## 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)}.

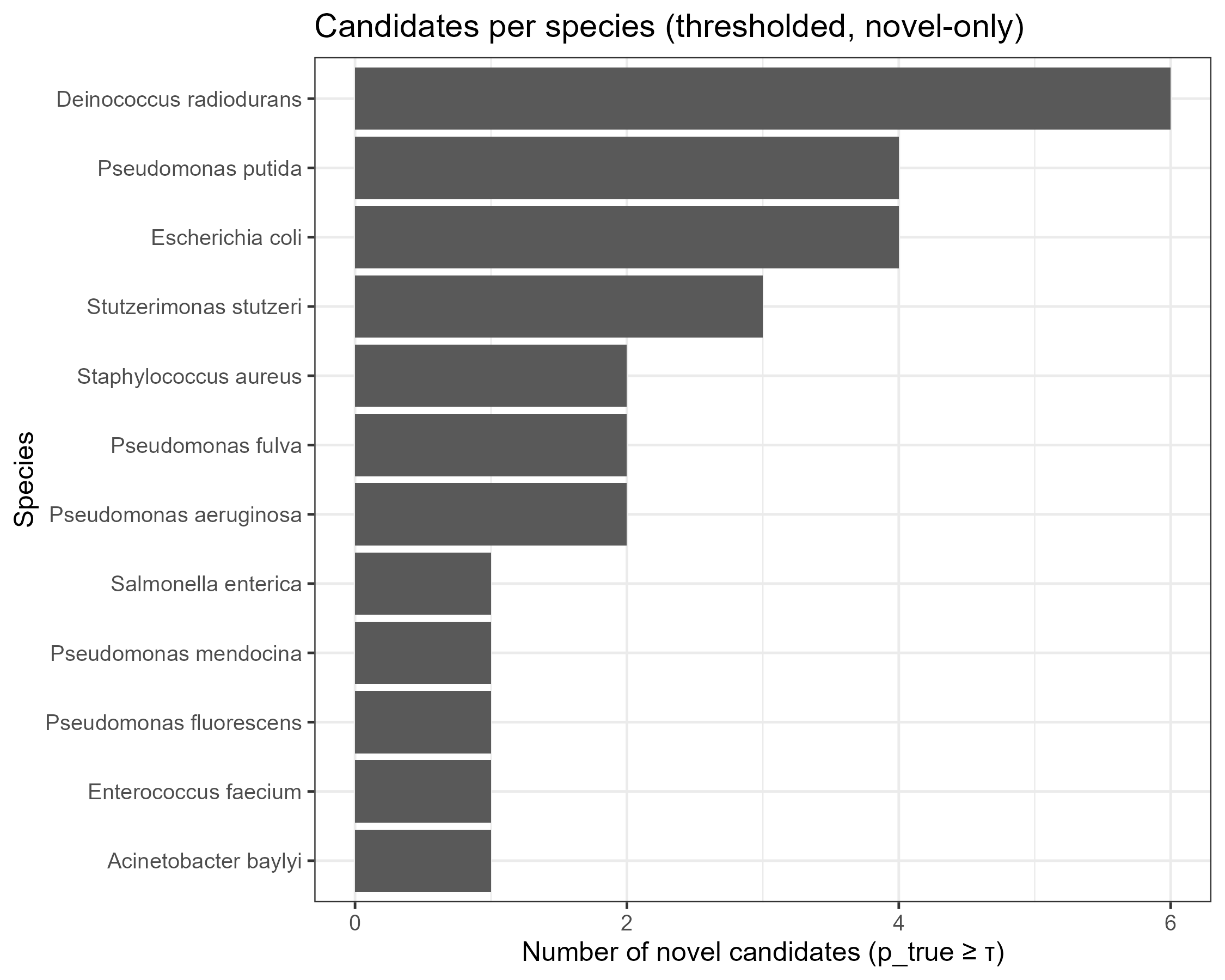
Filtering. We do not retrain the model. We filter the scored candidates by excluding anything in S.

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.

