Final Research Report

Course Code: BIOX7011  
Student Name: Yu Sun  
Student Number: 45105764  
Supervisor: [Jan Engelstaedter, Affiliation]  
Date: [Submission Date]

# Abstract

This section provides a concise summary of the entire project (200–300 words). It should clearly outline the research aims, the main methodologies used, the key findings, and the major conclusions derived from the study.

# 1. Introduction

Antimicrobial Resistance (AMR) has become one of the most pressing threats to global public health in the 21st century. The World Health Organization estimates that by 2050, more than 10 million people will die annually from AMR-related infections [1].AMR not only makes it more difficult to treat infections, but also significantly raises healthcare costs and length of hospitalization, placing greater pressure on public health systems in low-income countries [2]. Among the many drug-resistant pathogens, drug-resistant tuberculosis (DR-TB) is of particular concern. approximately 450,000 people worldwide will have rifampicin-resistant tuberculosis (RR-TB) in 2022, with the majority of cases also showing resistance to isoniazid, thus constituting multidrug-resistant tuberculosis (MDR-TB) [3].

# Although Mycobacterium tuberculosis is the most intensively studied model for rifampicin resistance, rifampicin is exposed to a wide range of bacteria, both in clinical use and through environmental contamination. Therefore, this study aims to predict cross-species mutation patterns, which may provide insights into the evolutionary convergence of antibiotic resistance.

# Materials and Methods

Step 3.supervisedMl

Data and Representation

Object: Multi-species rpoB site-level mutation matrix (species × mutation; binary values/frequency are unified into binary values: presence = 1, absence = 0).

Filtering: Remove extremely low-frequency/single-sample mutation sites; retain a sufficiently informative site set (used for the "Full Mutation Map" and "Top-30 Mutation Map" visualizations).

Confounding Assessment and Stratification (confounder score → high / mid-high)

Calculate the confounder score for each species (based on comprehensive metrics such as known source bias, sequencing depth/sample size, and publication bias).

Stratification by Thresholding:

High group: High confounder score (confounder score >0.7 ,potentially more biased, independent stability assessment first).

Mid-high group: Second-highest confounder score (0.3<score< 0.7,serves as a control stratum for analysis parallely).

Clustering Strategy Grid (algorithms × distances)

Algorithms: K-means, GMM, hierarchical clustering (HDBSCAN is optional for robustness testing).

Distance/Similarity: Euclidean, Cosine, (optional) Jaccard/Manhattan.

2D Grid Combination: Run the full combination for both the high and mid-high groups.

Model Selection and Stability Assessment

Primary Score: Silhouette (cosine distance is preferred for evaluating "pattern similarity" for binary/sparse data).

Stability: Re-run concordance of subsampling/bootstrapping (median and IQR of NMI/ARI can be reported).

Determining the Top Three: Sort by primary score. If the scores are close, weight the decision based on stability and interpretability (biological plausibility).

**Visualization and Comparison**

Heatmap 1 (Full Mutation): A binary matrix of species × all loci; rows are annotated with cluster groups, and columns are loci.

Heatmap 2 (Top-30): Top 30 sites filtered by overall frequency/information gain (easier to identify patterns).

**Upsetplot 3(top10)&all:**

Step4.

## ***1. Data Preparation***

## Known RIF-resistance mutations are used as positive samples (Positive).

## Unknown/unreported mutations are used as unlabeled samples (Unlabeled).

## A binary matrix (X\_dense) of species × mutations is formed.

## ***2. Model Selection and Training***

## Random Forest is used as the base classifier for supervised learning.

## Positive samples and unlabeled samples are combined using the PU-learning framework to build a model.

## During training, some known mutations are randomly masked to evaluate the model's resilience.

## ***3. Validation Strategy: Mask-then-Recover***

## Mask some known mutations.

## Use the model to predict candidate mutations.

## Calculate Recall@K to assess whether the masked mutations can be recovered in the top K predictions.

## ***4. Candidate Mutation Generation***

## **Top-K Strategy:** Select the top K mutations with the highest predicted probability for each species.

## Threshold Strategy: select mutations with p\_true ≥ τ(0.7) Generate a list of candidate mutations for subsequent analysis and experimental verification.

5. Novelty Filtering

Definition of novel. A mutation is “novel” if it does not appear in our compiled set of previously observed non-lab mutations after mapping all records to a unified E. coli rpoB amino-acid coordinate

Construction of the non-lab set. We parse the Google Sheet (Origin ≠ “Lab mutant”), map records to E. coli coordinates and build a blacklist S of observed mutations (two interchangeable modes):

**Global de-duplication**: S = {Mutation} across all species (default).

**Species-specific de-duplication** (optional): S = {(Species, Mutation)}.

Filtering. We do not retrain the model. We filter the scored candidates by excluding anything in S.

For Top-K, we refill from the remaining pool (highest p\_true) until reaching K or the pool is exhausted.

For Threshold, we simply keep items with p\_true ≥ τ and Mutation ∉ S.

# Results

作者对物种×突变矩阵（X\_dense）在三种距离度量（Euclidean、Manhattan、Cosine）与四种聚类算法（HDBSCAN、k-means、DBSCAN、GMM）的所有组合进行了系统比较。每个组合首先在统一参数（n\_neighbors=15, min\_dist=0.3, seed=123）下通过 UMAP 进行二维嵌入，然后在嵌入空间上执行聚类（HDBSCAN/DBSCAN 使用 minPts 与 eps，k-means 设定 k=4，GMM 由模型自动择优）。为避免噪声干扰，仅在有效样本（簇标签为正且簇数≥2）上计算平均轮廓系数作为质量指标。根据轮廓系数从高到低选取 top\_k 个方案作为 Top methods；若设置 restrict\_metric 则在指定度量内择优，否则在全体组合内筛选。所有方案的评分与标签均导出存档以保证可重复性。  
对入选的 Top methods，我们展示其 UMAP 投影及对应的聚类热图（全突变与 Top-30 版本），以对比不同算法在突变谱空间得到的分群一致性与差异性。

