## Method Process and Result Analysis Based on SupervisedML (Outline)

## Methodology

## ***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)}.

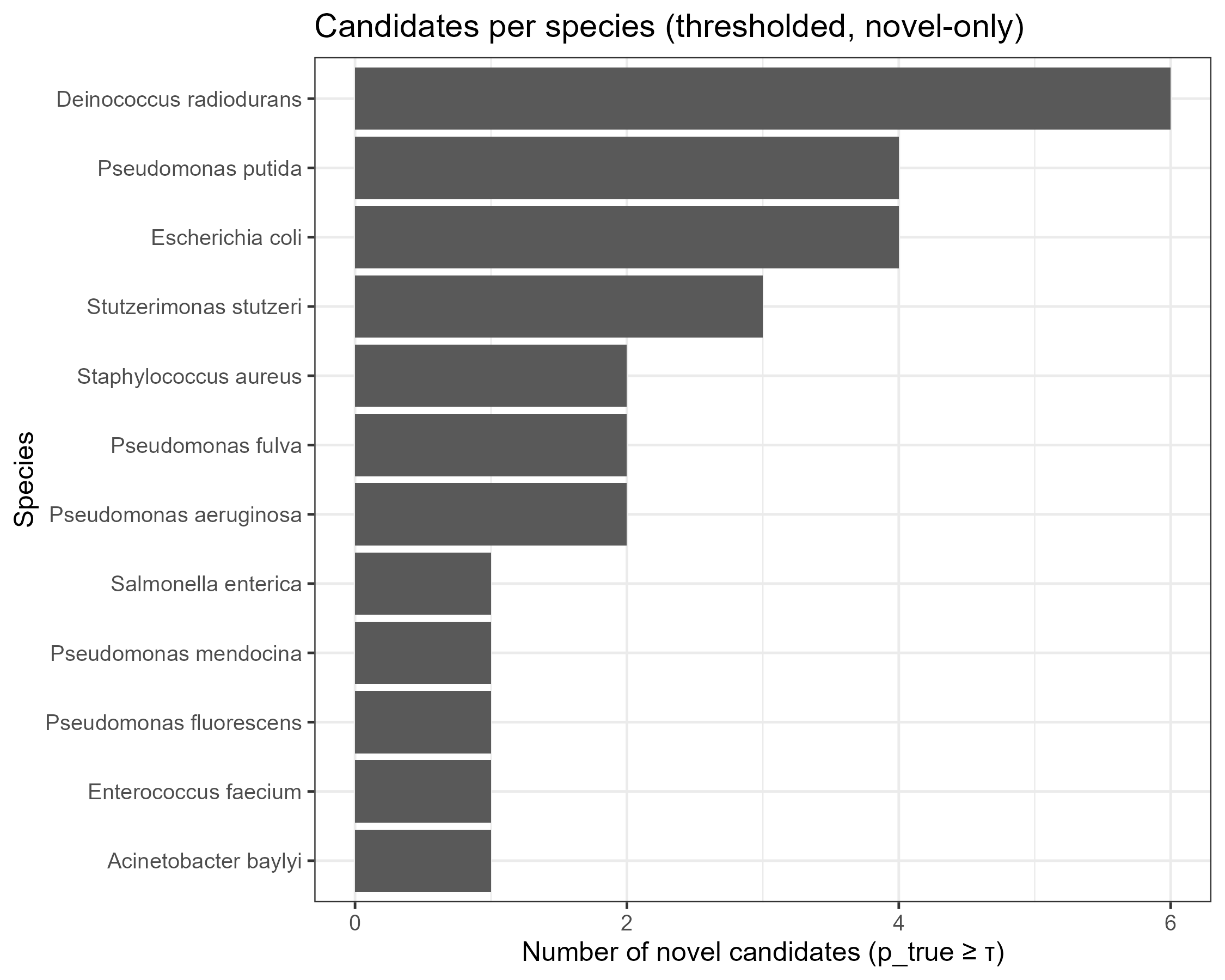
Filtering. We do not retrain the model. We filter the scored candidates by excluding anything in S.

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.