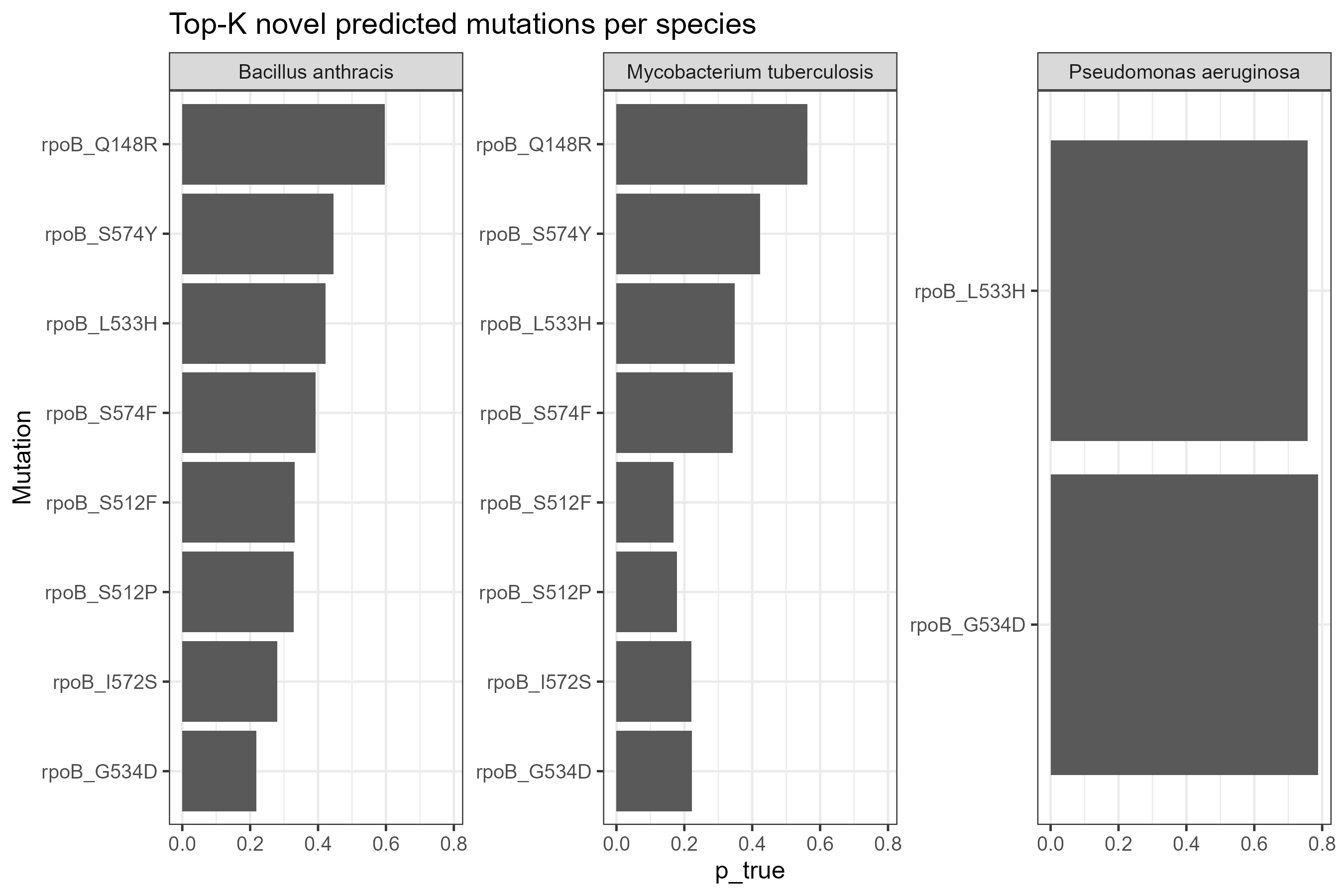
**Analysis of Figures：**

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

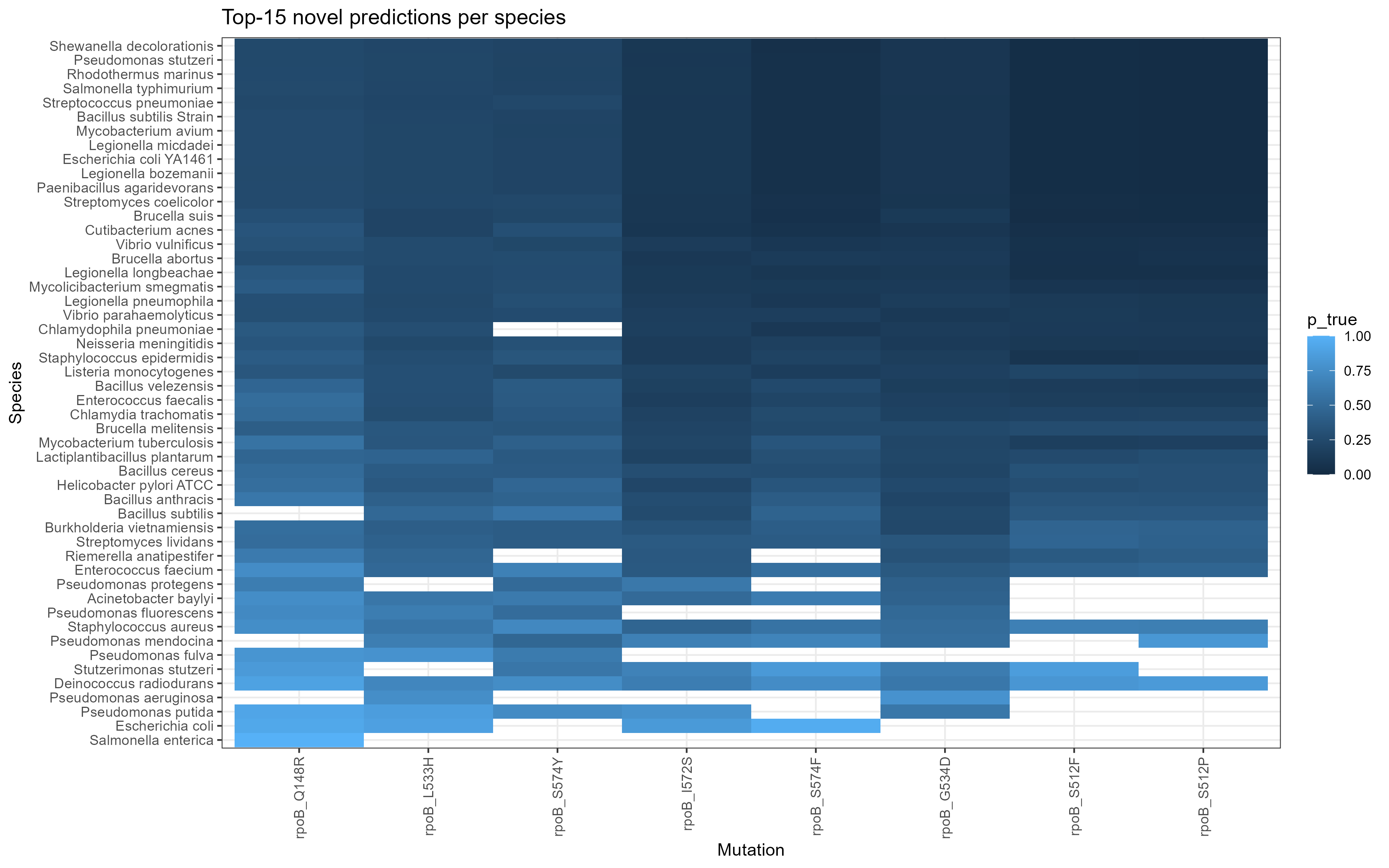


Both B. anthracis and M. tuberculosis ranked rpoB\_Q148R very highly (p\_true ~0.7–0.8), along with the S574 series and L533H.