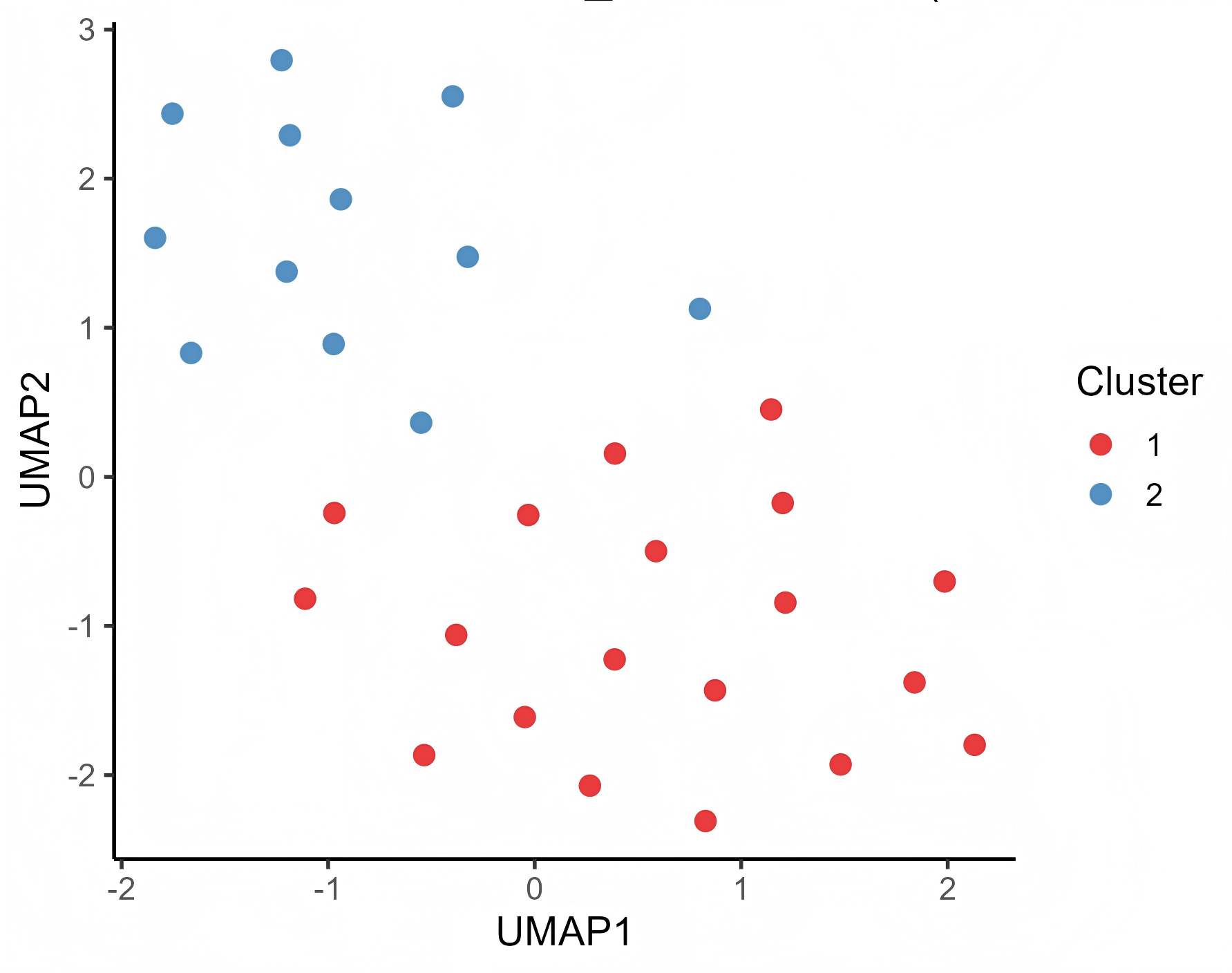
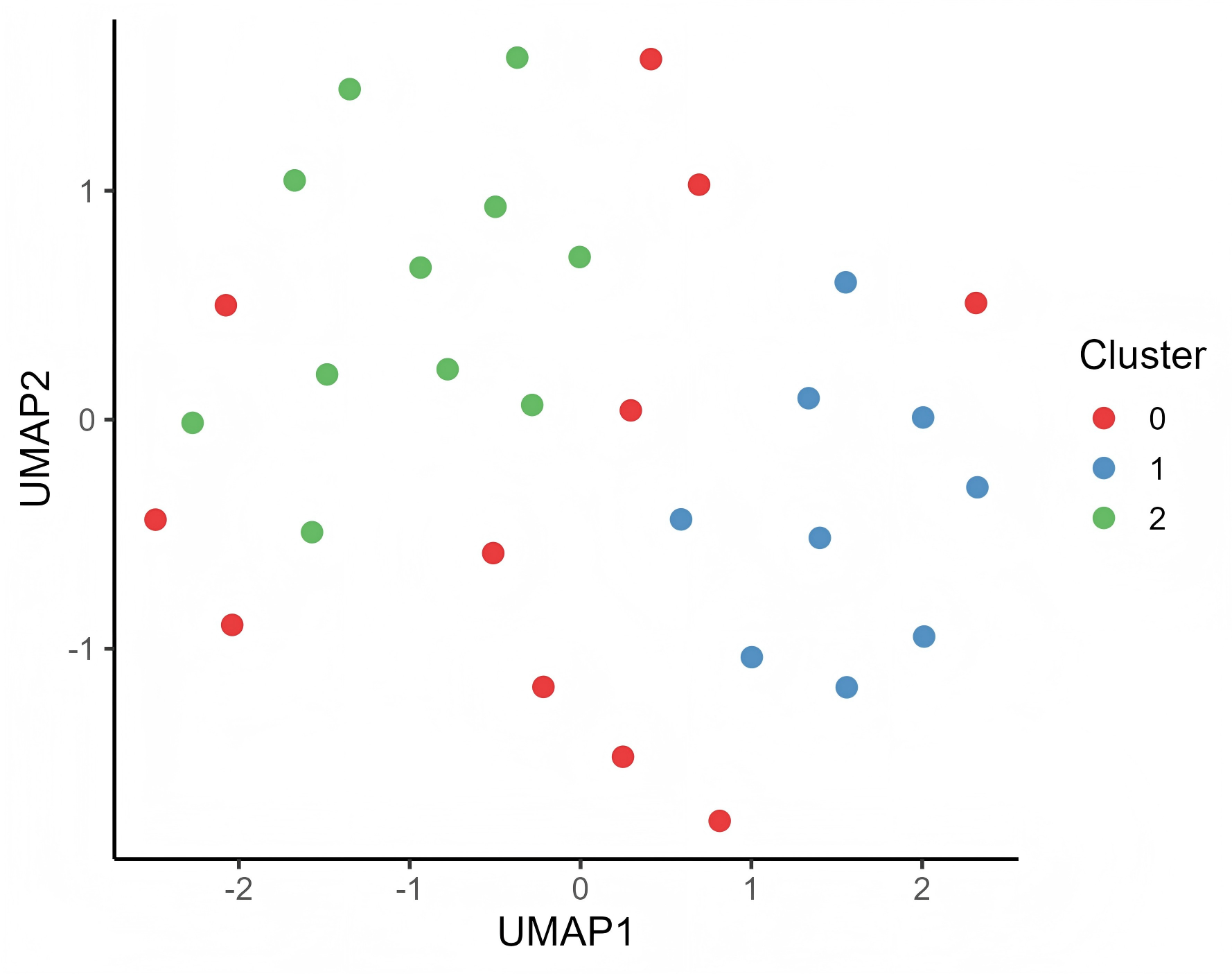
基于 X\_dense\_midhigh 矩阵的参数，本次研究对物种间突变分布模式进行了多算法聚类与可视化分析。为验证聚类结果的稳健性与一致性，分别采用三种表现最优的聚类方法：**Cosine–HDBSCAN**、**Euclidean–GMM** 以及 **Euclidean–HDBSCAN**。  
其中，HDBSCAN 算法能在非球形数据中自动识别簇并排除噪声点，而 GMM （高斯混合模型）通过概率密度建模来捕捉数据的潜在连续分布。三种方法均以标准化后的突变存在矩阵为输入，降维可视化部分采用 UMAP （Uniform Manifold Approximation and Projection）算法，以保留样本间的局部拓扑关系。

