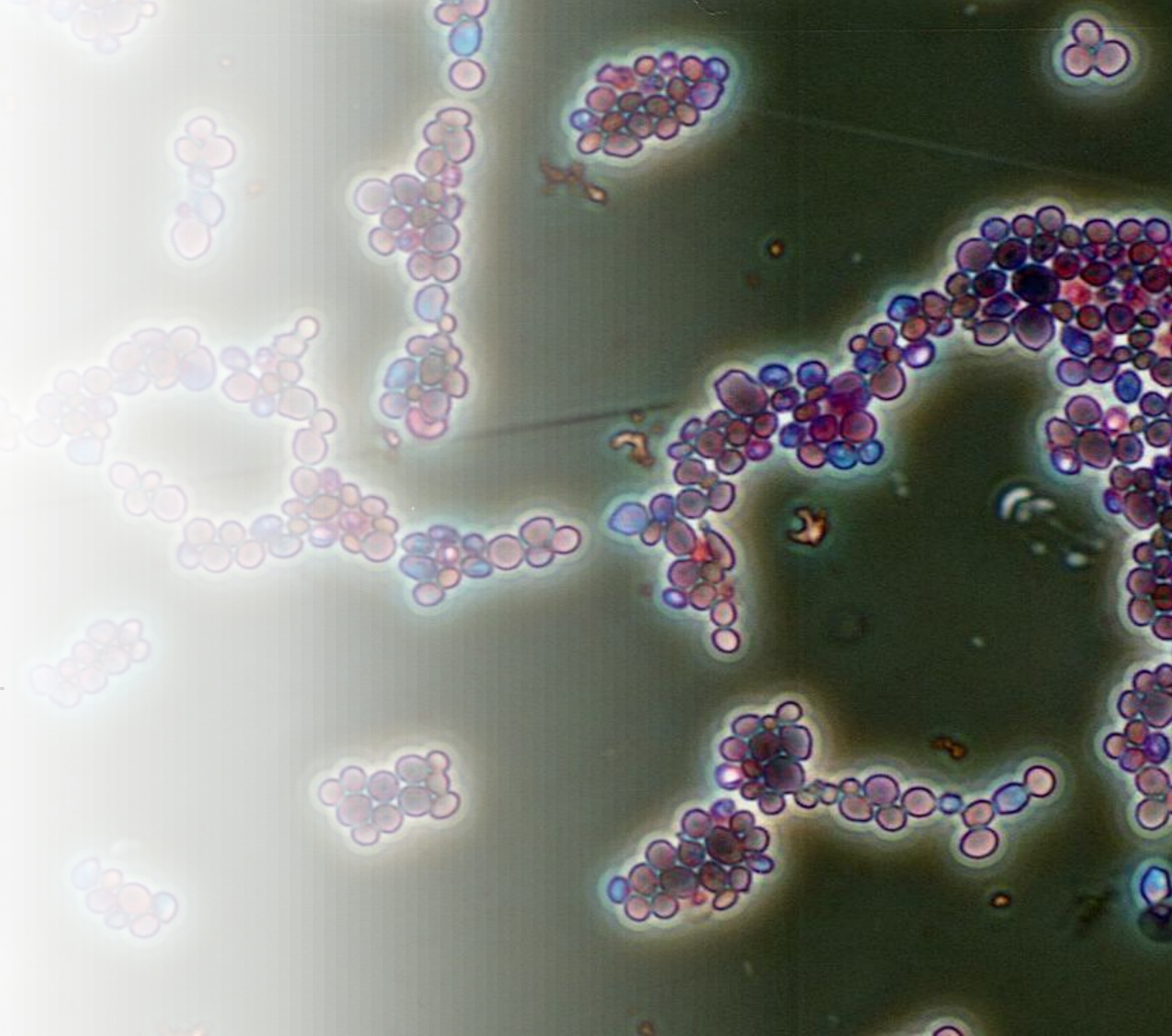
Species-Level Clustering of *Candida* Using ITS and ERG11 Gene Sequences

Sabrina Saiphoo

https://github.com/snsaiphoo/binf6210_assignment4.git



Species Clustering & Antifungal Resistance Concern in *Candida*

Candida species cause a wide range of human fungal infections and can vary greatly in antifungal resistance (Whaley et al., 2017).

This project explores how high- and low-resistance Candida species cluster using two fungal marker genes.

The species are:

- **High resistance concern** → Candida auris, Candida glabrata, Candida krusei (Chowdhary et al., 2017; Arendrup et. Al., 2017),
- **Low resistance concern** → Candida albicans, Candida parapsilosis, Candida tropicalis (Nascimento et al., 2024)

The fungal markers are:

- **ITS** – internal transcribed spacer, fungal barcode (Schoch et al., 2012)
- **ERG11** – involved in ergosterol synthesis, drug-target gene linked to azole resistance (Whaley et al., 2017)

Understanding these clustering patterns may provide insight into whether gene choice (ITS vs ERG11) reflects clinically meaningful resistance categories, and if high or low resistance concern strains can be grouped.

Resistance patterns are known, but clustering patterns are not.