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

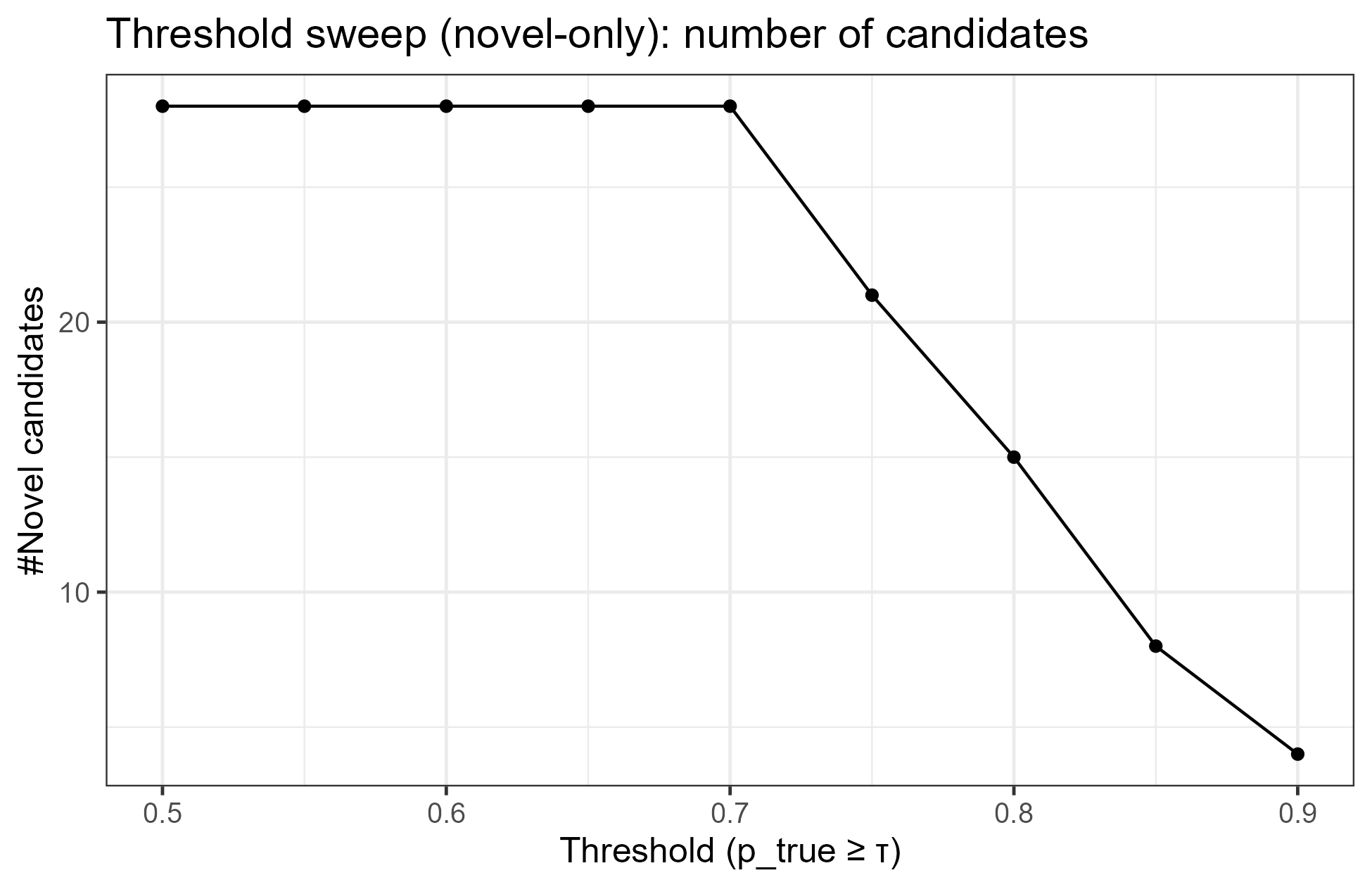
Implications: Cross-species co-occurring sites (148, 533, 574, 512, 572, etc.) appear repeatedly, and the model considers these sites "unreported but likely to occur."

