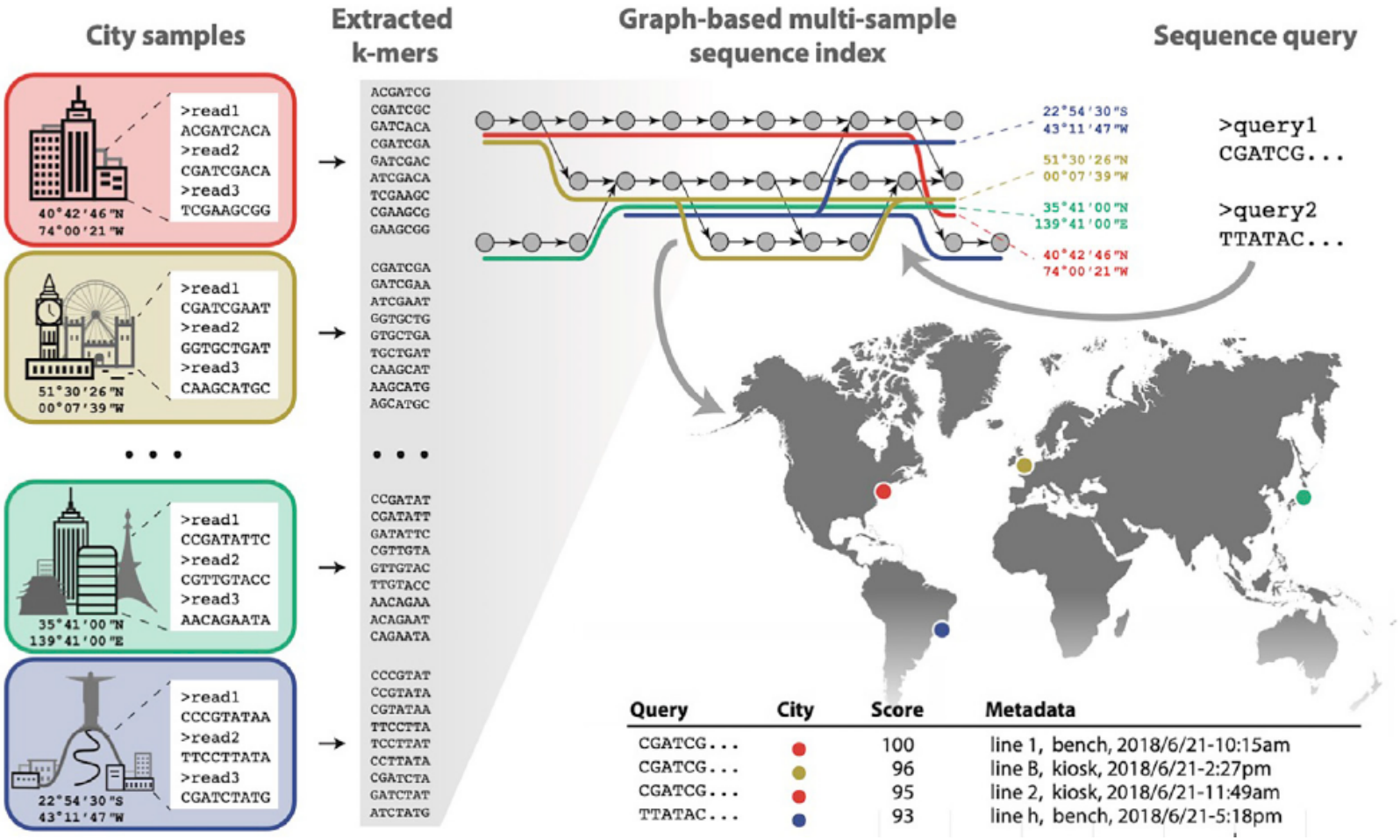
1. The data is uniformly distributed on Riemannian manifold
2. The Riemannian metric is locally constant (or can be approximated as such)
3. The manifold is locally connected

From these assumptions, it is possible to model the manifold with a fuzzy topological structure. The embedding is found by searching for a low dimensional projection of the data that has the closest possible equivalent fuzzy topological structure.