High and low-resistance Candida species are well documented in clinical antifungal susceptibility studies (Whaley et al., 2017).

However, it is unclear whether these resistance categories also emerge when analyzed with unsupervised clustering.

Although the ITS and ERG11 genes are widely used in fungal genomics, their ability to reflect resistance groups in clustering analyses is something to be explored (Schoch et al., 2012; Whaley et al., 2017).

Gap in knowledge:

- Do high-resistance species cluster differently than low-resistance species when applying unsupervised sequence-based methods?

Understanding this relationship will help determine whether gene choice in clustering influences species separation patterns.

Unsupervised clustering to compare high- and low-resistance *Candida* species.

The project is an exploratory, unsupervised clustering analysis.

Objective:

- Compare whether high and low-resistance Candida species show distinct clustering patterns when analyzing ITS and ERG11 gene sequences.

Research Question:

1. Are there differences between high resistance concern species vs lower resistance concern species?
2. Does ITS vs ERG11 give different clustering patterns?

Approach: Obtain ITS and ERG11 gene data from the 6 chosen Candida species and perform unsupervised clustering from pairwise distance matrices and display results from NMDS, PCA, and evaluate cluster quality.

Candida Data NCBI

Data Source:

- ITS and ERG11 gene sequences from six Candida species were obtained from NCBI Nucleotide Database (GenBank) through R's rentrez package.
- The analysis presented is based on the dataset retrieved on December 5th, 2025, as the data get continuously updated on each run of the script.

Dataset Properties:

- 500 Raw ITS and ERG11 sequences per query with length filters:
  - **ITS**: 400-800 bp – (values were chosen from Schoch et al., 2012)
  - **ERG11**: 1000-2500 bp – (values were chosen from de Oliveira Ceita et al., 2014)
- After filtering for a balanced dataset per gene:
  - 84 ITS sequences (14 samples per species)
  - 30 ERG11 sequences (5 samples per species)

Key Variables & NCBI Search Terms

Key variables:

- Gene of Interest: ITS or ERG11)
- Species of Interest:
  - High resistance concern: Candida auris, Candida glabrata, Candida krusei
  - Low resistance concern: Candida albicans, Candida parapsilosis, Candida tropicalis

Search Terms:

```
> candida6_ITS_term
[1] "(\"Candida auris\"[ORGN] OR \"Candida glabrata\"[ORGN] OR \"Candida krusei\"[ORGN] OR \"Ca
ndida albicans\"[ORGN] OR \"Candida tropicalis\"[ORGN] OR \"Candida parapsilosis\"[ORGN]) AND
\"internal transcribed spacer\"[All Fields] AND 400:800[SLEN]"
```

```
> candida6_ERG11_term
[1] "(\"Candida auris\"[ORGN] OR \"Candida glabrata\"[ORGN] OR \"Candida krusei\"[ORGN] OR \"Ca
ndida albicans\"[ORGN] OR \"Candida tropicalis\"[ORGN] OR \"Candida parapsilosis\"[ORGN]) AND
(ERG11[Gene] OR \"lanosterol 14-alpha demethylase\"[Title] OR \"lanosterol 14 alpha demethylase
\"[Title] OR CYP51[Gene]) AND 1000:2500[SLEN]"
```

Reference: https://www.ncbi.nlm.nih.gov/nucleotide/

Methods

Retrieved ITS (400-800 bp) and ERG11 (1000-2500 bp) sequences for 6 Candida species using rentrez.

Standardized species names. Balanced the dataset by selecting an equal number of sequences per species.

Measured distance matrices using both alignment-based distances and simple sequence features.

Performed hierarchical clustering, PCA, and NMDS to visualize how sequences cluster by species and resistance group.

Used silhouette indices to quantitatively compare clustering performance between ITS and ERG11.