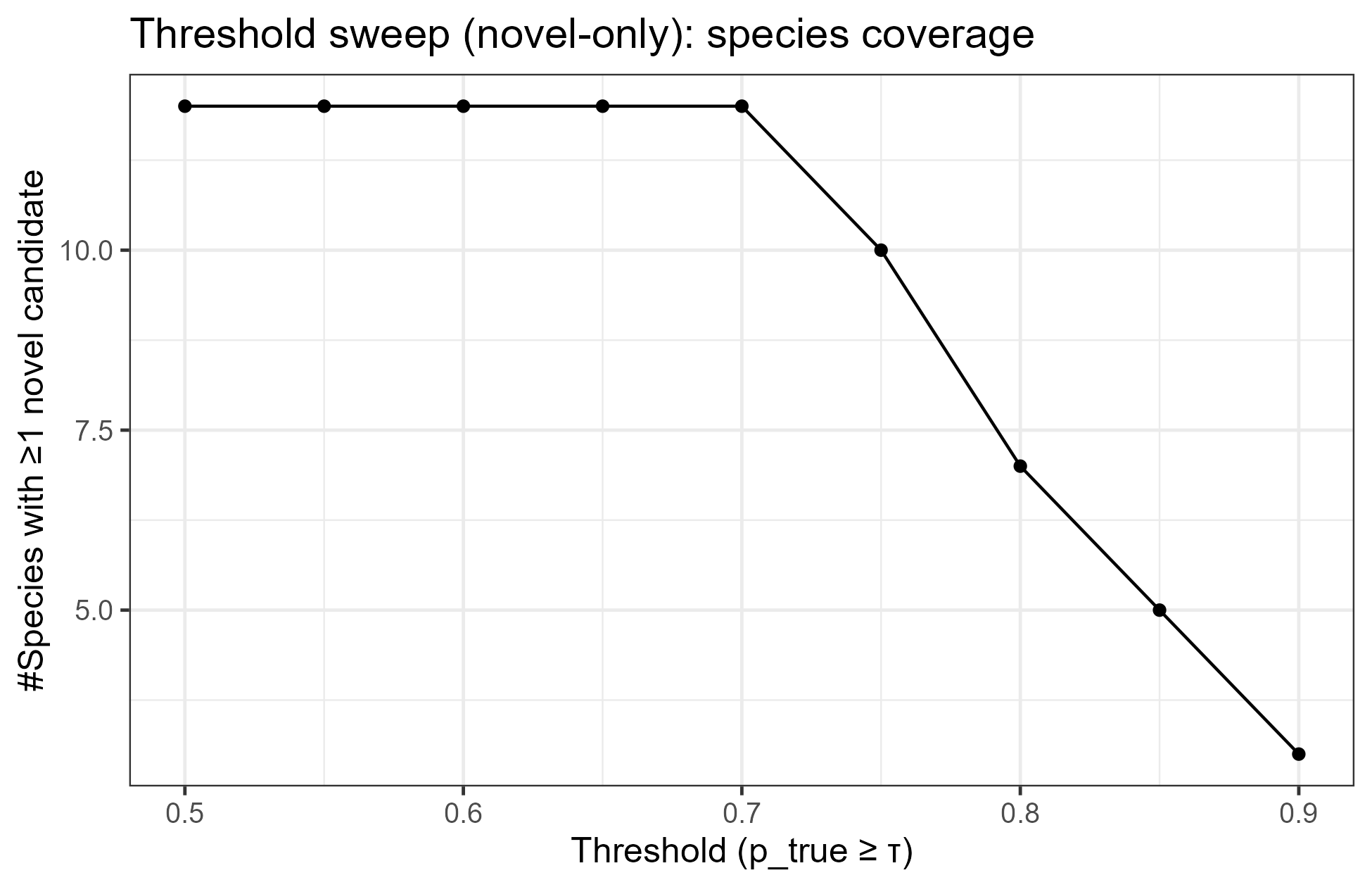
Fig3 – novel Top-15 in species heatmap



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

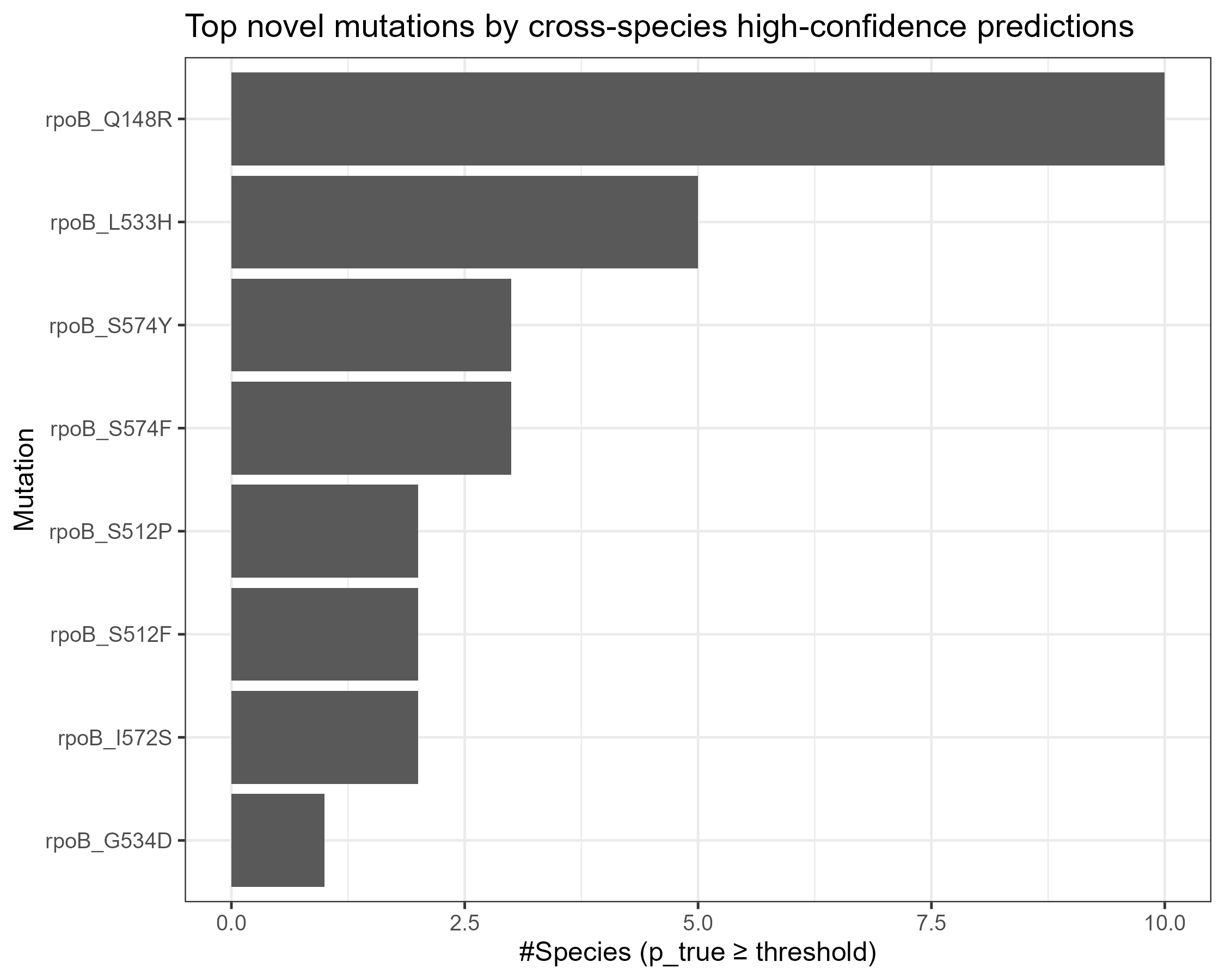




Conclusion: The number of candidates and coverage remain high at 0.7 (but decrease significantly after 0.75).

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.

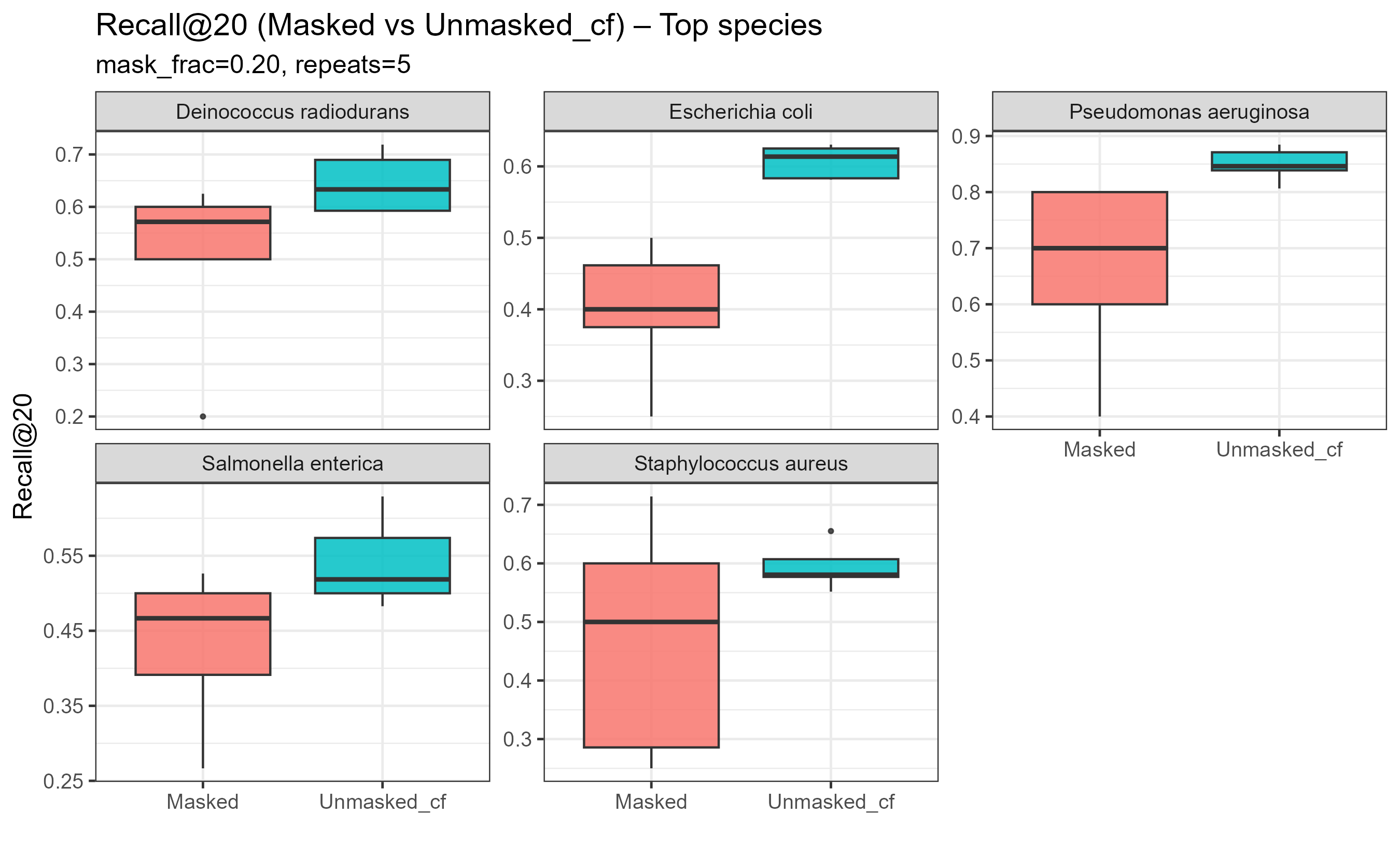
**Unmasked\_cf (counterfactual)**: These true positives are not masked (still participate in training as positive examples), but we use the scores of the same model to put them into the unlabeled pool, and compare them with the unlabeled scores to calculate "how high they would be ranked if they were 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).·



