# Microbiome Geographic Population Structure

Chandrashekar CR

28th March 2025

## 1 Abstract

Over the past decade, large microbiome projects have provided a wealth of sequencing data, revealing that microorganisms are distributed unevenly across different regions. This raises important questions such as:

- How do these microbes move?
- How do antimicrobial resistance genes spread?

To answer these questions, we must first identify which microbes are local to an area and which ones have originated elsewhere. However, existing tools lack the accuracy required to map microbes to specific locations, limiting their practical applications in fields such as medicine and ecology.

In this study, the researchers analyzed microorganisms from urban areas, soil, and marine environments. They discovered that certain microbial species are specific to certain regions, making them useful as *biogeographical markers*. Utilizing this insight, they developed a machine learning tool called **Microbiome Geographic Population Structure (mGPS)**. This tool uses the relative abundances of microbial sequences to determine the origin of a sample.

The mGPS tool achieved impressive results:

- It correctly identified the city of origin for **92%** of samples.
- Within cities, it pinpointed the specific location for **82%** of samples, sometimes within a few hundred meters.
- For soil samples, it was **86%** accurate, and for marine samples, **74%** accurate.

Additionally, mGPS distinguished between local and non-local microorganisms and traced the spread of antimicrobial resistance genes across the globe. Its ability to pinpoint locations (such as water bodies, countries, cities, and transit stations) opens new possibilities for tracking microbiomes in forensics, medicine, and epidemiology.

# 2 Introduction

## 2.1 What is Trace Evidence Analysis?

- A forensic discipline that examines tiny material transfers (e.g., hair, fibers, soil) between objects, people, and environments.
- These transfers can link individuals to specific locations, objects, or events.

## 2.2 Challenges in Tracing Geographic Origins of Organisms

- Identifying biological material to link individuals with locations is difficult due to:
  - Complexity of biological interactions.
  - Lack of dynamic data on recent movements.
- Human DNA is static—it identifies origins but doesn't track recent movements.
- Eran's GenoChip tool.

## 2.3 Importance of Microbiomes

- Microbiomes (bacteria, fungi, viruses, etc.) provide dynamic, spatiotemporal information that changes with an individual's environment.
- Microbiomes can be used as biogeographical markers for tracing movements.

## 2.4 Broader Applications of Microbial Biogeography

- **Forensics**: Predicting the geographic origin of samples.
- **Ecology**: Understanding biodiversity and environmental influences on microbial communities.
- **Medicine**: Addressing the spread of antimicrobial resistance (AMR), a major global challenge.
- **Policy Development**: Informing guidelines to reduce AMR risks in human mobility, trade, and food distribution.

## 2.5 Antimicrobial Resistance (AMR) and Microbial Tracing

- AMR is spread through human travel, migration, and global trade of goods (e.g., food and animals).
- Example: AMR bacteria spread through trade routes, like illegally traded species, can be traced using microbiome data.

## 2.6 Introducing mGPS

- A machine learning–based tool to identify the fine-scale geographic origin of microorganisms using microbial relative sequence abundances (RSAs).
- Tested on urban, soil, and marine microbiomes.
- mGPS successfully distinguishes local from nonlocal microorganisms and traces global AMR gene spread.

## 2.7 Why this matters?

- Understanding microbial biogeography helps manage AMR risks, improve forensics and epidemiology and study human environment interactions.

# 3 Results

## 3.1 The mGPS Tool Implementation

1. **Goal of mGPS:** The goal is to predict the geographic origin of microbiomr samples using microbial relateve sequence abundances (RSA) and metadata.
2. **Key Features of mGPS:** It does not require users to tweak complex setting. It mainly focuses on Geographically Informative Taxa. These are sepcific microbial taxa that show unique patterns in certain regions, making them useful for location predictions.
3. **Process for building the model:**
   (a) **Selecting the Best Taxa:**
      i. A method called **recursive feautre elimination** is used to rank microbial taxa based on how useful they are for predicting geographic locations.
      ii. Only the most informative taxa (GITs) are kept for building the model.
   (b) **Training the model:**
      i. The model uses RSA data (the proportion of sequences belonging to each taxon).
      ii. It starts by predicting broad regions like continents.
      iii. Each prediction is then added as extra input to train models for finer levels (e.g., countries, latititude and longitutde)
      iv. This step-by-step approach helps the model to get better at predicting exact coordinates.
   (c) **Final Output:** The model predicts the latititude and longitutde of the microbiome samples.
4. **Testing mGPS:** To check how well mGPS works, it was tested on different types of microbiome data.
   (a) Urban biome: Data from citites around the world. (Next-generation data)
   (b) Soil biome: Data from soil microbes collected globally. (16srRNA data)
   (c) Marine biome: Data from ocean microbes. (Shotgun Sequencing data)
5. **Ensuring Reliability:** Data came from different sequencing methods and platforms, esuring the tool is not tied to specific ways of collecting of proecessing data. The rigorous testing showed that mGPS works across various datasets.

