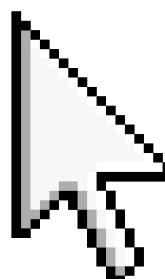


Home Content Contact

How long is a

Barcode?

EXPLORING SEQUENCE LENGTHS AND
DEPOSITOR PATTERNS IN LYCOSIDAE



Start

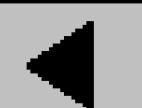




WHAT ARE DNA BARCODES?

DNA barcodes are short, standardized genomic regions used to identify species, relying on marker sequences that enable accurate classification.

Using the correct marker length ensures sufficient phylogenetic signal without introducing errors from over-trimming, mislabeling, or sequencing artifacts. Deviations from expected length can flag potential quality control issues in public databases.



MARKER TYPES

COI-5P / COI-3P ~660 bp (Aly, 2014)

- The gold standard for animal barcoding; long enough to distinguish species, short enough for efficient sequencing.

ITS1 / ITS2 ~300 bp (Zhao, 2018)

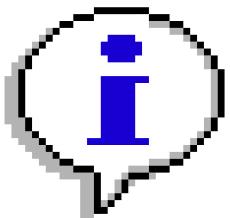
- Commonly used in fungi and some animals; shorter, with high variation between species.

18S rRNA ~1800 bp (Rix, 2008)

- Highly conserved; useful for broad taxonomic resolution, but too long for standard barcoding workflows.



WHY DOES SEQUENCE LENGTH MATTER?



- Sequence length can act as a proxy for barcode quality.
- Short sequences may lack the resolution needed to distinguish between species.
- Long sequences might reflect artifacts or poor trimming.
- If some labs consistently submit shorter or longer sequences, this could point to:
 - Protocol variation
 - Instrument differences
 - Data-cleaning issues
- This analysis asks whether such hidden biases exist, and whether they matter.



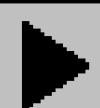
PUBLIC BARCODE DATABASES RELY ON TRUST

Understanding who submits sequences and whether they follow expected standards can help maintain data reliability and detect quality issues before they spread through downstream research.



WOLF SPIDERS

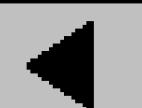
- Lycosidae, commonly known as wolf spiders, comprise over 2,400 species distributed worldwide. They are ground-dwelling, fast-moving predators found in diverse habitats, from forests and grasslands to deserts and wetlands (Piacentini, 2019)
- Wolf spiders play a vital role in ecosystem balance, regulating insect populations and contributing to soil health.
- Morphological similarity between species makes traditional identification difficult. Subtle variations in coloration and size often overlap across genera.
- Molecular tools provide a reliable means of distinguishing closely related species. Accurate barcoding helps improve biodiversity databases, biogeographic mapping, and ecological monitoring.



HYPOTHESIS

If the Lycosidae dataset represents consistent and standardized barcoding practices, then the majority of depositor submissions should cluster tightly around the expected COI barcode length (~660 bp) with minimal skew.

By comparing depositor-specific distributions, I aimed to test whether barcode consistency reflects standardized sequencing practices or institutional variability.



OBJECTIVES

1. Examine distribution of barcode lengths across depositors.
2. Identify whether some institutions disproportionately contribute short or non-standard sequences.
3. Use sequence length bins to summarize depositor trends.
4. Explore whether depositor-level summaries could help flag data quality in large barcode datasets.



PHASE 1

Download Data

Imported Lycosidae dataset from BOLD

Viewed + Cleaned

dataset (removed NAs, standardized columns)

Generate Features

Calculated sequence lengths & created bins (Short / Medium / Long)

PHASE 2

Focus on COI-5P

Filtered dataset to main barcode marker

Define Bins

Clustered sequences around COI length
(0-620 | 620-700 | 700+)