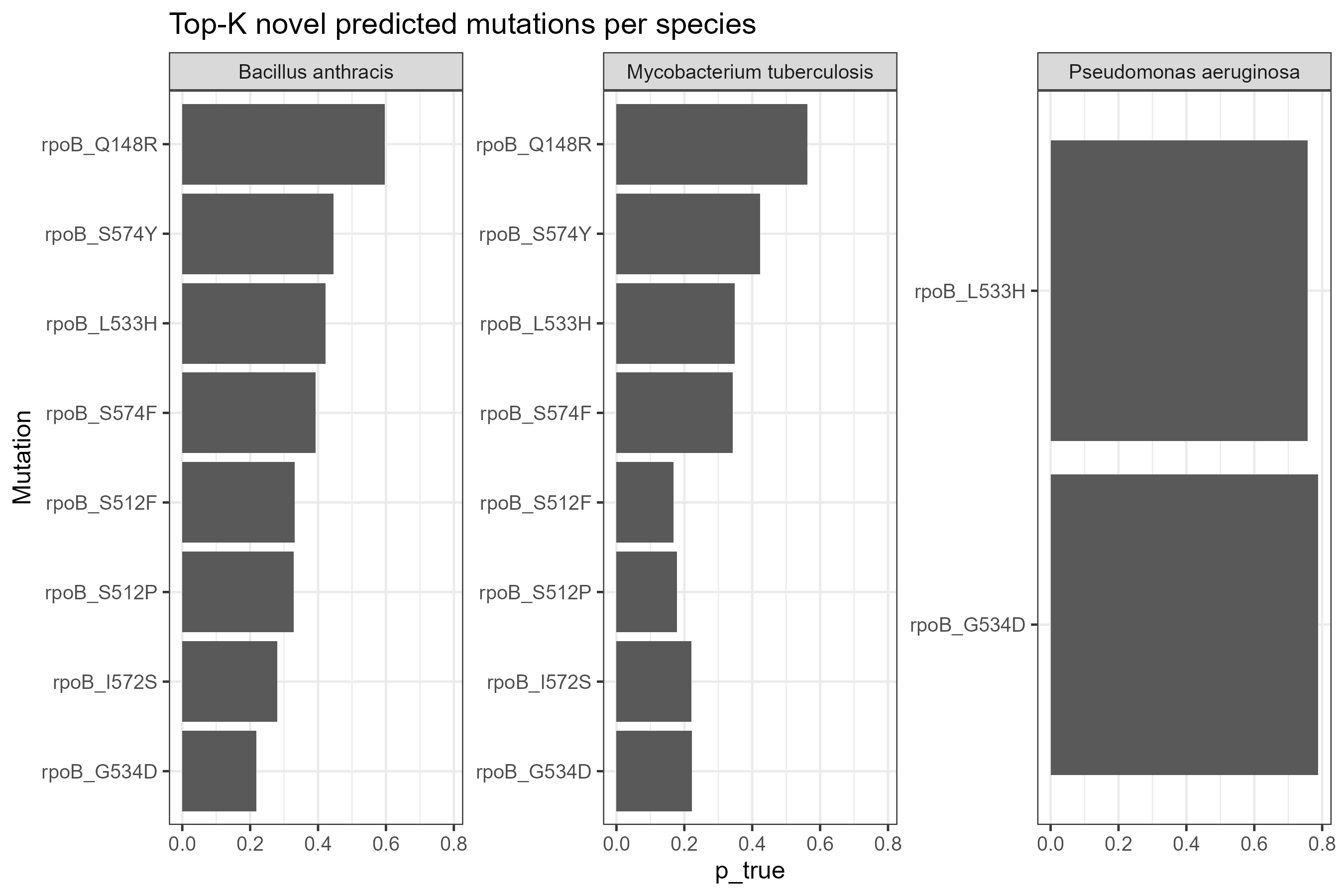
**Analysis of Figures：**

Fig1 – Novel threshold method: the number of candidate species



结论：Candidates were highly heterogeneous (Deinococcus radiodurans≈6, Pseudomonas putida / E. coli≈4,rest of species