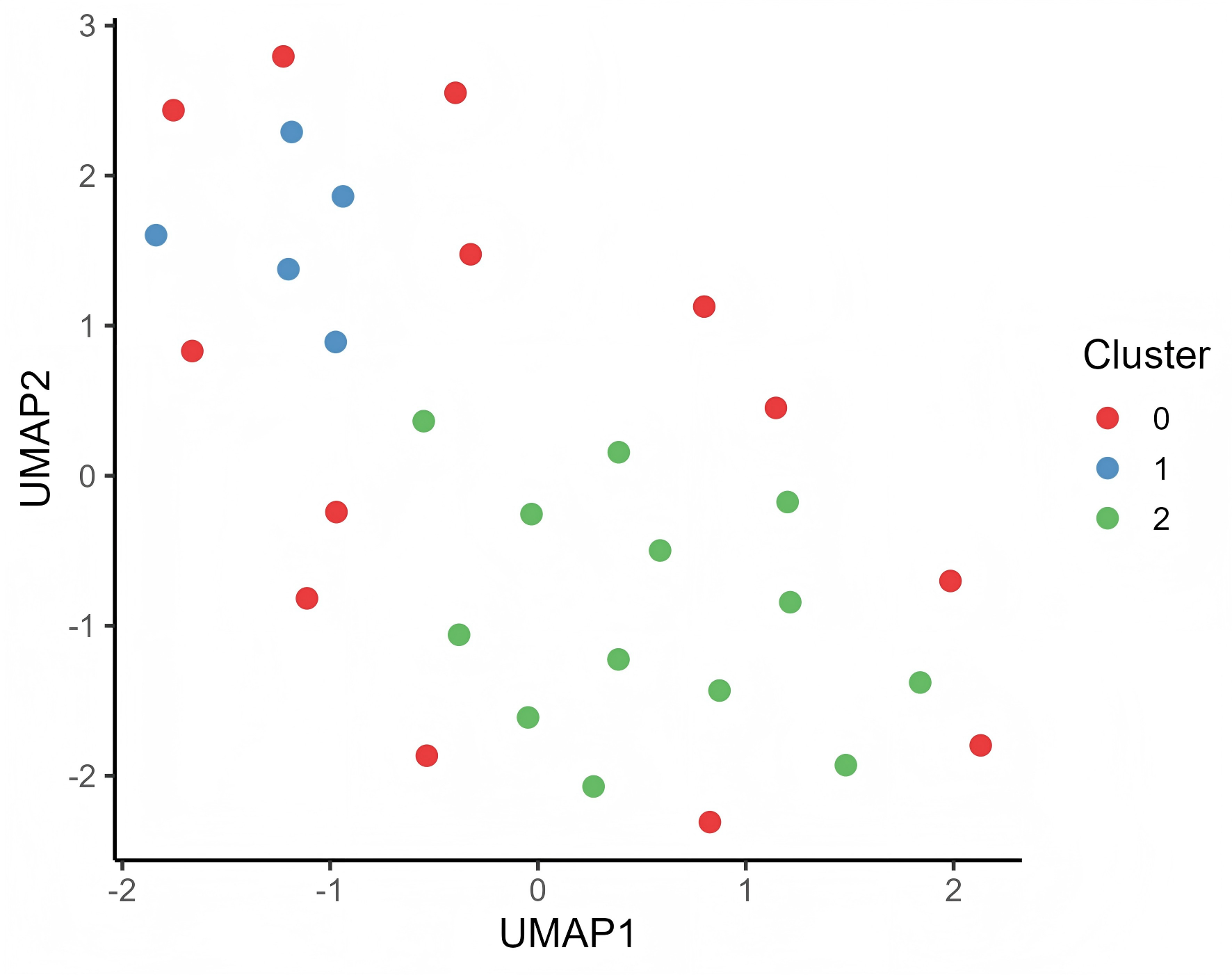
**UMAP 散点图**（图 X–1 至 X–3）展示了三种方法下各物种在二维潜空间中的分布与聚类分界。不同颜色表示不同聚类标签，可以直观反映物种间突变谱的相似性。例如，Cosine–HDBSCAN 能够较好地将高相似突变谱物种归为同簇；Euclidean–GMM 则展现出相对规则、边界清晰的分布；而 Euclidean–HDBSCAN 在处理中间型或噪声物种时表现出更高的分辨度。总体上，三种方法得到的聚类结构高度一致，说明突变模式在不同度量空间下具有稳定可重现性。

**Heatmap 热图**（图 X–4 至 X–6）进一步展示了各聚类内物种的突变共享情况。行表示物种，列表示突变位点，颜色表示突变的存在与否（1/0）。侧边的色条对应 UMAP 聚类结果的簇编号。可以看到，不同簇在突变分布上呈现明显的互补或特异性，例如 Cluster 0 主要集中在 RRDR 核心突变（如 D516、H526、S531 系列），而 Cluster 2 则富集边缘或低频突变（如 Q148R、L533R 等）。这些分群特征表明突变谱具有一定的系统发育与进化分化特征。