Group by Depository

Combined top 4 depositories + Other

PHASE 3

Boxplot by Marker Gene

Showed variation & outliers across markers

Bar Plot by Depository

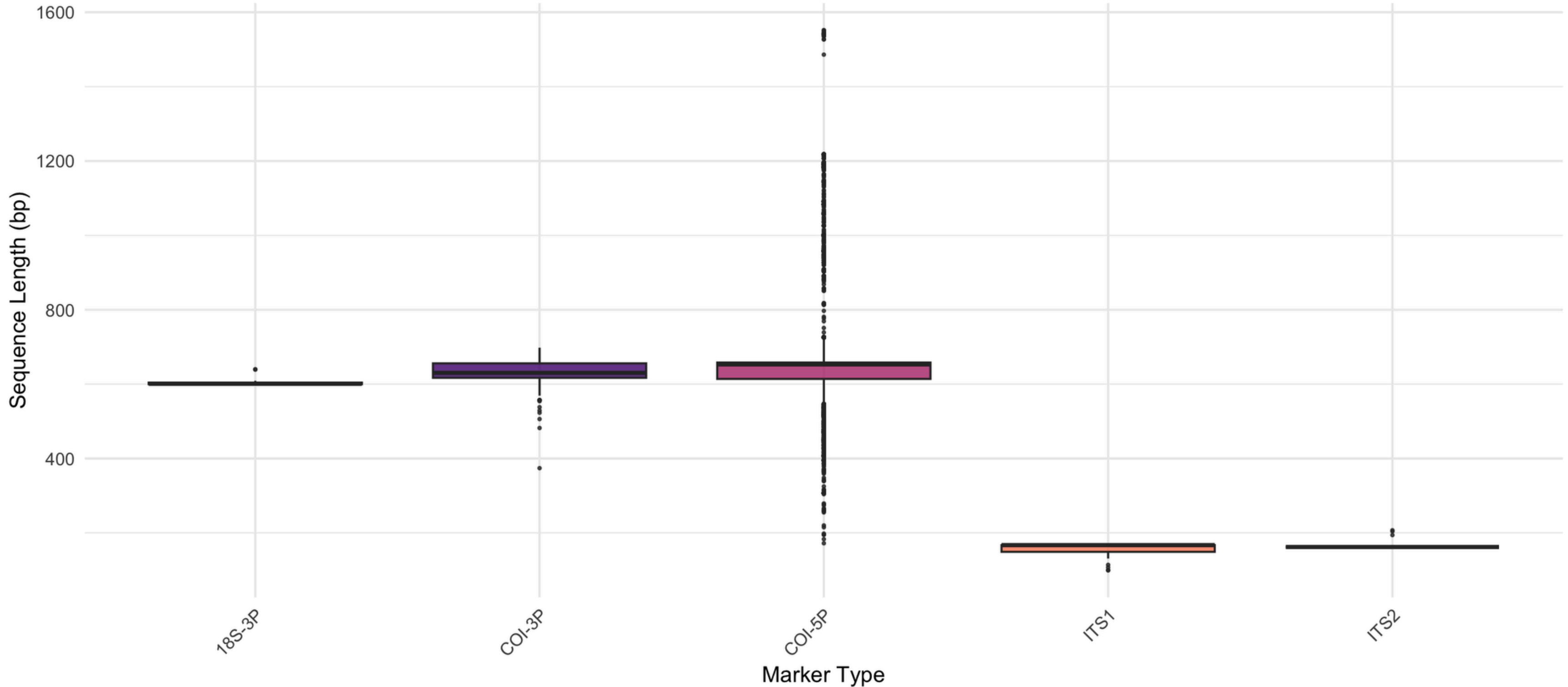
Compared raw counts in each length bin

Heatmap of Proportions

Visualized within-depository sequence-length distribution



SEQUENCE LENGTH DISTRIBUTION BY MARKER TYPE



DISCUSSION

This boxplot confirms that most markers cluster tightly around their expected lengths:

- COI-3P and COI-5P hover around the ideal ~660 bp, showing strong consistency and supporting barcode suitability.
- ITS1 and ITS2 are shorter, ~300 bp, as expected for nuclear ribosomal spacers.
- 18S-3P appears truncated (~660 bp), suggesting potential trimming or sequencing limitations, as full 18S is typically ~1800 bp.

These distributions reflect generally good data quality across markers.

Minimal spread and few extreme outliers suggest low sequencing error or trimming issues at the marker level.