Fig2 – novel per-species Top-K column figs

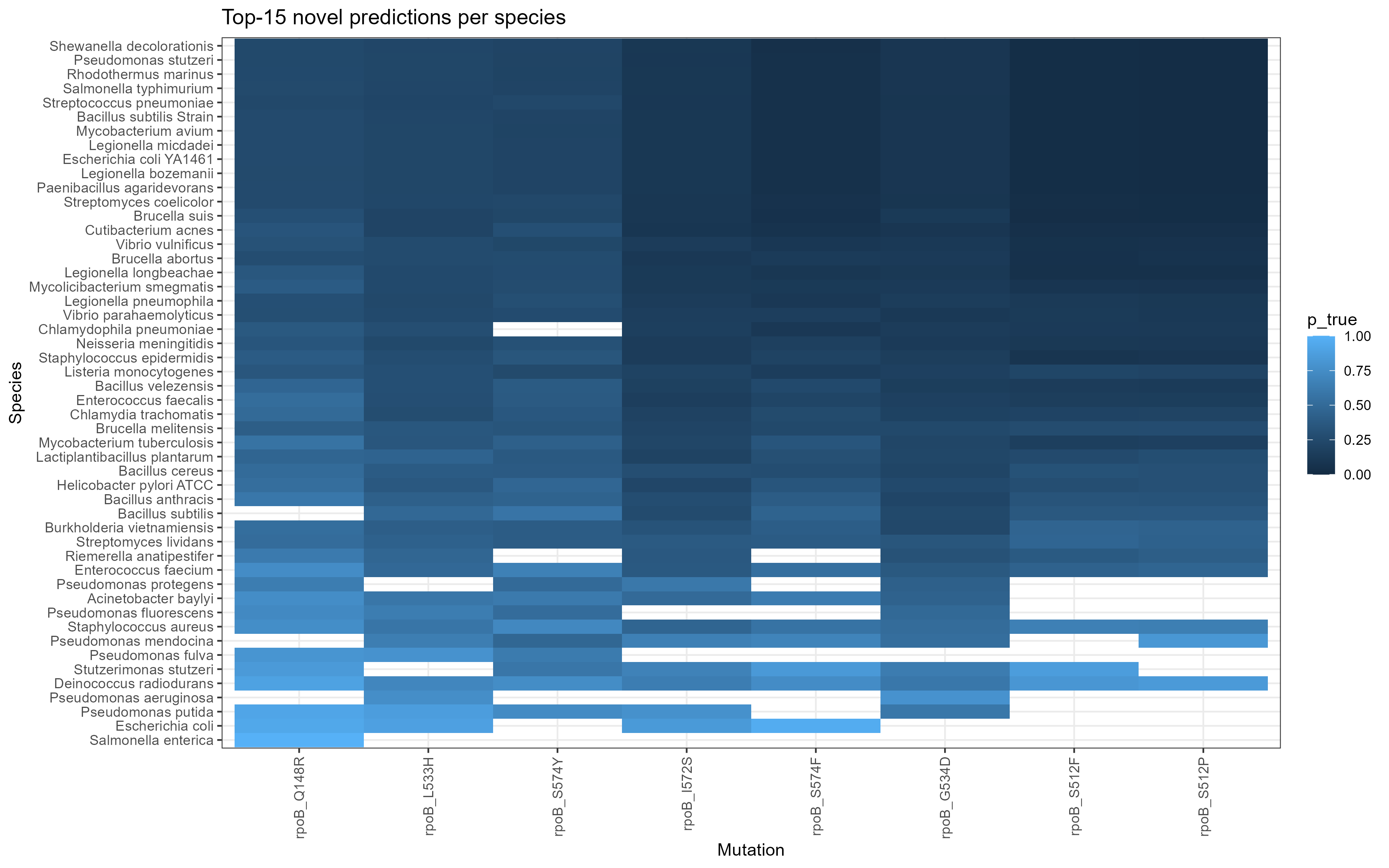


Both B. anthracis and M. tuberculosis ranked rpoB\_Q148R very highly (p\_true ~0.7–0.8), along with the S574 series and L533H.

