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

# 尽管结核分枝杆菌已经是利福平耐药性研究最为深入的模型，但利福平毕竟暴露于多种细菌中，无论是临床使用还是环境污染都确定存在。因此，本文进行了对其解跨物种突变模式的预测，这或许能为抗生素耐药性的进化趋同提供一些参考。

# 传统基于文献的整理和分子表征方法在整合分散在不同物种的数千份抗性报告方面，可扩展性有限。机器学习方法，尤其是自然语言处理和无监督特征提取，提供了一个可扩展的框架，可以系统地揭示隐藏的突变模式和跨物种相似性，而这是仅靠人工整理无法实现的。

# 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.

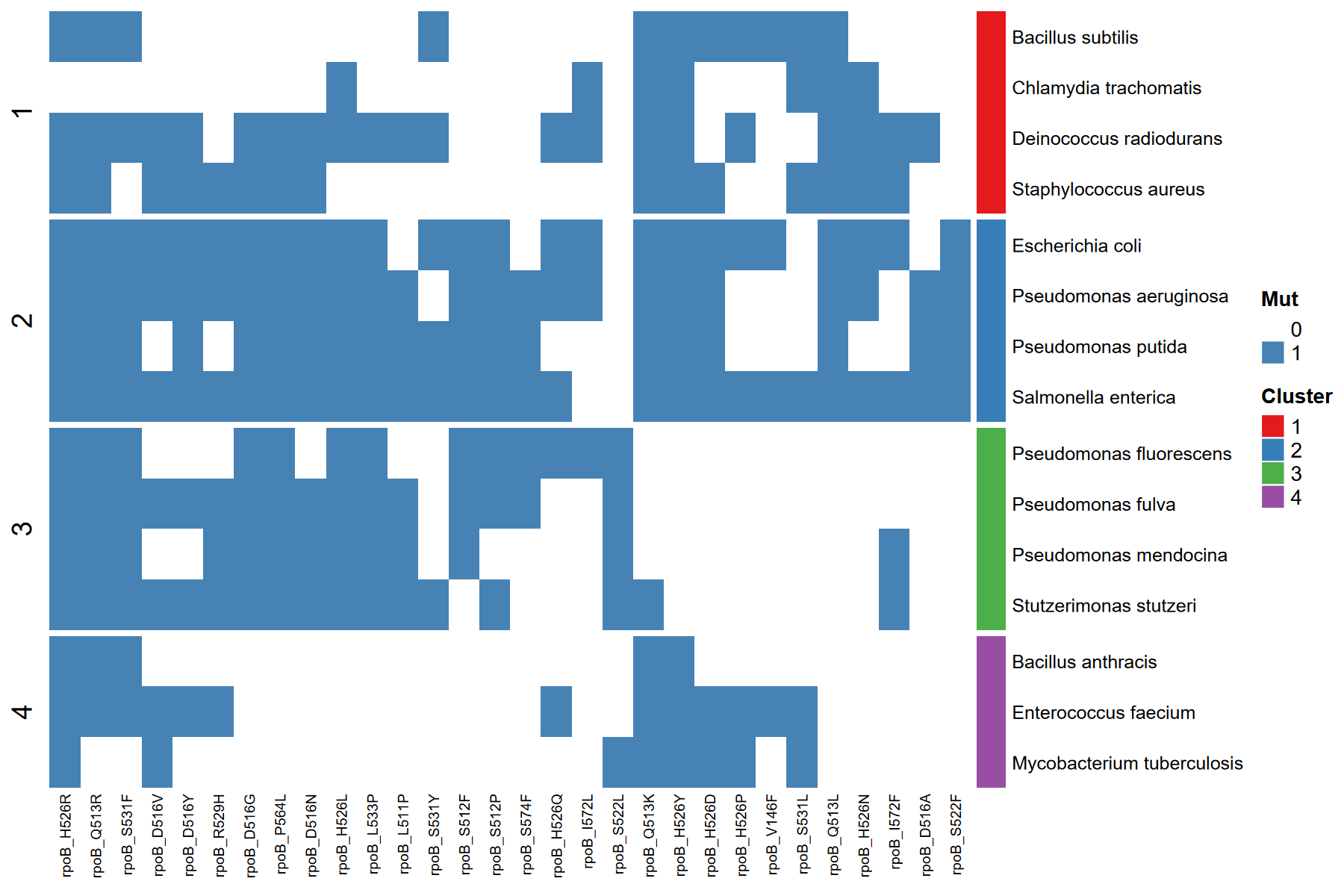
For Top-K, we refill from the remaining pool (highest p\_true) until reaching K or the pool is exhausted.