Figure 1 – ITS NMDS

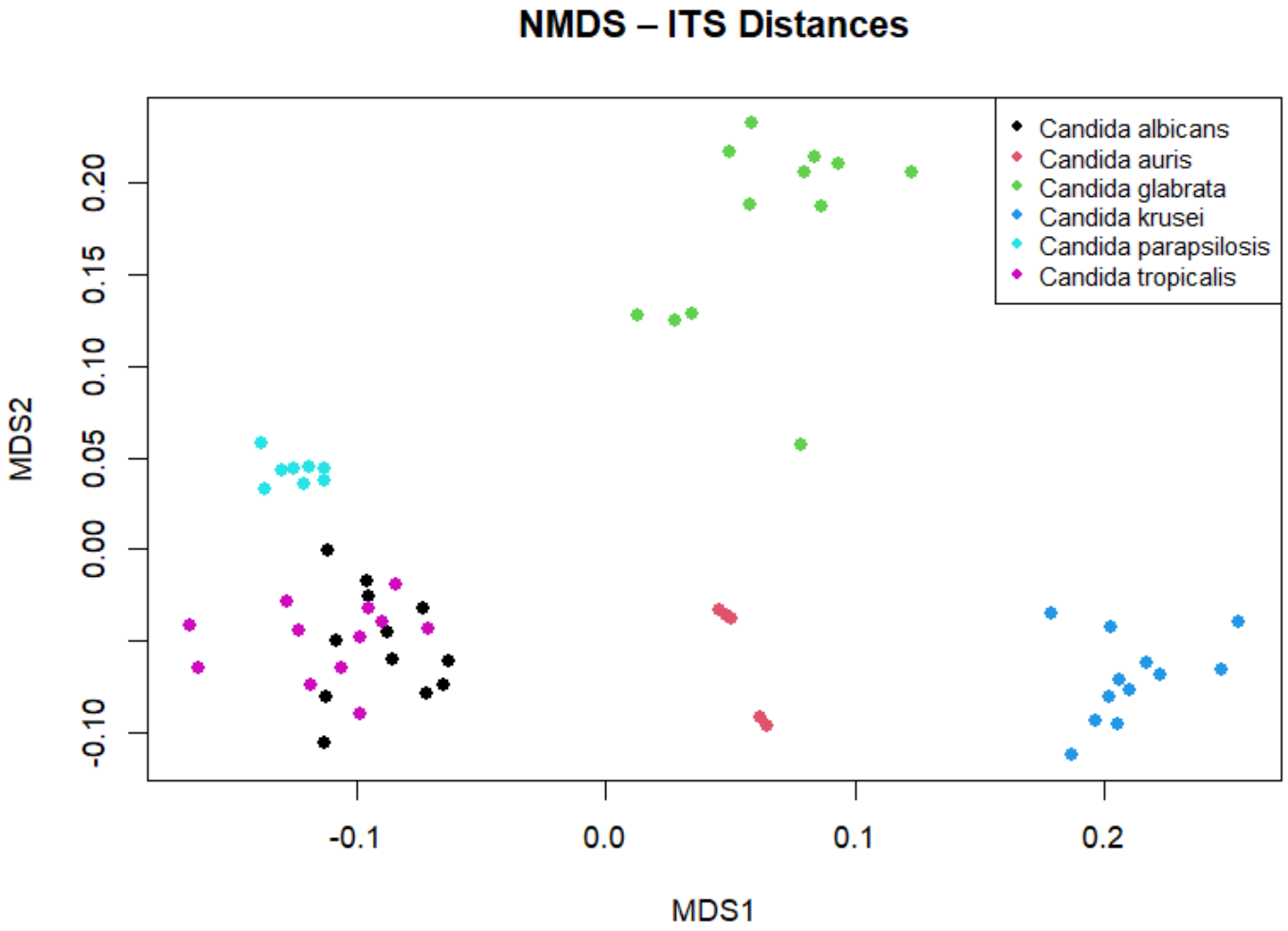
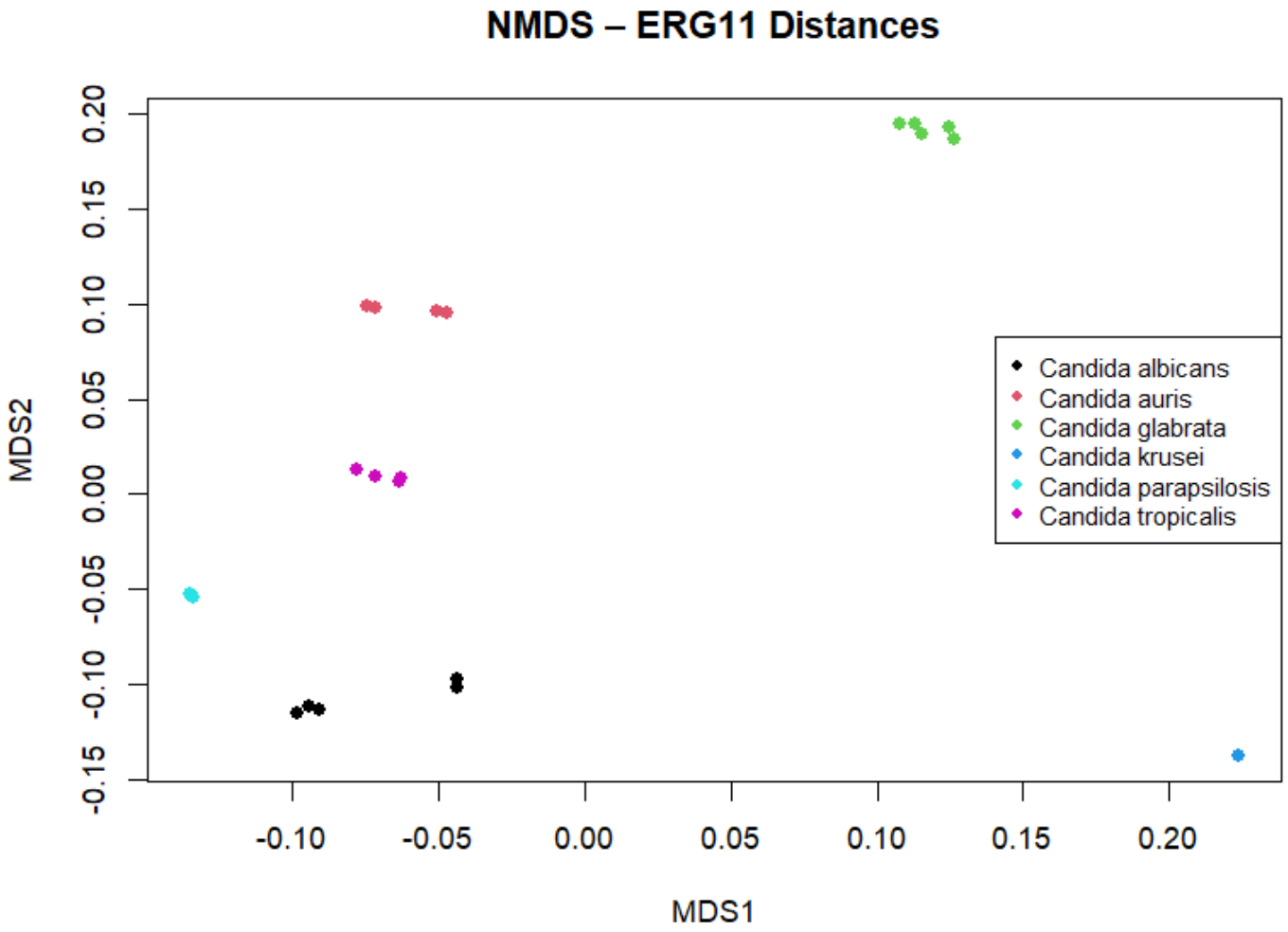