Clustering Framework Overview

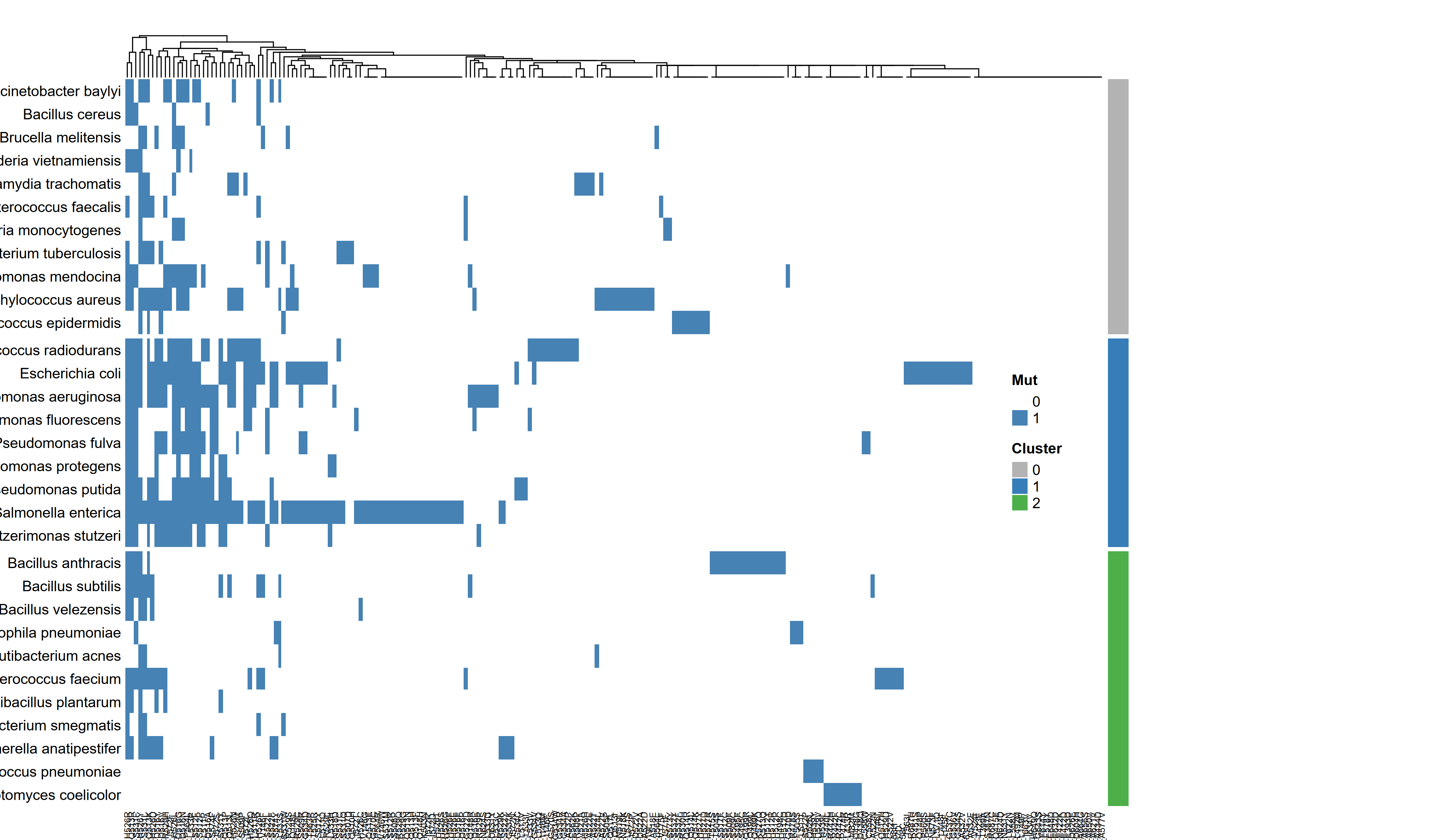
Based on the X\_dense\_midhigh matrix, the authors used three top-performing clustering methods: Cosine–HDBSCAN, Euclidean–GMM, and Euclidean–HDBSCAN. All three methods take a normalized mutation presence matrix as input and employ the Unified Mapping (UMAP) algorithm for dimensionality reduction and visualization, preserving local topological relationships between samples.

The UMAP scatter plots show the distribution and cluster boundaries of each species in the two-dimensional latent space using the three methods, with different colors representing different cluster labels.  




**Mutation Distribution and Intra-Cluster Characteristics**

The heatmap further illustrates the shared mutations among species within each cluster. Rows represent species, columns represent mutation sites, and colors indicate the presence or absence of mutations.

The color bars on the sides correspond to the cluster numbers in the UMAP clustering results. Different clusters exhibit distinct complementarities or specificities in their mutation distributions.  
********

****Comparison of Clustering Patterns at Different Confounder Levels****

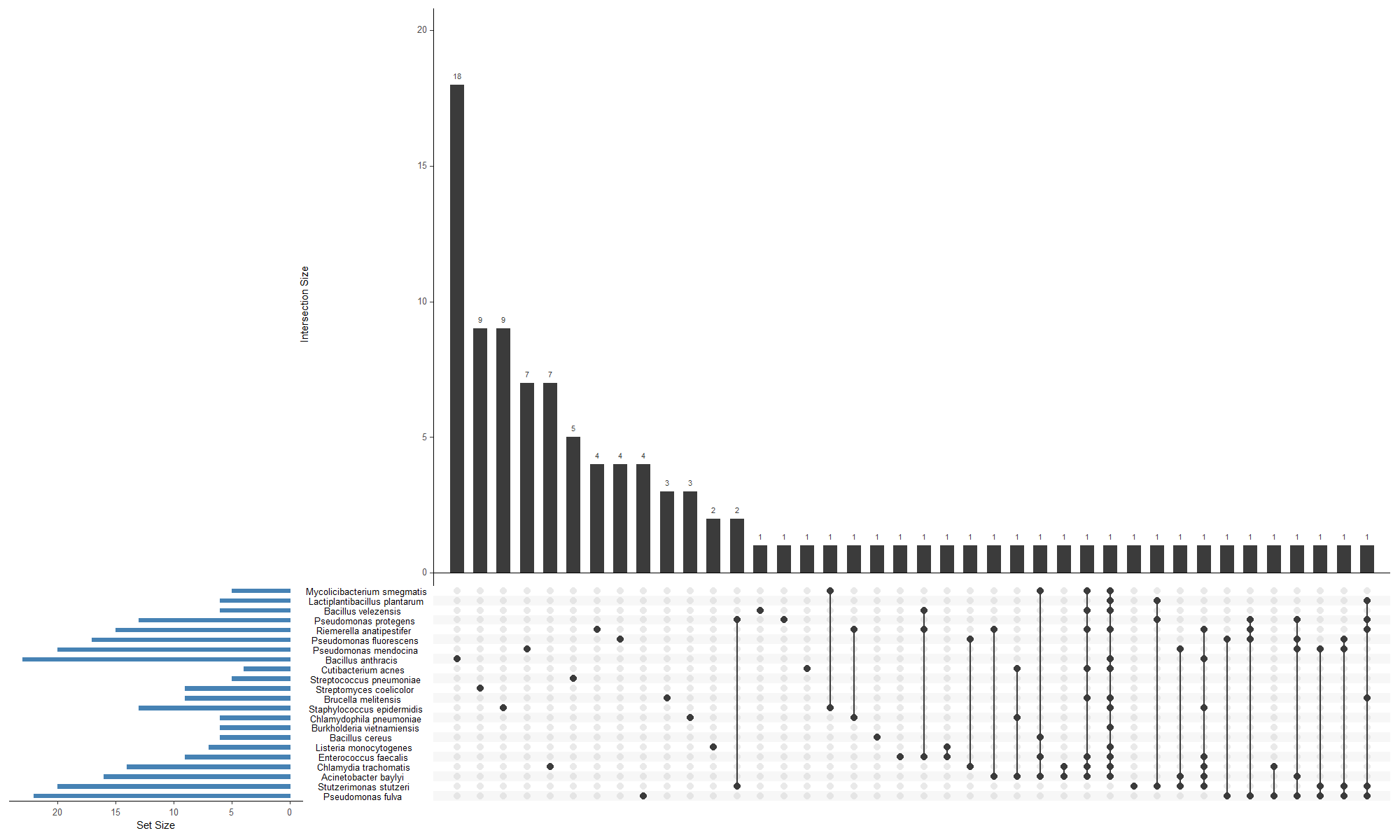
**To ensure that clustering results are not significantly biased towards highly studied species versus less studied ones, the authors specifically compared clustering results for data with high confounder and mid-high confounder levels.**