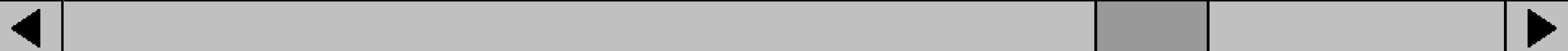
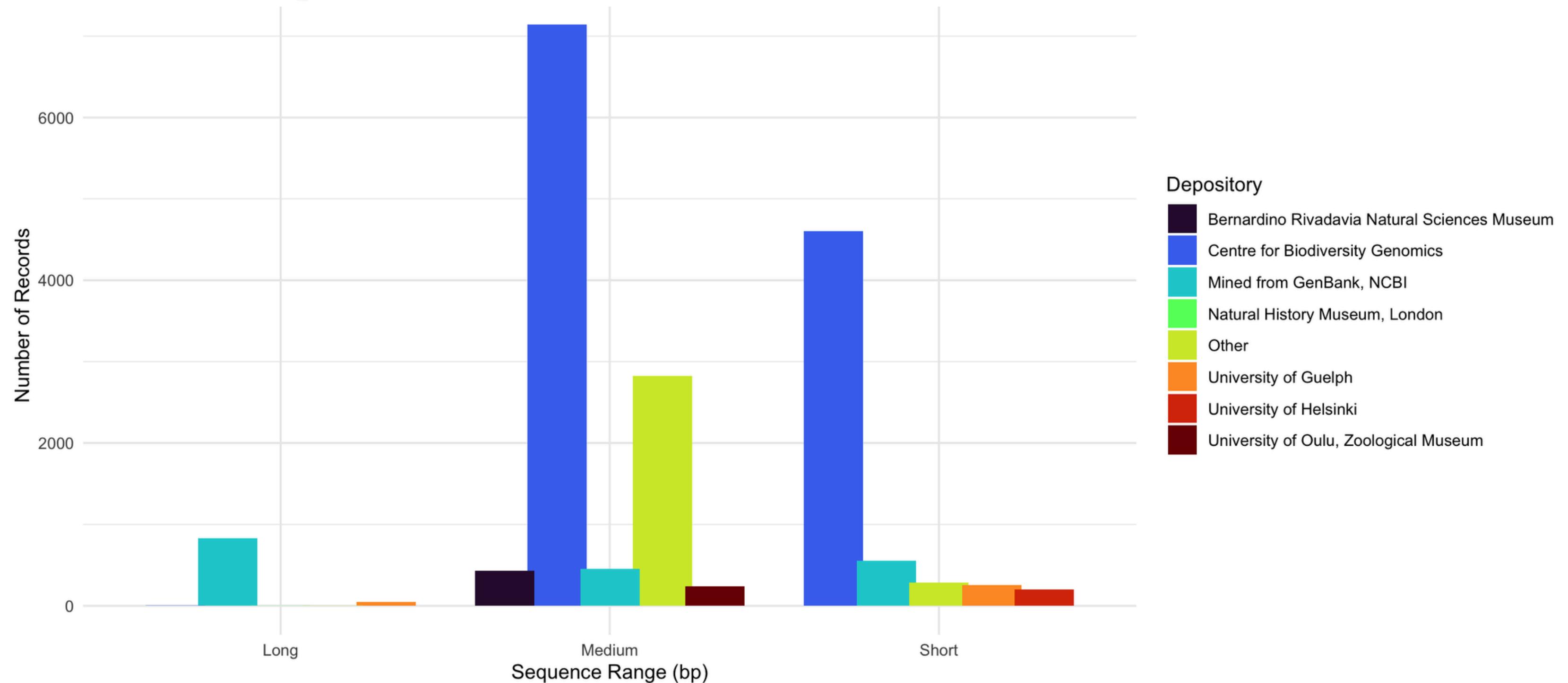
However, without linking sequences back to depositor IDs, we cannot identify which institutions may be contributing disproportionately short or long sequences.

This figure supports the hypothesis that sequence length patterns reflect marker-specific expectations, and that most submissions adhere to standard barcode guidelines.

Incorporating depositor overlays in future QC pipelines could enhance flagging of atypical submissions.



RAW SEQUENCE COUNT BY BIN AND DEPOSITOR



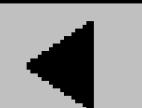
DISCUSSION

This bar plot explores how sequence length distributions vary across data sources (top 5 depositaries per range). Most records fall within the medium (620–700 bp) range, aligning with expected COI barcode lengths and confirming overall data reliability.

However, a few institutions show imbalances, some contributing noticeably more short or long sequences. For example, the Centre for Biodiversity Genomics (CBG) dominates submissions, producing a strong concentration of both, medium and short length reads. In contrast, GenBank includes an even distribution of reads from all three categories