Figure 2 – ERG11 NMDS



ERG11 separates species more clearly than ITS, which shows overlap between resistance groups.

Non-metric Multidimensional Scaling (NMDS):

- **ITS** → Most species from their own regions except Candida tropicalis and Candida albicans, these ones overlap. (Figure 1)
  - The high resistance concern strains were isolated, low resistance had overlap.
- **ERG11** → Species clusters are more compact and more clearly separated. (Figure 2)
  - All were quite separated, no overlap, but also smaller sample size.

Were the results expected?

ERG11 overall performed better than ITS for species separation, which was unexpected.

Answer to the question 1:

High- and low-resistance *Candida* species do show distinct clustering patterns. ITS shows high resistance concern strains for when MDS1 > 0, and therefore low resistance concern strains on the left MSD1 < 0.

ERG11 showed the high resistance concern strains in the top right, above the descending diagonal, so not as good at clustering resistance. But the strength of separation increases in the ERG11 NMDS plot, providing clearer separation than ITS.

Figure 3 – ITS Feature-Based Clustering with PCA

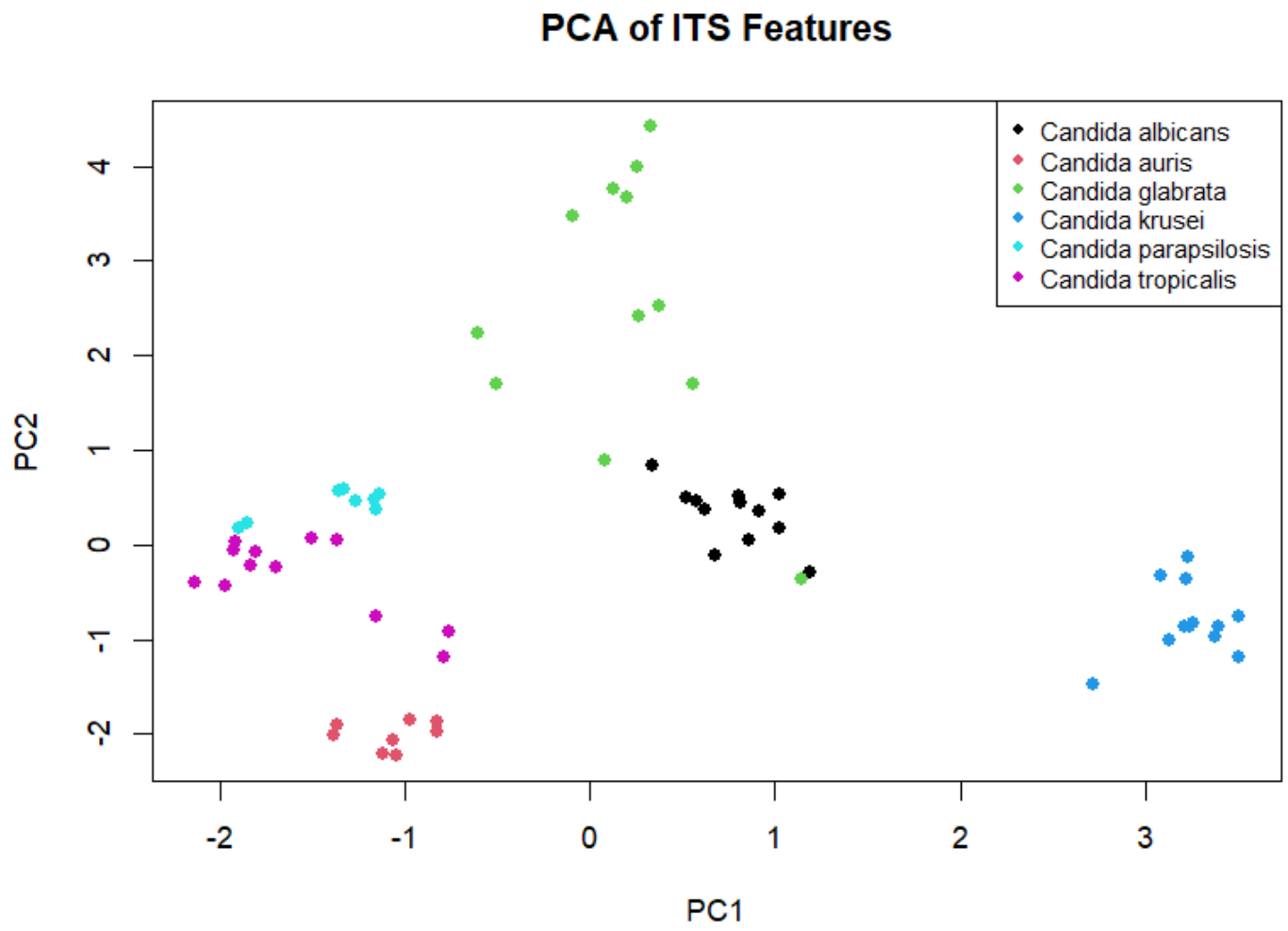
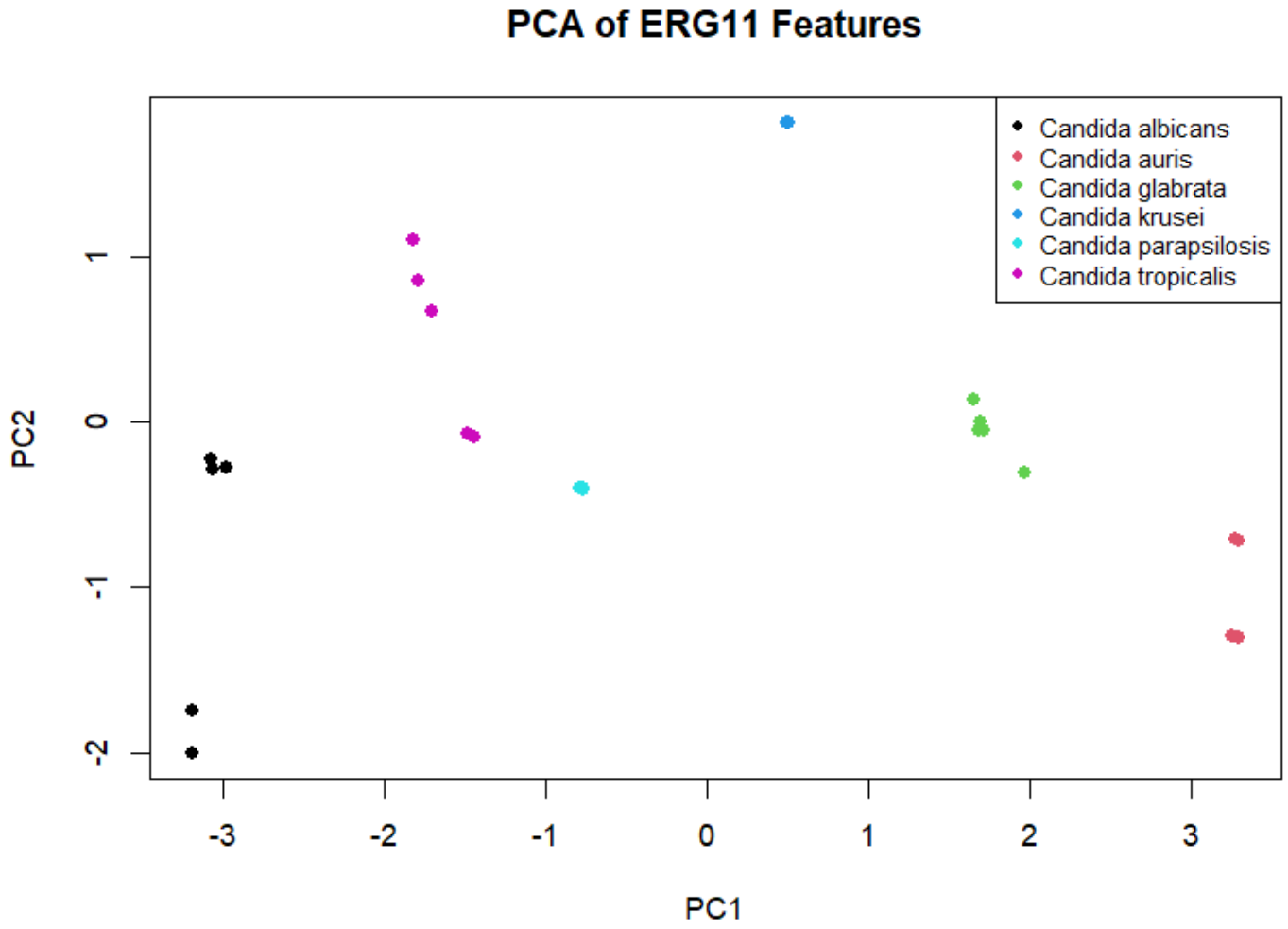


Figure 4 – ERG11Feature-Based Clustering with PCA



ERG11 PCA separates species more clearly than ITS, which shows overlap among several groups.