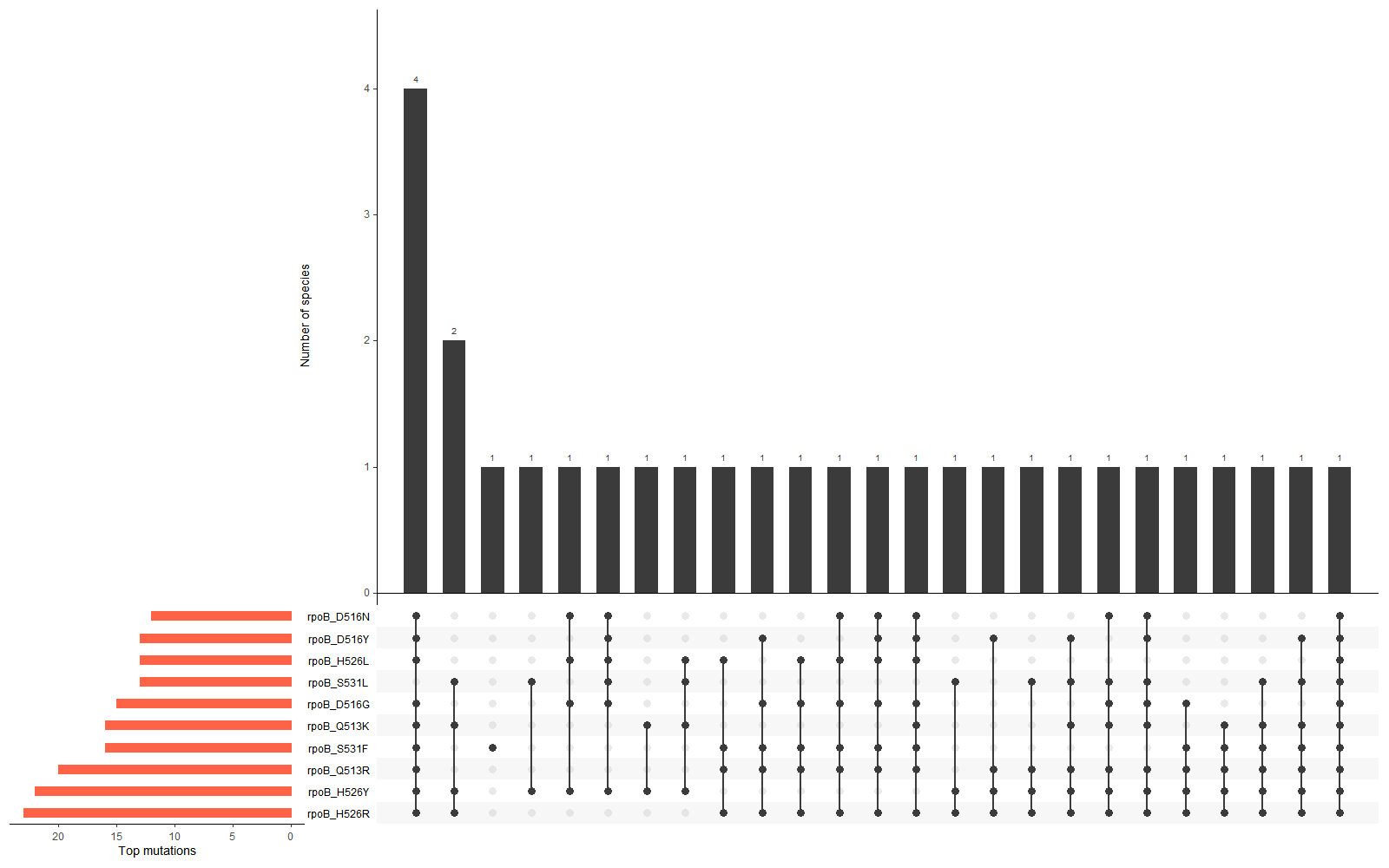
**Despite varying noise levels, the clustering structures obtained by the three methods are generally consistent, demonstrating that mutation patterns are robust and reproducible across different metric spaces.**