For all five species, Unmasked\_cf is higher than Masked (expected). Unmasked\_cf is a counterfactual bound for "unmasked true positives": they still participate in training as positive examples, resulting in a more optimistic score. Masked, on the other hand, inserts some true positives into the unlabeled pool and then "recovers" them, which is closer to the actual difficulty level.

P. Aeruginosa has the best performance, with a median Masked score of ~0.7; Unmasked\_cf is close to 0.85–0.9, with a small gap, indicating that the model's ranking of these species is very stable.

For E. coli / Salmonella / S. aureus, their masked score has a median score of 0.4–0.55, and S. aureus has a larger variance, indicating slightly weaker stability under repeated sampling (mostly due to the more sparse/unbalanced distribution of positive examples for this species).

D. radiodurans: Both groups have high scores (~0.6–0.7), and are generally stable.(this could be think as?）

Top20 high confidential prediction

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | p\_obs | p\_true | Observed | Mutation |
| Salmonella enterica | 0.81384 | 1 | 0 | rpoB\_Q148R |
| Escherichia coli | 0.81744 | 0.965289211584658 | 0 | rpoB\_S574F |
| Escherichia coli | 0.70424 | 0.9363897457717264 | 0 | rpoB\_Q148R |
| Pseudomonas putida | 0.6900000000000001 | 0.9174555898308692 | 0 | rpoB\_Q148R |
| Deinococcus radiodurans | 0.67304 | 0.8949047973619828 | 0 | rpoB\_Q148R |
| Escherichia coli | 0.71704 | 0.885905376910912 | 0 | rpoB\_L533H |
| Pseudomonas putida | 0.70056 | 0.8655442804428043 | 0 | rpoB\_L533H |
| Stutzerimonas stutzeri | 0.75088 | 0.8612413849353614 | 0 | rpoB\_S512F |
| Stutzerimonas stutzeri | 0.63464 | 0.8438463993192213 | 0 | rpoB\_Q148R |
| Deinococcus radiodurans | 0.74624 | 0.8431624618351278 | 0 | rpoB\_S512P |
| Escherichia coli | 0.71648 | 0.8401500938086304 | 0 | rpoB\_I572S |
| Stutzerimonas stutzeri | 0.70848 | 0.8366217711679127 | 0 | rpoB\_S574F |
| Pseudomonas mendocina | 0.72576 | 0.8200224971878515 | 0 | rpoB\_S512P |
| Deinococcus radiodurans | 0.71064 | 0.8150870682272338 | 0 | rpoB\_S512F |
| Pseudomonas fulva | 0.60496 | 0.8043825124986704 | 0 | rpoB\_Q148R |
| Pseudomonas fulva | 0.63808 | 0.7883500263574064 | 0 | rpoB\_L533H |
| Pseudomonas aeruginosa | 0.65568 | 0.788304318553429 | 0 | rpoB\_G534D |
| Pseudomonas putida | 0.6644 | 0.7790806754221388 | 0 | rpoB\_I572S |
| Pseudomonas aeruginosa | 0.61344 | 0.7579072219293621 | 0 | rpoB\_L533H |
| Staphylococcus aureus | 0.56896 | 0.7565152643335816 | 0 | rpoB\_Q148R |