P. aeruginosa only had two remaining entries (L533H and G534D), both of which were highly ranked (~0.75–0.8).

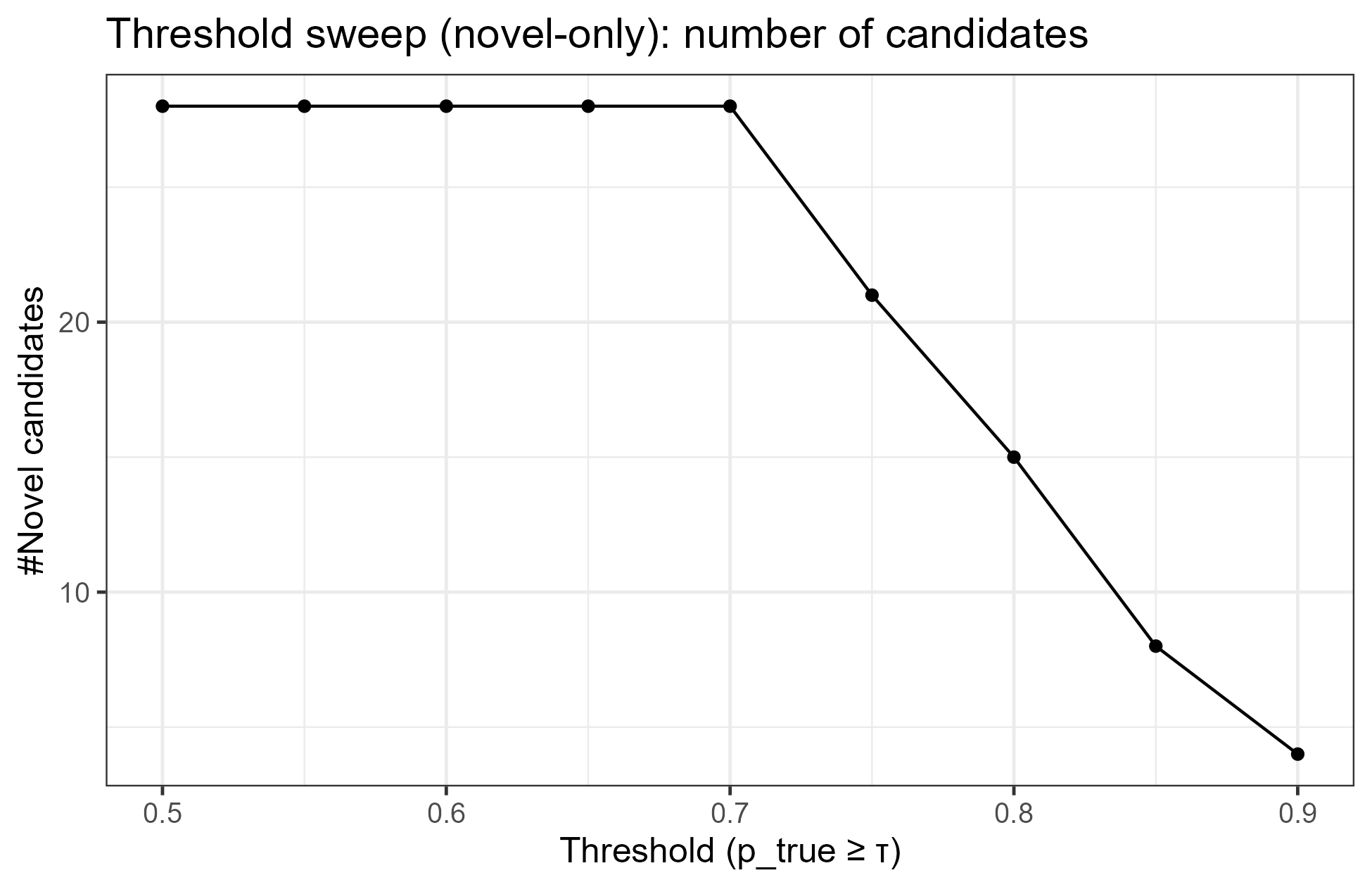
Implications: Cross-species co-occurring sites (148, 533, 574, 512, 572, etc.) appear repeatedly, and the model considers these sites "unreported but likely to occur."