****Mutation Intersection Relationships (UpSet Analysis)****

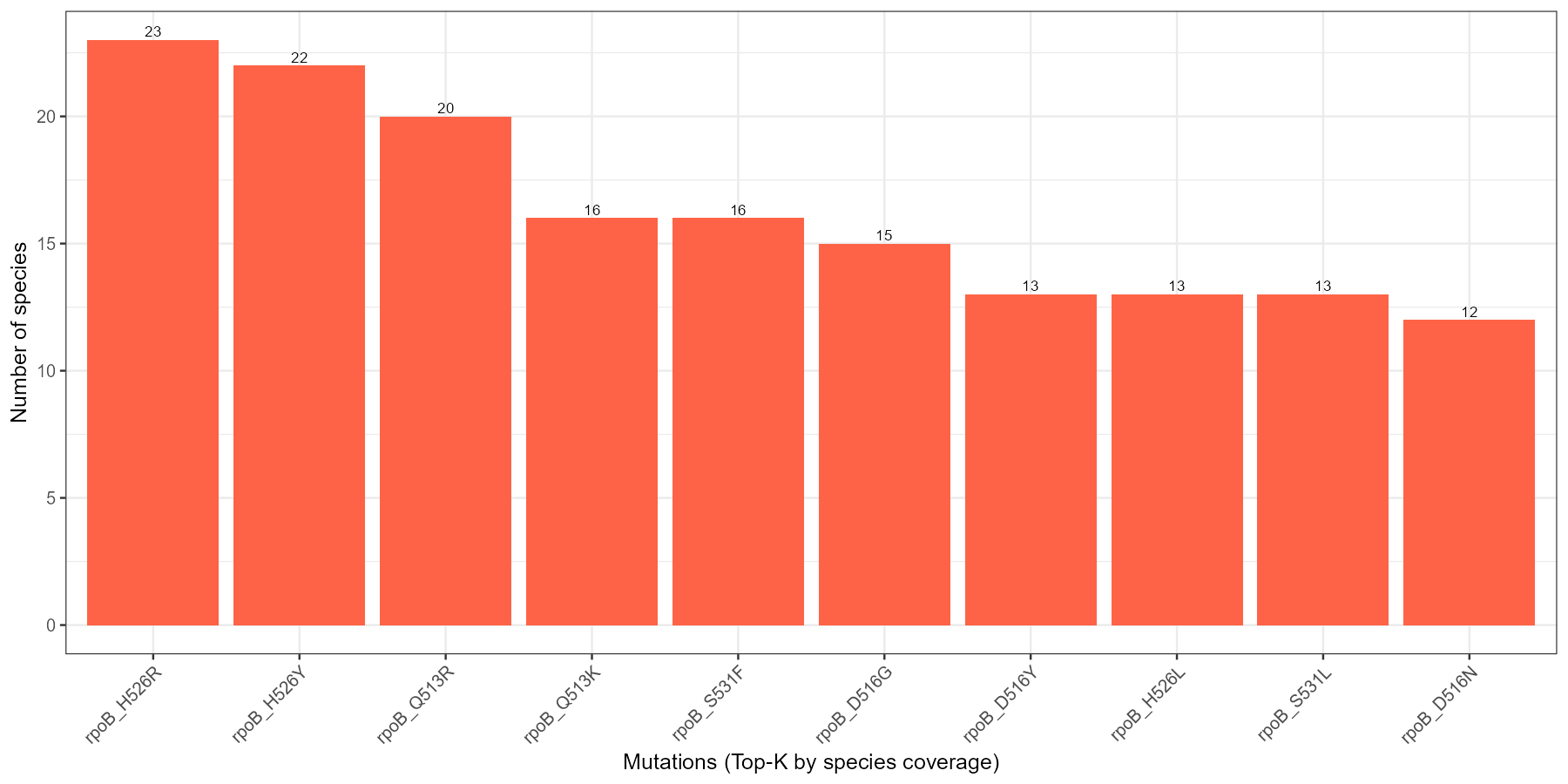
**UpSet plots demonstrate the shared mutation relationships between species. Species are set elements, and the bar graphs represent the size of the intersections.**

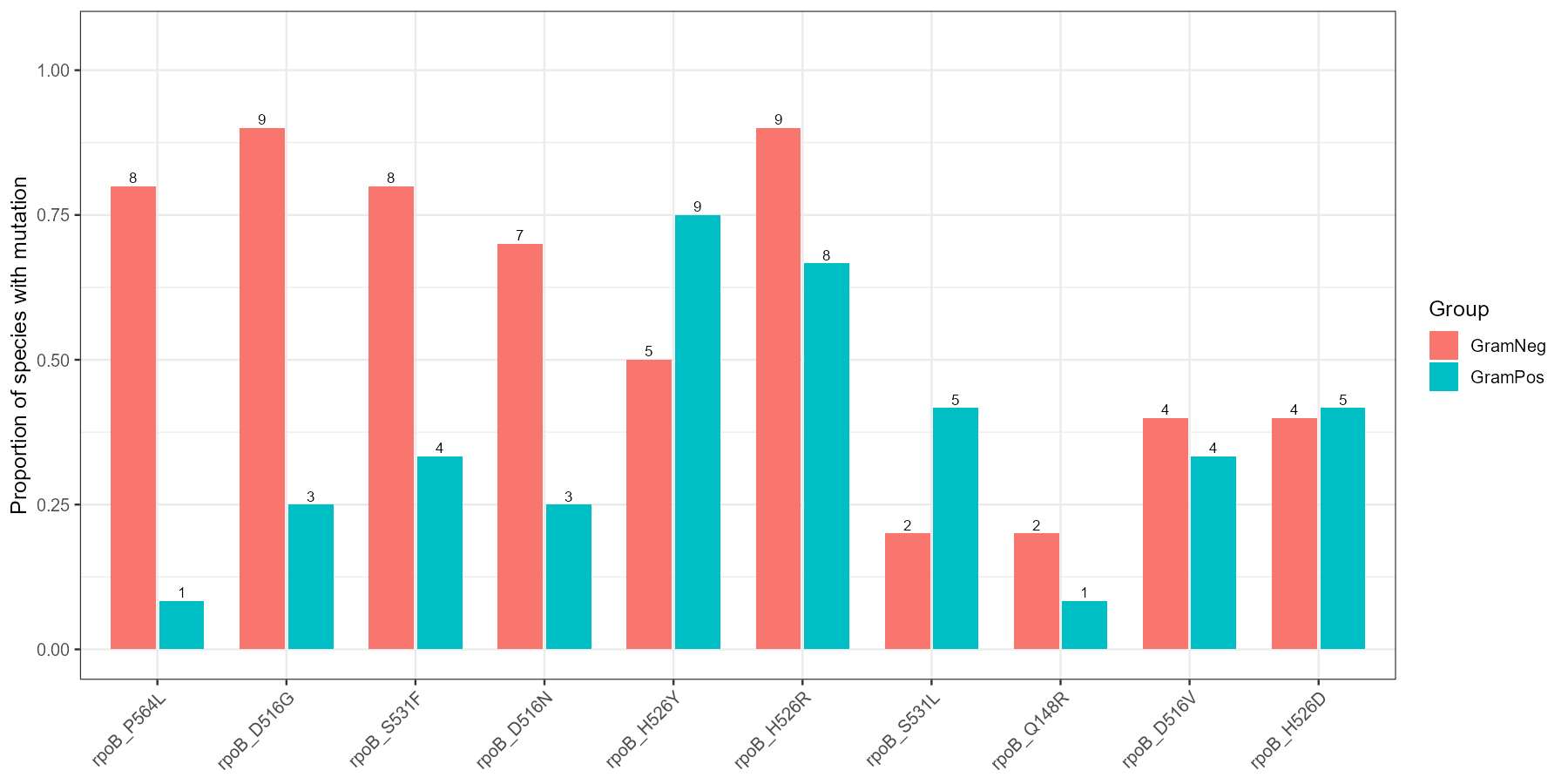
**In the mid-high dataset, some bacterial genera (e.g., Pseudomonas and Bacillus) share multiple high-frequency rpoB mutation sites, suggesting convergence in resistance mechanisms across lineages.**