# 4 Materials and Methods

## 4.1 Global Datasets

## 4.2 Microorganism DIstribution and Pathogenicity

## 4.3 mGPS Implementation

### 4.3.1 QC Procedure

- **Objective:** To identify informative, discriminating, and independent microbial taxa (Geographically Informative Taxa or GITs) for accurate geographic location prediction.
- **Data Preparation:** Datasets, encompassing microbial relative sequence abundance (RSA) from various geographic origins (cities, stations, countries, oceanic bodies) were utilized. Each dataset was randomly partitioned into training (80%) and testing (20%) subsets.
- **Recursive Feature Elimination (RFE) with Random Forest:** An initial random forest classifier was trained on the training data, using all avilable taxa as predictors. The predictive accuracy of th classifier was assessed using the testing data. Out-of-bag (OOB) prediction accuracy was recorded, and feature importance was calculated by permuting each variable and measuring the impact on prediction accuracy. Variables were ranked in descending order of importance based on this metric.
- **Iterative Subset Selection:** Five subsets of predictors variables () were iteratively created, representing the $i$ most important variables. The model was re-trained for each subset, and predictive accuracy was evaluated on the testing set. This process was repeated five times with different training/testing splits to mitigate selection bias.
- **Optimal Subset Determination:** The subset size that yielded the highest average classification accuracy was selected as the optimal size. The optimal number of features varied by dataset, for example 200 features were optimal for sail and marine data.
- **GIT Identification:** The most infromative variables, corresponding to the $i$ highest average importance values were designated as GITs.
- **Variable importance visualization:** Variable importance plots were generated to illustrate the contribution of each GIT to model prediction accuracy.
- **mGPS integration and user customization:** The feature selection process is integrated into the mGPS framework.

### 4.3.2 Model Training

1. **Hierarchial Geographical Prediction Framework:**
   - mGPS is desgined to exploit the inherent hierarchical structure of geographic data. This structure, exmemplified by continent, country, city, station and precise coordinates (latitude and longitude), provides a natural ordering for prediction.
   - The model employs a chained series of gradient-boosted decision trees (GBDT) to navigate this hierarchy, progressively refining location predictions.
2. **Chained Submodel Architecture:**
   - The model comprises a series of interconnected submodel (M1,M2,M3,M4) each dedicated to predicting a specific level of geographical hierarchy.
   - This chaining mechanism allows the model to leverage dependencies between hierarchical levels. For instance, the predicted continent from M1 influences the subsequent prediction of the country or city by M2, and so on.
   - For local studies, where the continental level is not relevant, the M1 submodel is excluded.
3. **Input Data and Feature Integration:**

- The initial input to M1 is a vector containing the relative sequence abundance (RSA) of the selected Geographically Informative Taxa (GITs).
- Subsequent submodels (M2, M3, M4) recieve as input the original RSA vector augmented with the predicted class probabilitis from the preceding submodel. This iterative input augmentation ensures that the model incorporates infromation from previous prediction stages.

4. **Submodel Functionality and Prediction Types:**
    - M1 is a classification model that predicts the highest-level geographical category (e.g., continent).
    - M2, also a classification model, predicts the next level of granularity (e.g., country, city or transit station.)
    - M3 and M4 are regression models that predict the continuous latitude and longitude, respectively.
5. **Gradient-Boosted Decision Trees (GBDT) with XGBoost:**
    - Each submodel utilizes the GBDT algorithm, a powerful ensemble learning methods that builds decision trees iteratively.
    - The XGBoost library, an optimized implementation of GBDT, is employed to enhance computational efficiency and model perfromace. XGBoost is known for its speed and ability to handle large datasets.
    - GBDT operates by sequentially constructing decision trees, where each subsequent tree aims to correct the erros made by its predecessors. This iterative process refines the model's predictive accuracy.
6. The XGBoost algorithm requires careful tuning of several hyperparameters, including:
    - learing rate
    - maximum depth of trees
7. The final model performance is evaluated based on predictions generated for the unseen test datasets.
8. All downstream analyses, including the analysis of AMR transfer patterns, are performed exclusively on these unseen test datasets, ensuring unbiased results.
9. The entire training procedure, including hyperparameter optimization, is repeated for each dataset indenpendently to account for dataset-specific characteristics and optimze performance.