Analysis with UMAP

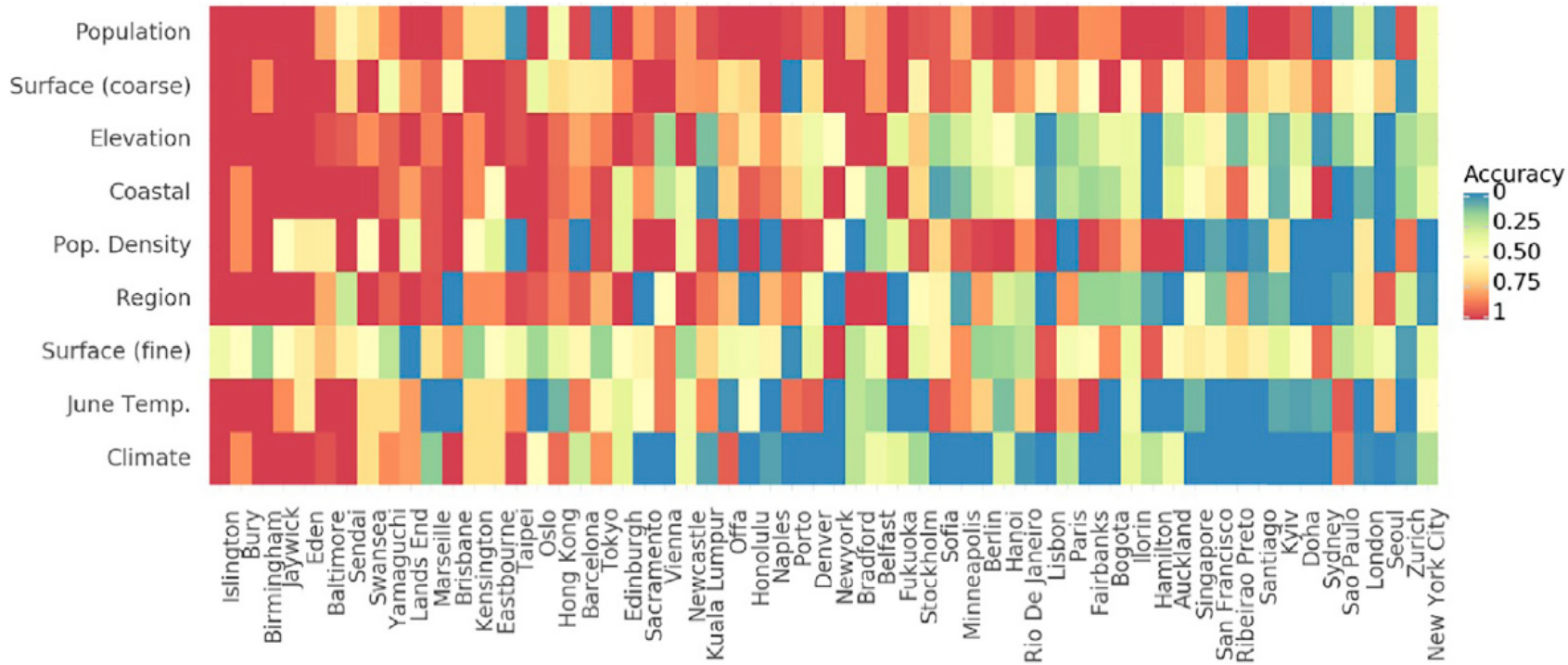


Microbial signatures



Schematic of GeoDNA representation using indexing on graphs

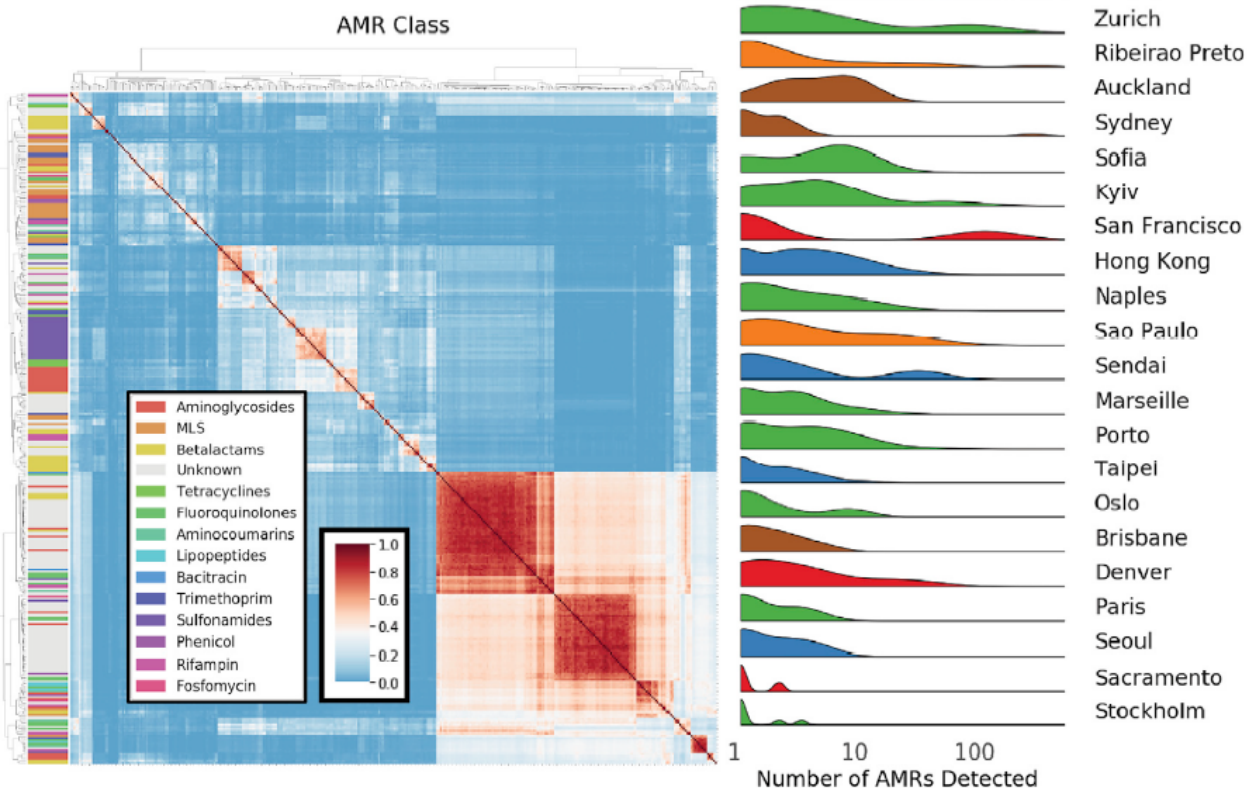
Prediction of features



Prediction accuracy of a random forest model for a given feature

Antimicrobial resistance

Using the MegaRES ontology and alignment software, one can map reads to known antibiotic resistance genes.



Co-occurrence of AMR genes (left), AMR genes by city (right)

Software and algorithms

Tool	Availability
AdapterRemoval v2.17	https://github.com/mikkelschubert/adapterremoval
Bowtie2 v2.3.0	https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.3.0
BLASTn	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST
KrakenUniq v0.3.2	https://github.com/fbreitwieser/krakenuniq
Centifuge	https://github.com/infphilo/centrifuge
MASH v2.1.1	https://github.com/marbl/Mash
HUMAnN2	https://pypi.org/project/humann2
DIAMOND v0.8.36	https://github.com/bbuchfink/diamond
metaSPAdes v3.8.1	https://github.com/ablab/spades/releases/tag/v3.8.1
MegaRes v1.0.1	https://megares.meglab.org/download/index.php
MetaBAT2 v2.12.1	https://anaconda.org/ursky/metabat2
CheckM v1.0.13	https://github.com/Ecogenomics/CheckM
dnadiff v1.3	https://github.com/mummer4/mummer
GTDB-Tk v1.0.2	https://github.com/jianshu93/GTDB_Tk
FastTree v2.1.10	https://anaconda.org/bioconda/fasttree
iTOL v5.5	https://itol.embl.de

References

1. D. Danko et al., A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell, 184(13):3376-3393, 2021.