********

********

****Gram-positive and -negative differentiation analysis****

****

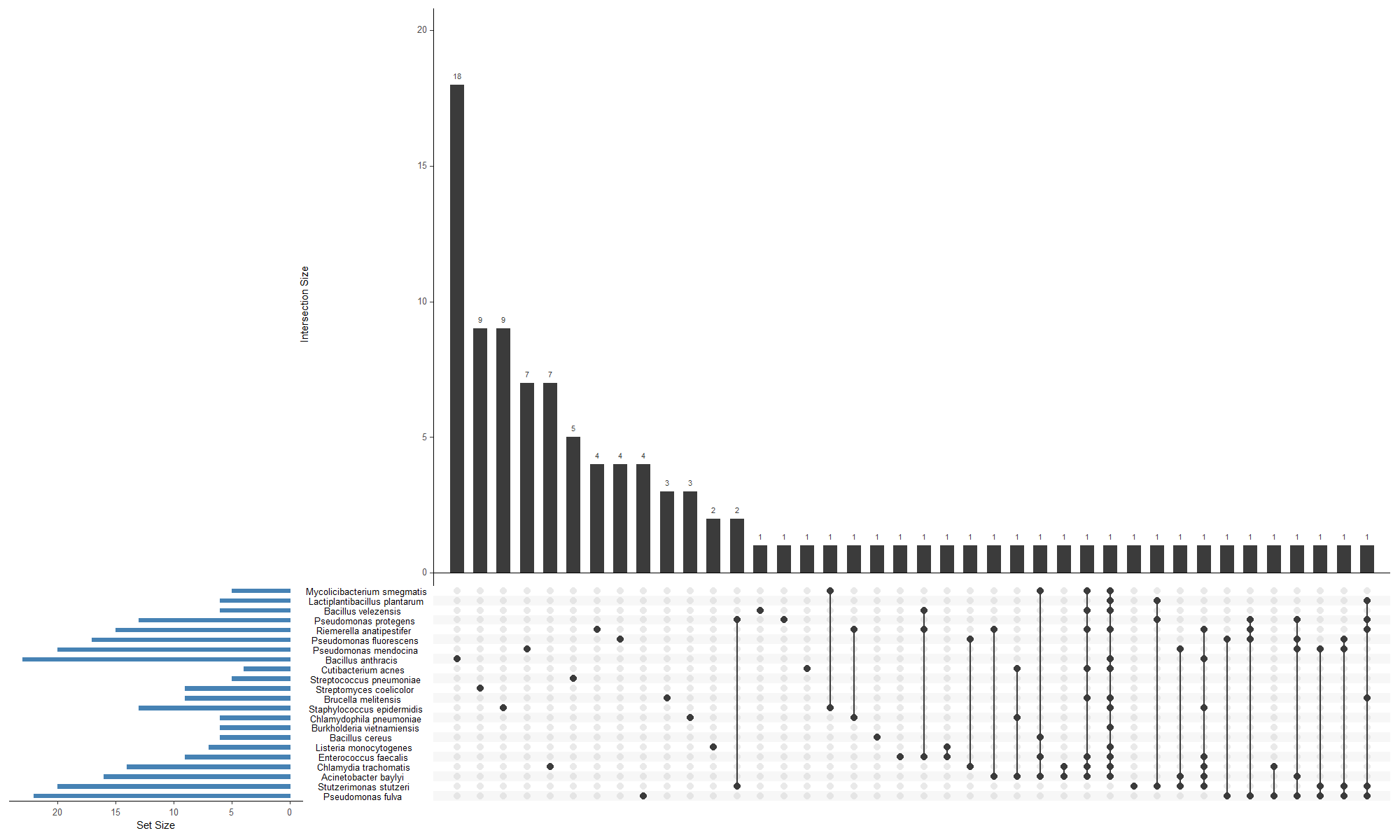
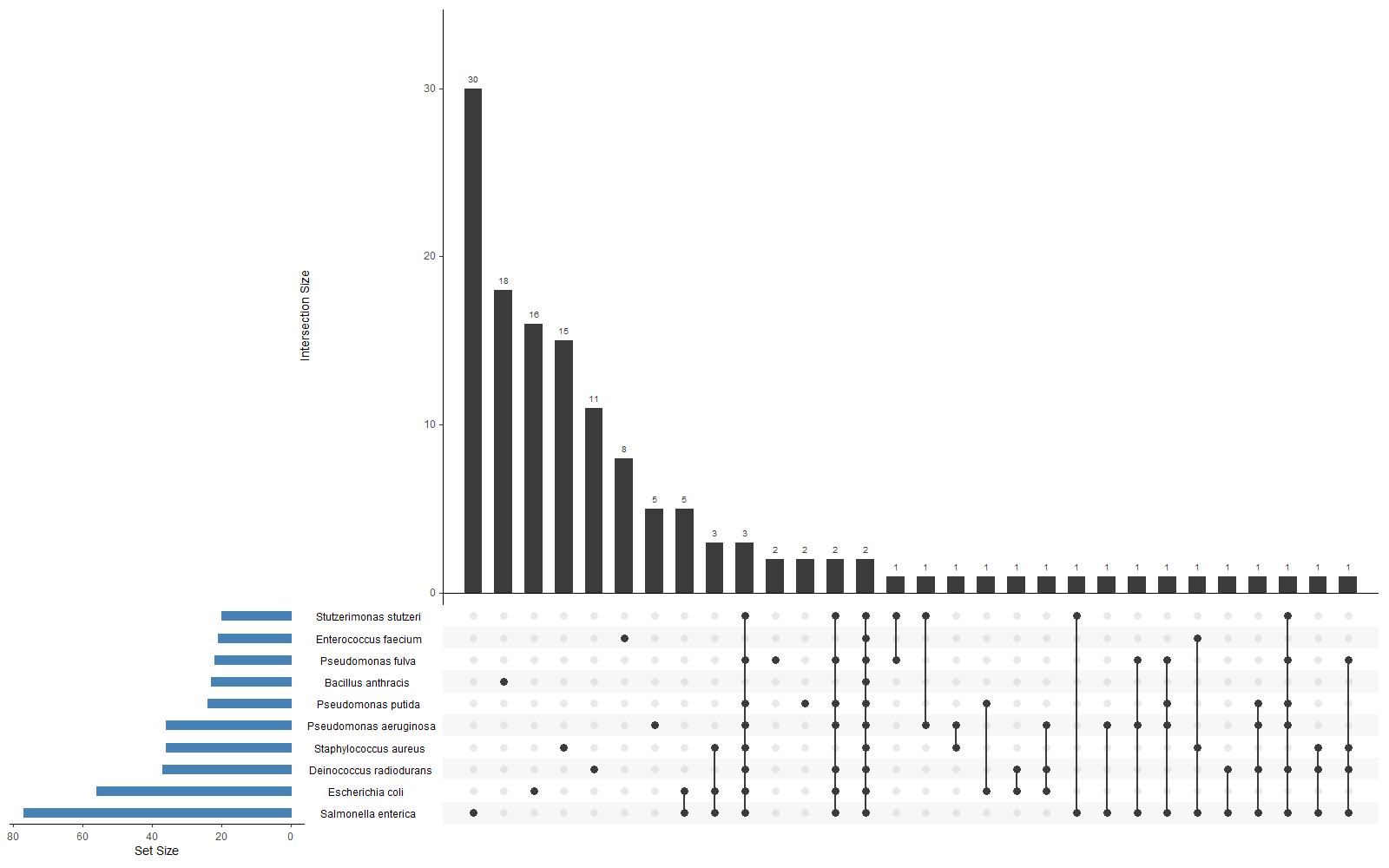
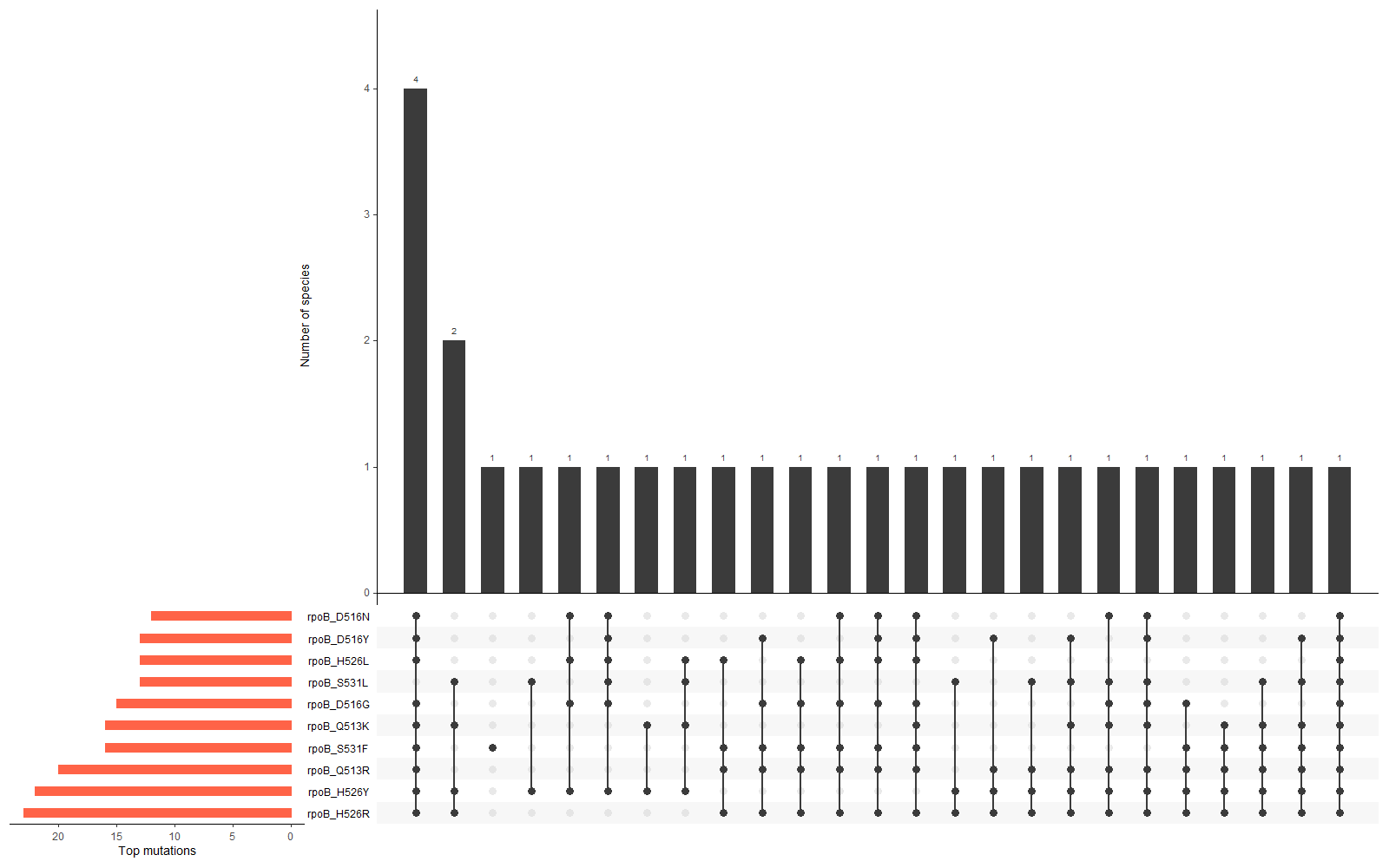
****