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

# Results

### ****heatmap_COSINE_GMM_top30_simple****

### ****heatmap_COSINE_KMEANS_top30_simple****

********

**Comparative Analysis of Result Figures:**

****Cosine + GMM****

**The most natural clustering results:**

**Cluster 1 (red): M. tuberculosis, B. anthracis, E. faecium → corresponding to the actinomycetes/Gram-positive group, with mutations concentrated in the L533R, S531L, and S522F regions.**

**Cluster 2 (blue): E. coli, Pseudomonas, Salmonella → Gram-negative group, with mutations concentrated in the D516V/H526Y/S531F regions.**

**The boundaries between clusters are clear and consistent with known phylogenetic distributions.**

****Cosine + K-means****

**The Gram-negative and Gram-positive groups can still be distinguished, but some boundaries (such as between Bacillus anthracis and Staphylococcus aureus) are slightly blurred.**

**The mutation pattern is slightly fragmented.**

****Euclidean + K-means****

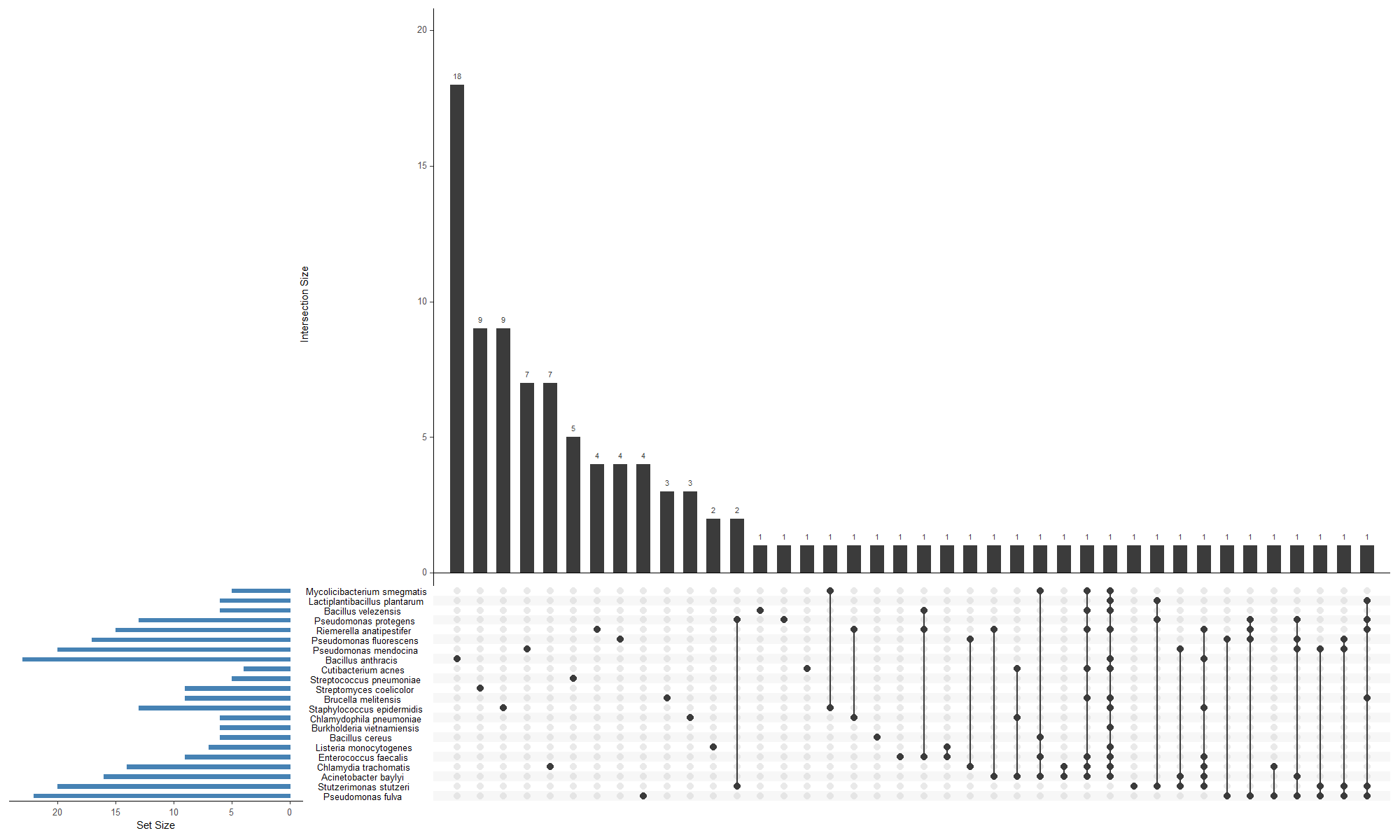
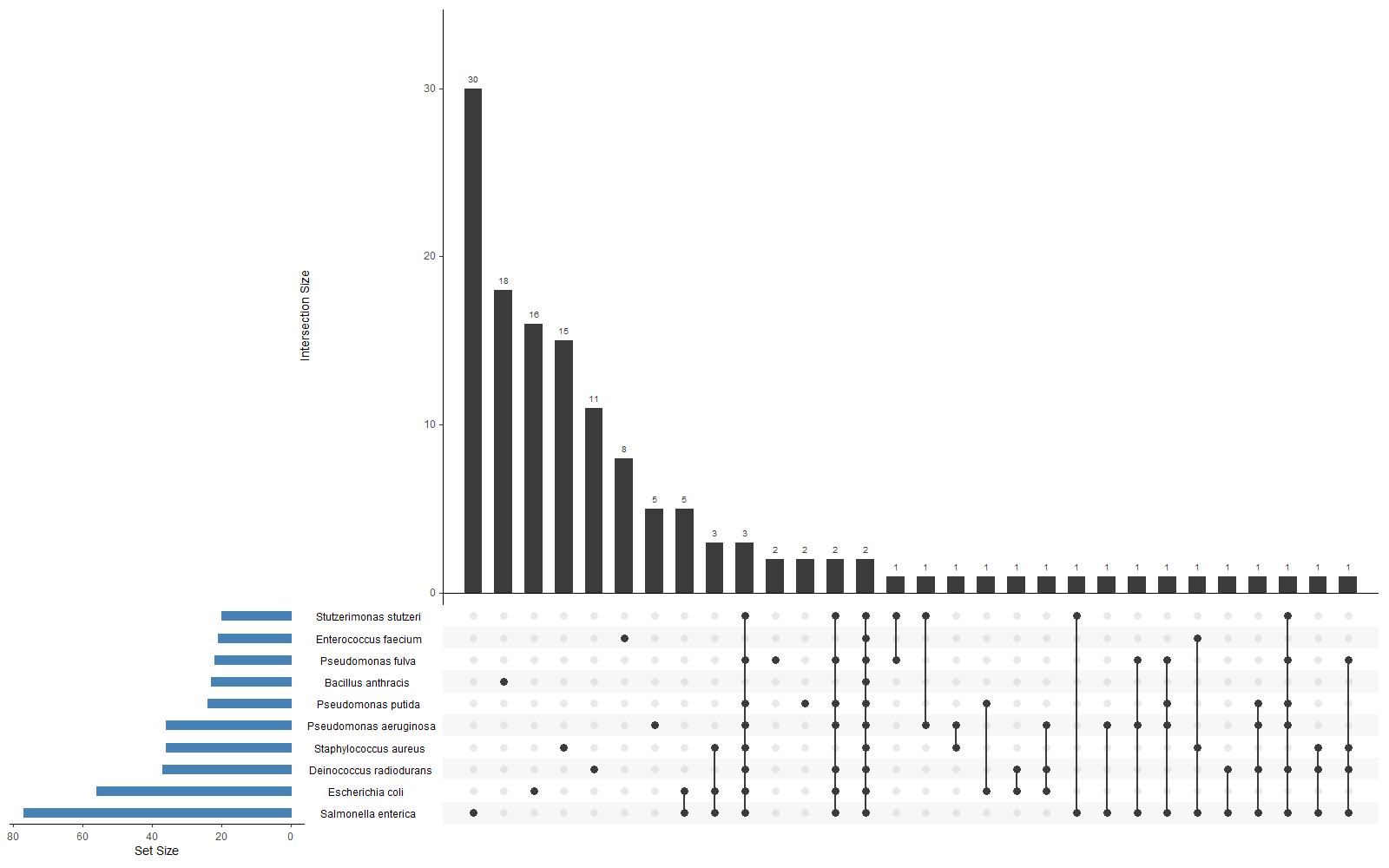
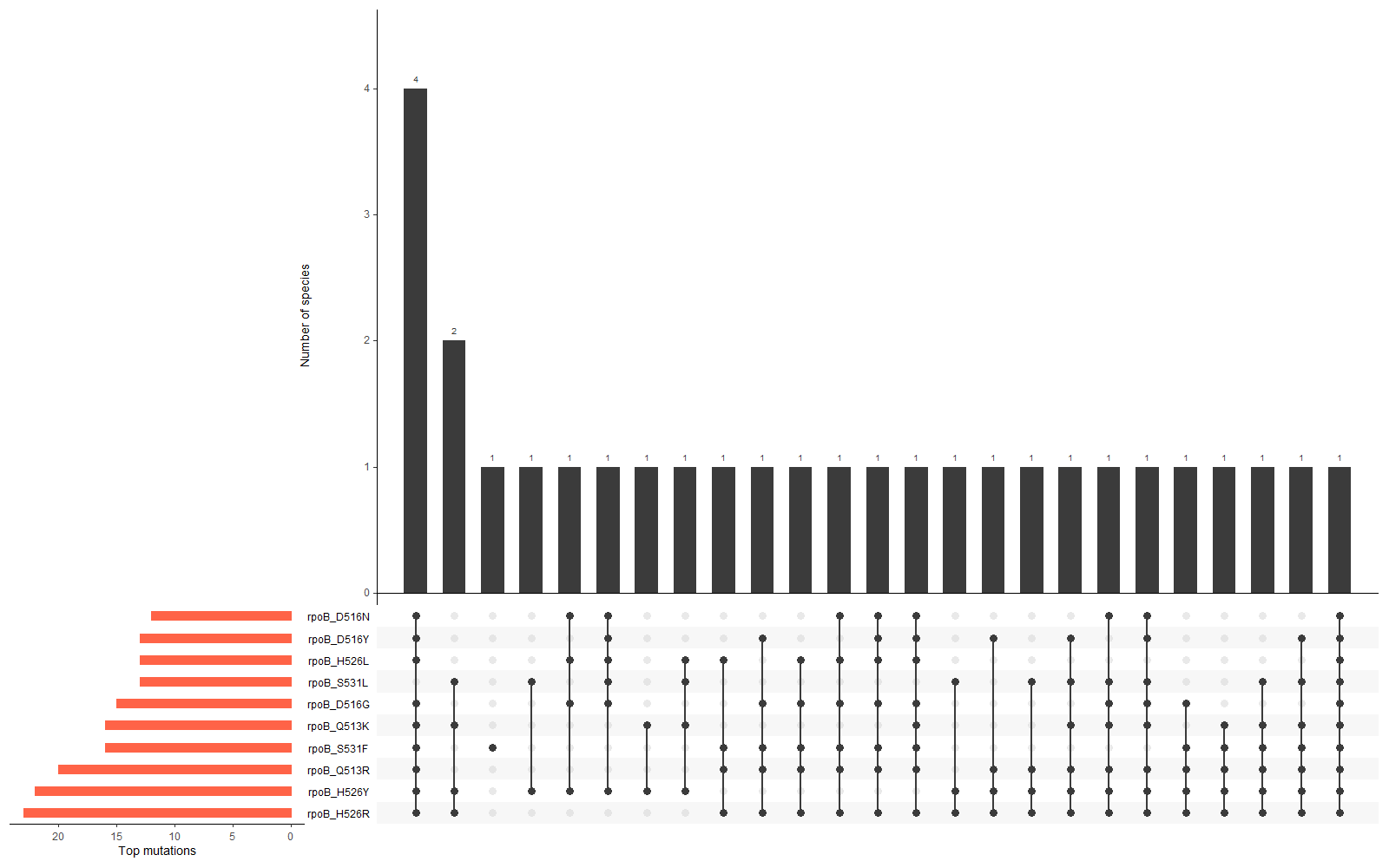
**Clustering is primarily driven by the number of mutations between samples rather than pattern similarity,**

**resulting in species with broad mutation spectra clustering together and those with fewer mutations clustering separately.**

**This results in the weakest biological interpretation.**

**Of the three clustering strategies, the Cosine distance + GMM model yielded the most stable results, successfully distinguishing actinomycetes/Gram-positive groups represented by M. tuberculosis and B. anthracis from Gram-negative groups represented by E. coli and Pseudomonas.**

**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.**



在前10物种图中，几乎所有交集都集中在那几个“经典耐药”属：  
Enterobacteriaceae, Pseudomonas, Staphylococcus。  
这些携带 rpoB 的 **RRDR 区域突变（如 H526、S531、D516）**。

而在 mid-high 全体物种图中，作者可以看到新的菌属进入（如 Mycolicibacterium, Streptomyces, Acinetobacter, Listeria, Burkholderia 等），  
这些多数拥有非典型突变位点（outside RRDR）。

交集柱子变矮、分布更均匀 → 表示这些突变几乎不再跨属共享，**提示耐药演化进入“独立发生阶段”**

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.