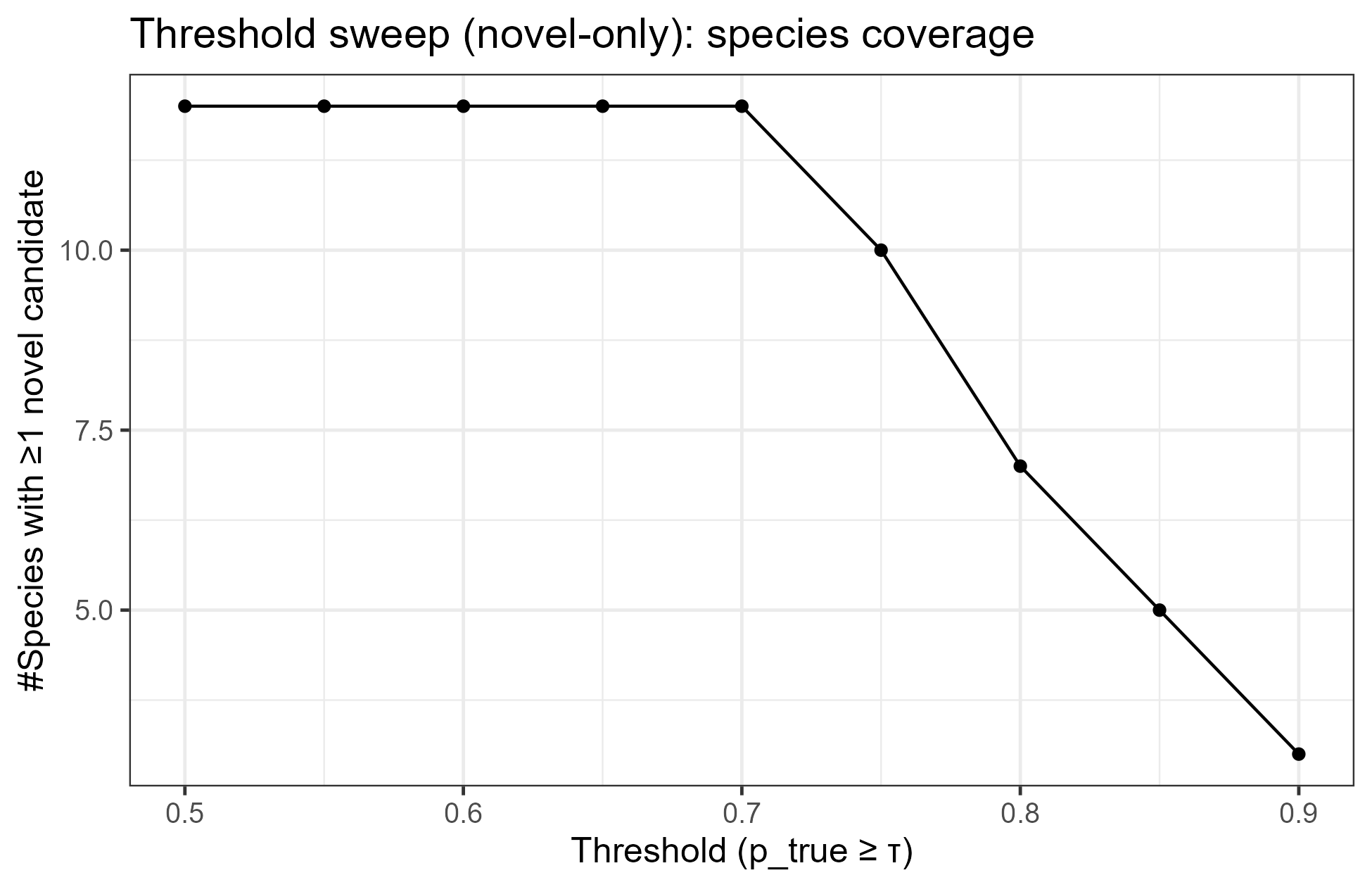
Fig3 – novel Top-15 in species heatmap



The columns are basically the hotspot sites mentioned above. Many species have dark blocks (high scores) in these columns, and white blocks indicate that the species does not list the mutation in the top 15.