Feature-Based Clustering with PCA:

- **ITS** → Candida glabrata and Candida albicans overlap substantially. Candida tropicalis and Candida parapsilosis show partial overlap. Candida krusei has an isolated cluster. (Figure 3)
- **ERG11** → Has a small sample size, and species clusters do not overlap. Displays compact clusters. (Figure 4)

Were the results expected?

It was unexpected that ERG11, a conserved functional gene, would separate species more clearly than the highly variable ITS region.

Answer to the original question 1:

Candida species show clearer separation in ERG11 PCA than ITS PCA, indicating that feature-based clustering strongly depends on gene choice.

The high or low resistance concern species are isolated not clustering with each other in the ITS graph. ERG11, shows that a PC1 > 0 has the high resistance concern strains and PC1 < 0 has the low resistance concern strains.

Figure 5 – ITS Silhouette plot

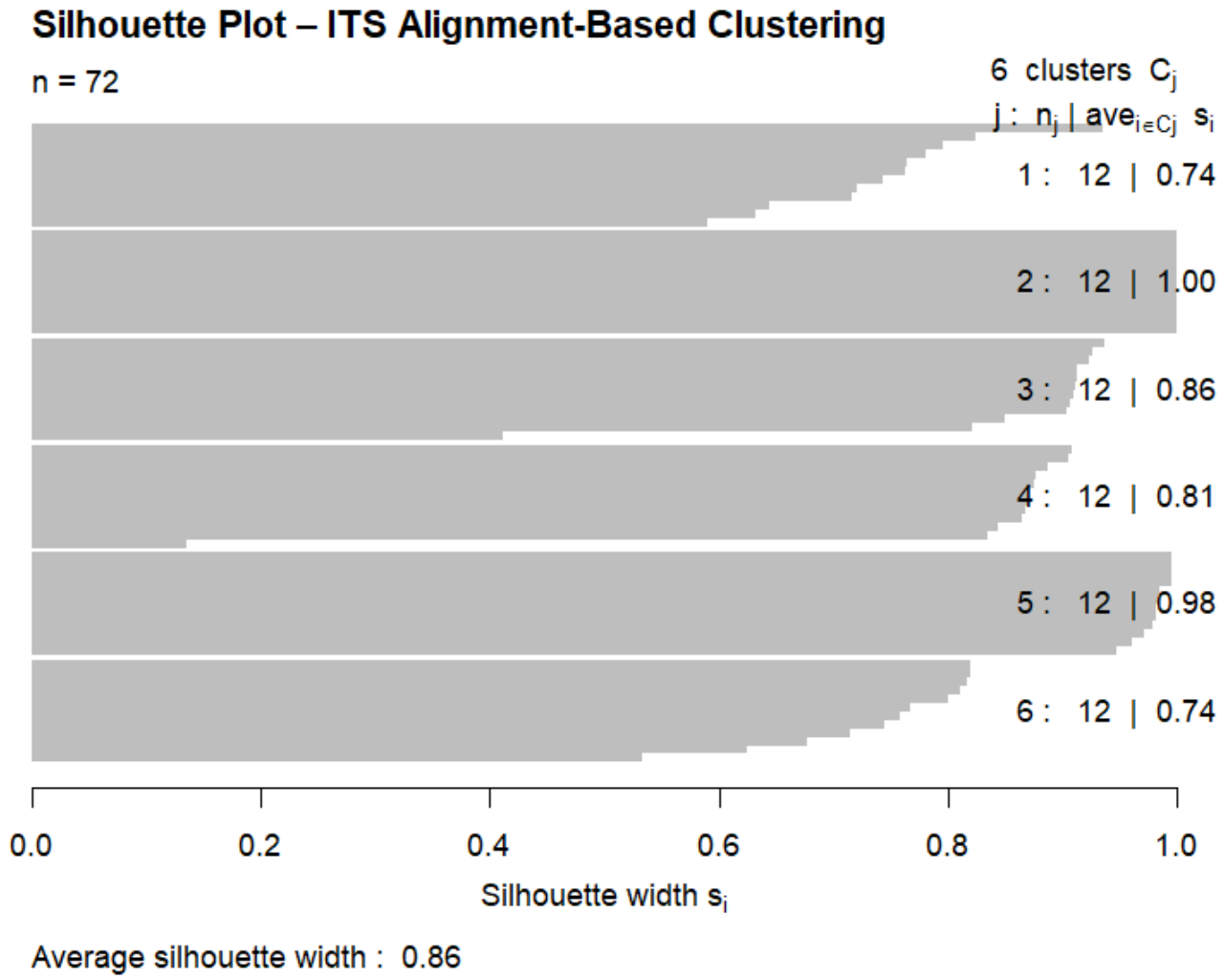
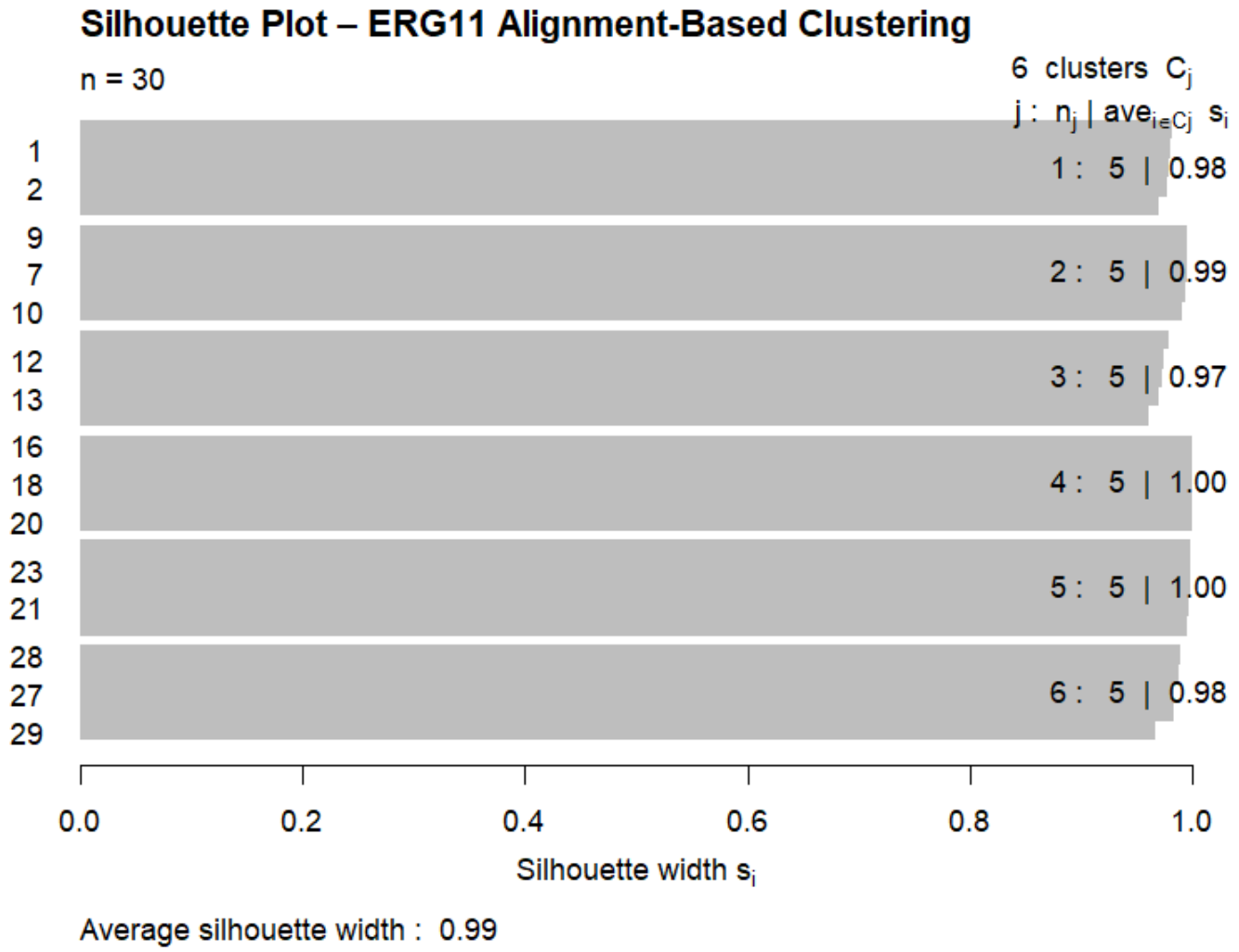


Figure 6 – ERG11 Silhouette plot



Silhouette Analysis Shows Stronger Clustering for ERG11 Than ITS.

Silhouette Analysis:

- **ITS** → Showed a good but variable performance 0.74 – 1.00, reflecting partial overlap among several species. (Figure 5)
- **ERG11** → Achieved a 0.99 silhouette width, which each species cluster scoring between 0.97-100. (Figure 6)

These scores indicate that ERG11 provides the strongest and most confident species separation, consistent with the NMDS and feature-based clustering.

Were the results expected?

it was unexpected that ERG11, a more conserved functional gene, would outperform ITS, the highly variable fungal barcode, by producing nearly perfect silhouette scores.

Answer to the original question 2:

Both genes show distinct clustering patterns between *Candida* species, but ERG11 separates species far more clearly and consistently than ITS.

Interpretation is limited by uneven sample sizes and database bias

One limitation in this project was the big **difference in sequence availability** between the two genes. ITS is a standard fungal barcode, so there were plenty of sequences to work with (Schoch et al., 2012).

ERG11, on the other hand, is mainly studied in the context of azole resistance (Whaley et al., 2017), so it appears less often in public databases. This created a **noticeable imbalance in sample size**.