Rpob-Q148R 7

|  |  |  |  |
| --- | --- | --- | --- |
| CC261 | Q148R | CAG → CGA | This work |

Q148R, a mutation associated with rifampicin resistance, also affects transcription slippage.

Transcriptional slippage occurs at mononucleotide repeats, causing RNA polymerase to synthesize a population of mRNAs with varying lengths. These mRNAs exhibit variations in nucleotide number, leading to the translation of multiple distinct protein products. If these proteins lack distinct functions, slippage is generally disadvantageous, leading to selection pressure to eliminate most slippage-prone sequences within coding regions.

Analysis of slippage efficiency data for similar mutations, such as P564L and Q513P, highlights Q148R, along with other mutations in the same region, as "highly concentrated around the RNA–DNA hybrid region, critical amino acids for controlling RNAP slippage." These mutations may increase slippage by weakening the DNA clamp of RNAP, making the RNA–DNA hybrid more easily unwound.

Rpob-L533H 4

|  |  |  |  |
| --- | --- | --- | --- |
| CC461 | L533R | CTC → CGA | This work |

The mutation also appears in the fork region of the β subunit, close to the RNA–DNA hybrid, a key region for transcriptional slippage. It is one of 25 mutations clustered around the rifampicin binding pocket that are of particular interest, suggesting that this site may alter the stability of RNA polymerase on the RNA–DNA hybrid, thereby affecting transcriptional slippage.

rpoB\_S512P 2

rpoB\_S512F 2

**Reason of L533H/Q148R been selected as good candidate**

In my training data, these two sites (particularly around 533, at the end of the classic RRDR window) have been reported in several species and are strongly correlated with other typical RRDR sites (such as S512, S531/533, and S574). PU-RF learns a position-level prior that "changes at these sites ⇒ high resistance probability."

Calculating the proportion of rpoB RRDR Cluster I (approximately 507–533 aligned to E. coli): Approximately 32–34% of candidates in the top3\_per\_species and high\_conf are concentrated in the RRDR region, significantly higher than the "whole protein random" baseline. L533H falls right on the edge of the RRDR "hotspot," and is therefore naturally more frequently promoted by the model.

**Substitution type preference:**

In all three tables, R (Arg) and H (His) are highly represented as substituted amino acids (R > H > Y). This suggests that the model is biased towards predicting substitutions that introduce positive charge/aromaticity—substitutions near pocket sites that are more likely to perturb rifampicin binding (altering hydrogen bonds, electrostatics, or steric hindrance).

**Of note:**

Q148 is not a classic RRDR, yet it appears very frequently across species in my predictions (47 times/top-3 list). This is generally due to two reasons:

The training focused on positive samples in certain bacterial communities where Q148 is already present (or sites that strongly co-occur with it);

In my species clusters, many "neighboring species" have feature combinations similar to these positive samples, leading PU-RF to identify Q148R as a transferable secondary hotspot.

This may provide a strong candidate for a "non-RRDR but consistently occurring" mutation, worthy of separate discussion.

Species analysis:

Escherichia coli--4 muts

Gensu Pseudomonas--8 muts

Pseudomonas is a [genus](https://en.wikipedia.org/wiki/Genus" \o "Genus) of [Gram-negative bacteria](https://en.wikipedia.org/wiki/Gram-negative_bacteria" \o "Gram-negative bacteria) belonging to the family [Pseudomonadaceae](https://en.wikipedia.org/wiki/Pseudomonadaceae" \o "Pseudomonadaceae) in the class [Gammaproteobacteria](https://en.wikipedia.org/wiki/Gammaproteobacteria" \o "Gammaproteobacteria).