Fig4a/4b – novel threshold scanning

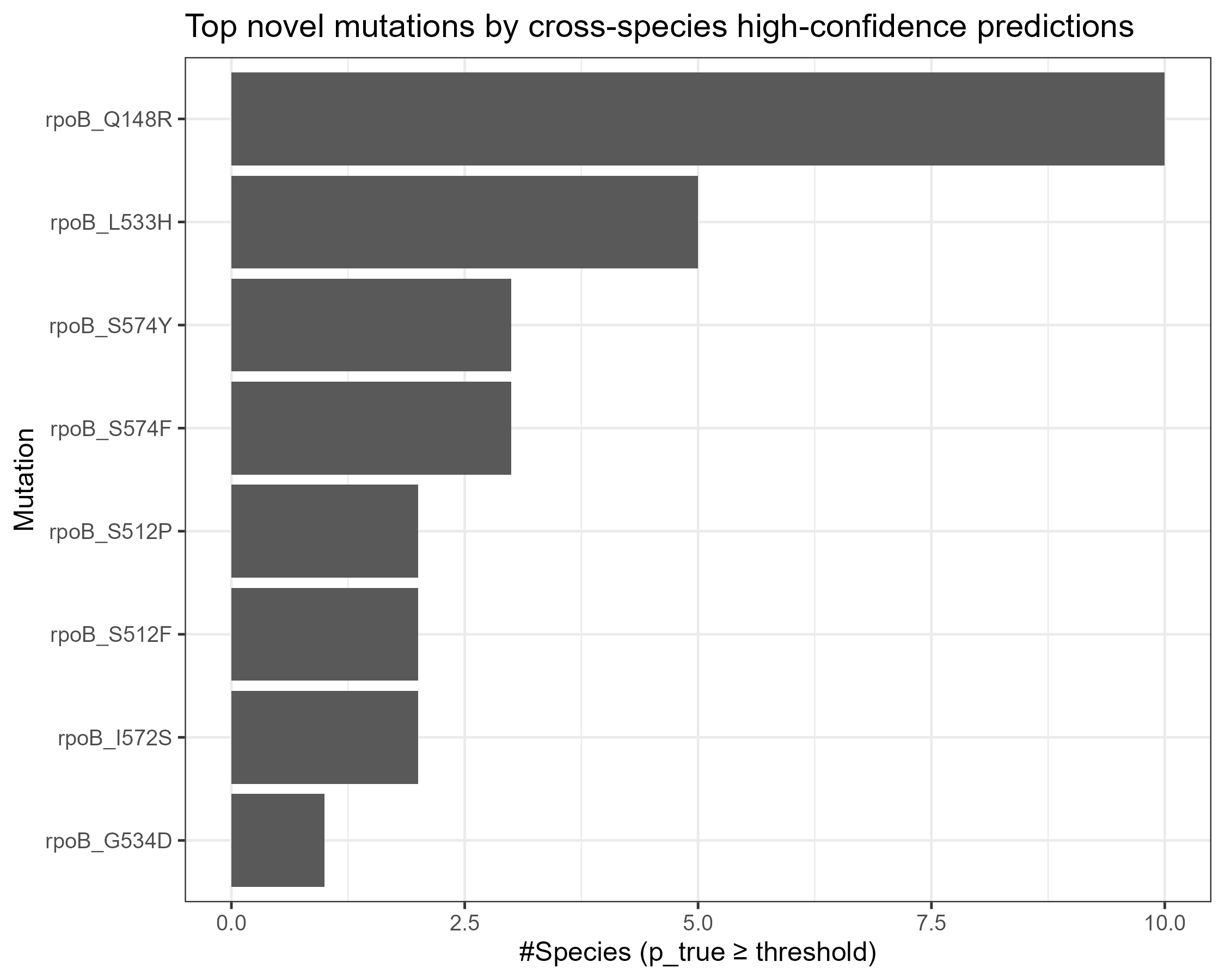




Conclusion: The number of candidates and coverage remain high at 0.7 (but decrease significantly after 0.75).

Thus, in the paper, I can say that τ≈0.7 is a good balance; >0.8 significantly sacrifices coverage.

Fig5



rpoB\_Q148R entered high scores in the most species (~10), followed by L533H, S574Y/F, S512P/F, I572S, and G534D.

**Masked vs Unmasked\_cf description**

**Masked**: Mask some known true positives as unlabeled (0), then train and retrieve them in the unlabeled pool, and calculate Recall@K. → Closer to the difficulty of "real discovery", should be more conservative.