**Leveraging the highly recognized Gram-positive/-negative differentiation method, the authors further compared mutation patterns in the mid-high data by Gram staining group (Gram+/Gram–). They found that the two bacterial groups differed in their preferences at typical rpoB loci.**

Most classical RRDR mutations (e.g., P564L, D516G, S531F) were enriched in Gram-negative species, whereas a few peripheral variants (S531L, Q148R) showed relative enrichment in Gram-positive taxa.

This suggests that Gram-negative species rely predominantly on canonical RRDR substitutions conferring strong resistance, whereas Gram-positive taxa accumulate peripheral or compensatory variants that may fine-tune rifampicin susceptibility with reduced fitness costs.

**This grouping trend is consistent with the phylogenetic clustering reported by Bolourchi et al. (2025), indicating that mutational spectrum structure exhibits reproducible evolutionary clustering across species.**



从最佳clustering方法分析， mid-high 数据的最佳聚类是 **cosine–HDBSCAN**，其次是 **euclidean–GMM**。  
这两个算法都倾向于捕捉**密度差异**和**多模态分布**。

从 UpSet 的变化可以看到这种“密度差异”：  
前10物种 → 一两个密度峰（高共享突变）；  
mid-high → 多个平缓峰（局部共性但整体稀疏）。

这说明 UpSet 图的拓宽物种维度本身已经**反映出你的聚类结构在突变层面的稀疏化趋势**。

# 4. Discussion

Interpret the results critically in the context of existing literature. Highlight how the findings support or contradict previous studies, and explore possible explanations. Discuss the implications of the results, the limitations of the study, and suggest directions for future research.

# 5. Conclusions

Summarize the most significant findings and their implications. State whether the research aims were achieved and highlight potential applications or follow-up studies.

# 6. References

All cited literature should appear here in a consistent referencing style (APA, Harvard, or a scientific journal format). Example:  
Foster, L., Mouse, M., & Christ, J. (1972). The effect of hypoxia on free divers. J. Irrep. Res., 23, 490–512.

# 7. Figures and Tables

All figures and tables should be labeled and inserted close to where they are first mentioned in the text. Each figure should include a descriptive caption below, while tables should have a title above.

# 8. Appendices

Include any supplementary materials, extended data tables, code snippets, or detailed methodology not essential to the main text.