Because of that, even though ERG11 looked like it produced cleaner clusters, it's hard to say whether this reflects true biological separation. This could just be an outcome of having fewer sequences. Smaller sample sizes are known to inflate silhouette widths and make clusters appear more distinct than they actually are (Wani et al., 2024).

So overall, while the ERG11 results were clustering the high and low resistant concern strains, future work should attempt to confirm this by **analyzing a much larger ERG11 dataset** or using curated fungal genome resources to validate the patterns.

Future work requires larger ERG11 datasets and testing more antifungal genes

Conclusion

- For NMDS, the ITS showed high resistance concerned strains MSD1 > 0, and low resistance concern strains MSD1 < 0, but ERG11 had a better species separation overall.
- The PCA plots had a similar results but with high resistance concern strains having a PC1 > 0 for ERG11 and the low resistance having PC1 < 0. The ERG11 showed better species separation overall and by resistance concern.
- Because the ERG11 sample size was small, the next step would be to collect more ERG11 sequences to test whether these patterns hold.

Preliminary Finding Worth Following Up

- I would also want to analyze additional antifungal genes to see whether functional genes consistently separate species better than ITS.
- The fact that each high and low-resistance species clustered cleanly is a valuable signal and worth exploring further in the field of antifungals.
- As antifungal resistance continues to rise, understanding how these genes vary across species is important, and this project provides an early look at those patterns.

Reflection

I chose to focus on clustering for this project because I already have more experience with supervised machine learning, and I wanted to push myself into something less familiar.

At the same time, I've recently started learning more about *Candida* and found the topic of antifungal resistance genuinely interesting, so this project felt like the perfect way to challenge myself while exploring a new research area. One of the biggest improvements for me was finally implementing a functions file having clean, reproducible code made everything so much more organized and easier to follow.

I definitely feel that my R skills have developed significantly through this course and these assignments. Overall, this project helped me grow both technically and personally, and I'm excited to bring these concepts and workflows into BINF*6999 and eventually into my future career.

Acknowledgements slide

I would like to acknowledge Sharon, for being readily available to answer questions and share ideas with. Sharon was also helpful in reminding me of specific rubric requirements.

I would also like to acknowledge Saira, for being responsive with emails and running my code prior to ensure my functions worked correctly!

References

Arendrup, M. C., & Patterson, T. F. (2017). Multidrug-resistant *Candida*: Epidemiology, molecular mechanisms, and treatment. *The Journal of Infectious Diseases, 216*(suppl_3), S445–S451. https://doi.org/10.1093/infdis/jix131

Chowdhary A, Sharma C, Meis JF (2017) *Candida auris*: A rapidly emerging cause of hospital-acquired multidrug-resistant fungal infections globally. PLoS Pathog 13(5): e1006290. https://doi.org/10.1371/journal.ppat.1006290

de Oliveira Ceita, G., Vilas-Boas, L. A., Castilho, M. S., Carazzolle, M. F., Pirovani, C. P., Selbach-Schnadelbach, A., Gramacho, K. P., Ramos, P. I., Barbosa, L. V., Pereira, G. A., & Góes-Neto, A. (2014). Analysis of the ergosterol biosynthesis pathway cloning, molecular characterization and phylogeny of lanosterol 14 α-demethylase (ERG11) gene of Moniliophthora perniciosa. *Genetics and molecular biology*, *37*(4), 683–693. https://doi.org/10.1590/S1415-47572014005000017

Nascimento, T., Inácio, J., Guerreiro, D., Diaz, P., Patrício, P., Proença, L., Toscano, C., & Barroso, H. (2024). Susceptibility patterns of candida species collected from Intensive Care Units in Portugal: A prospective study in 2020–2022. *Infection Prevention in Practice*, *6*(4), 100403. https://doi.org/10.1016/j.infpip.2024.100403

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., Fungal Barcoding Consortium, & Fungal Barcoding Consortium Author List (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), 6241–6246. https://doi.org/10.1073/pnas.1117018109

Wani A. A. (2024). Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions. *PeerJ. Computer science*, *10*, e2286. https://doi.org/10.7717/peerj-cs.2286

Whaley, S. G., Berkow, E. L., Rybak, J. M., Nishimoto, A. T., Barker, K. S., & Rogers, P. D. (2017). Azole Antifungal Resistance in *Candida albicans* and Emerging Non-*albicans Candida* Species. *Frontiers in microbiology*, *7*, 2173. https://doi.org/10.3389/fmicb.2016.02173

Programming Resources:

prcomp function | R Documentation. (n.d.). Www.rdocumentation.org. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp

cluster.stats function - RDocumentation. (2024). Rdocumentation.org. https://www.rdocumentation.org/packages/fpc/versions/2.2-13/topics/cluster.stats

metaMDS function - RDocumentation. (2025). Rdocumentation.org. https://www.rdocumentation.org/packages/vegan/versions/2.7-2/topics/metaMDS

hclust function - RDocumentation. (n.d.). Www.rdocumentation.org. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust

R Functions. (n.d.). Www.w3schools.com. https://www.w3schools.com/r/r_functions.asp