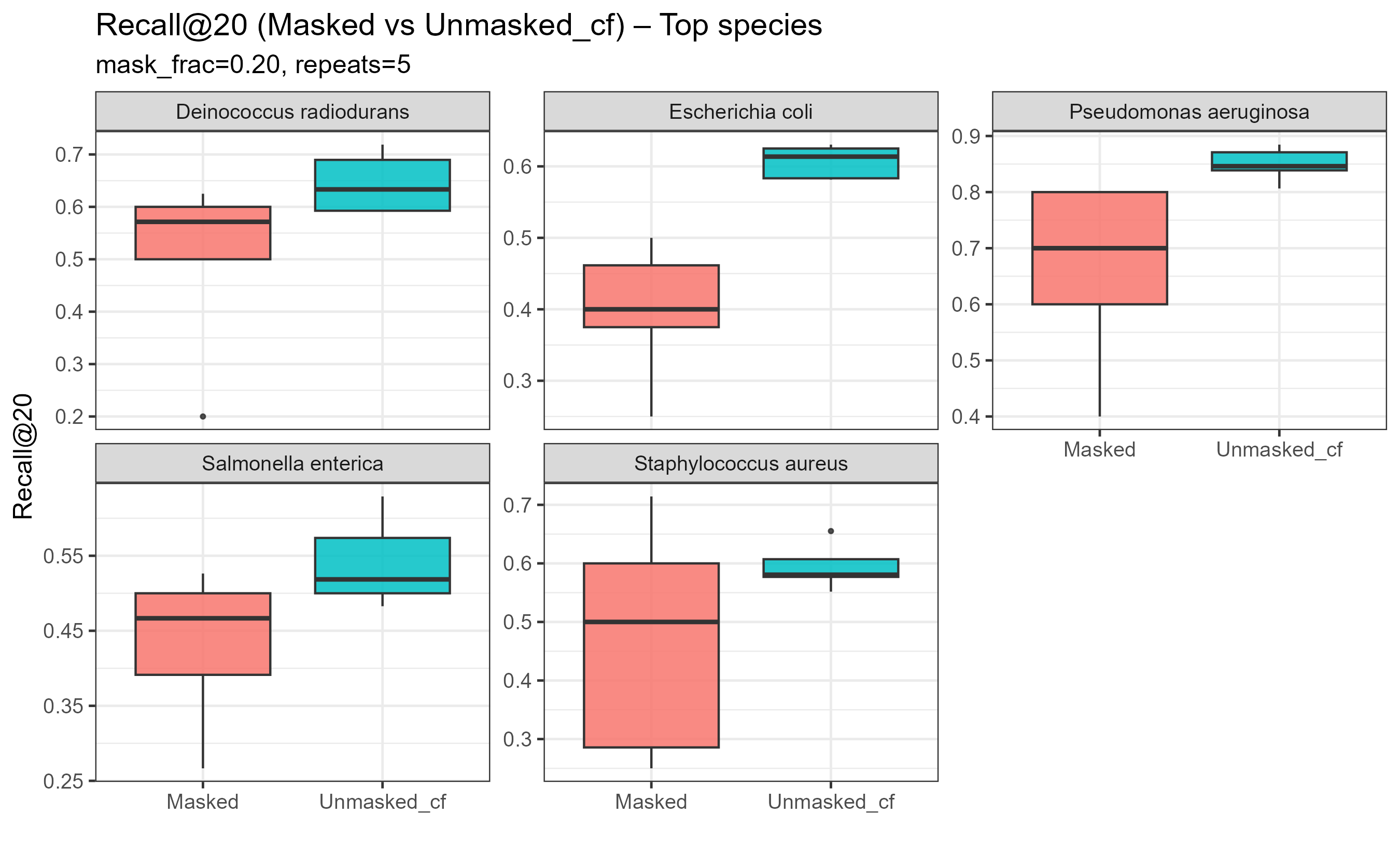
**Unmasked\_cf (counterfactual)**: These true positives are not masked (still participate in training as positive examples), but we use the scores of the same model to put them into the unlabeled pool, and compare them with the unlabeled scores to calculate "how high they would be ranked if they were also unlabeled." This is an "upper bound/optimistic" reference, usually higher than Masked.

Box plot (gray/pink/blue rectangle): The interquartile range of the group (Masked or Unmasked\_cf) under multiple repeats.

Therefore, the "height" of the box reflects the dispersion of the distribution (IQR = Q3 − Q1); the "higher" means the greater the fluctuation.

"Whiskers" (thin lines outside the box): The minimum/maximum values ​​extending from the box to non-outliers (usually the extreme values ​​in the range Q1 − 1.5 × IQR to Q3 + 1.5 × IQR).

Black dots: Outliers, i.e., individual replicates that fall outside the whiskers (significantly lower or higher under certain folds/samples).·



· 五个物种里，Unmasked\_cf 全都高于 Masked（预期现象）。Unmasked\_cf 是“未被遮蔽的真阳性”的反事实上界：它们仍作为正例参与训练，所以得分更乐观；而 Masked 是把一部分真阳性塞进未标注池再去“找回”，更接近真实难度。

· · P. aeruginosa：表现最好，Masked 的中位也在 ~0.7；Unmasked\_cf 接近 0.85–0.9，gap 小，说明模型对该物种的排序很稳。

· · E. coli / Salmonella / S. aureus：Masked 中位在 0.4–0.55，且 S. aureus 方差较大，说明重复采样下稳定性稍弱（多半因为该物种的正例分布更稀/不平衡）。

· · D. radiodurans：两组都高（~0.6–0.7），整体稳

·