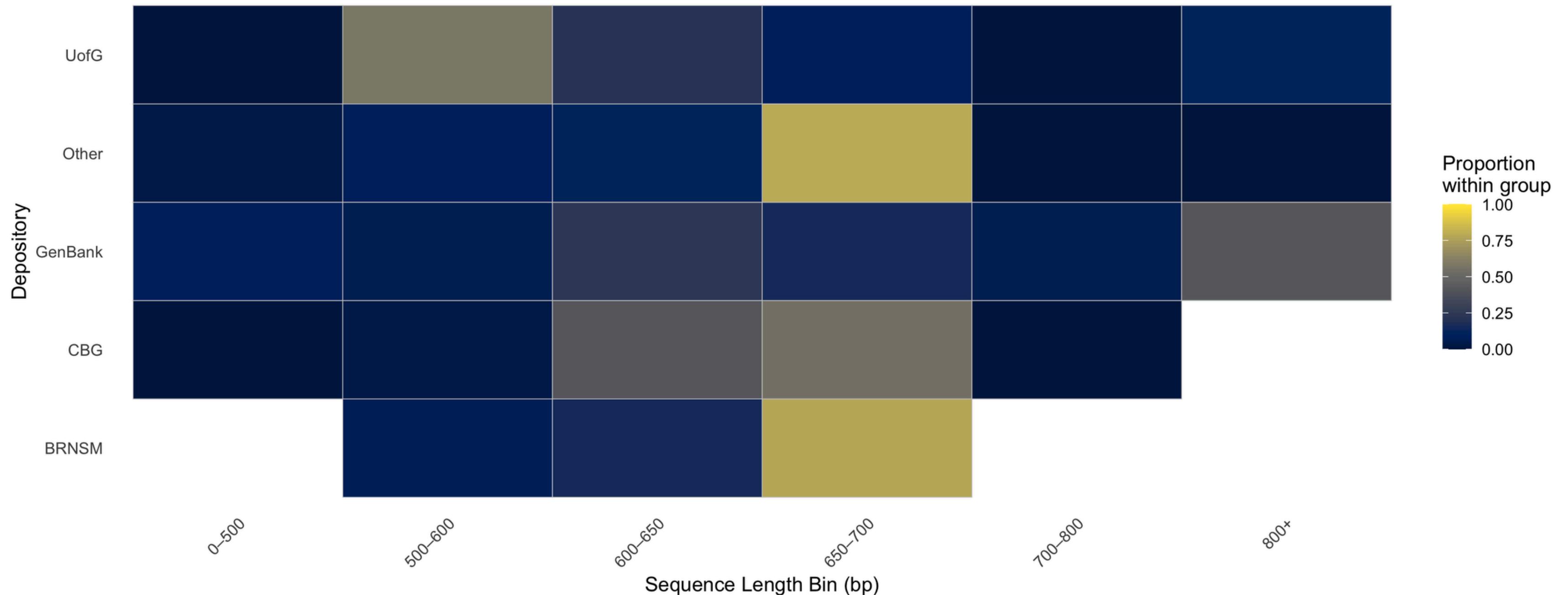
This plot reveals that while marker consistency is strong, submission practices differ across depositaries, hinting that institutional workflows, limitations in technology, or sequencing pipelines can influence barcode completeness.

The next visualization (heatmap) narrows in on COI-5P specifically to examine these depositor-level trends in finer detail.



COI–5P SEQUENCE LENGTH DISTRIBUTION BY DEPOSITORY (TOP 4 + OTHER)

Each row normalized by total submissions (proportion within depository)



DISCUSSION

After identifying depositor-level differences in sequence length distributions, this heatmap focuses exclusively on COI-5P, the dominant barcode marker, to visualize proportional quality patterns across the top four contributors.

Each row represents a depository, normalized by its own submissions. Brighter tiles mark higher proportions of sequences in that length bin

clear contrasts emerge: Bernardino Rivadavia Natural Sciences Museum (BRNSM) and CBG show strong clustering within the 650–700 bp range – indicative of consistent, high-quality barcoding workflows.

In contrast, GenBank and University of Guelph display more fragmented profiles, with heavier representation in shorter or longer bins, suggesting mixed data imports, legacy sequences, or less standardized protocols.



REFLECTION

This visualization demonstrates that while the overall Lycosidae dataset is largely reliable, barcode completeness and sequence quality vary by source. These depositor-specific biases emphasize the potential for harmonized submission standards and the integration of automated quality-control (QC) pipelines within BOLD. Such pipelines could flag sequences that deviate from expected length ranges, improving data consistency before downstream use (Kress, 2008).

With a larger dataset, an expanded boxplot comparing sequence length by depositor across the top five markers could provide a more comprehensive quality overview. This single figure would effectively summarize everything achieved through this three-step analysis—showing which depositories contribute in-range, truncated, or over-extended sequences, and how much variability exists within each.

The main drawback is visual complexity: displaying multiple markers and depositories would require a sufficiently large dataset and careful design to avoid clutter. However, the advantage is clear: a single, high-density figure that captures marker-specific consistency, depositor behavior, and overall data reliability at a glance, providing a powerful tool for large-scale barcode data validation.



ACKNOWLEDGEMENTS

- My sister - for reviewing my slides, finding all grammatical and logic errors, endured all spider pictures and listened to me talk about this for 13+ hours
- Yumna - for checking my code on her computer to ensure it ran smoothly and was ready for submission, and for always offering helpful edits and moral support.
- Stephanie - for introducing me to how heatmaps work and helping me understand how to interpret them properly.
- Brendan - for teaching me how to use grep() and other search commands in R and saving me from hours of confusion.

<https://ggplot2.tidyverse.org/reference/index.html>

<https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/filter>

https://dplyr.tidyverse.org/reference/group_by.html

<https://dplyr.tidyverse.org/reference/summarise.html>

<https://www.pluralsight.com/courses/r-ggplot2-data-visualization>

<https://stackoverflow.com/questions/62094782/how-to-draw-a-grouped-barplot-of-two-dataframes>

https://ggplot2.tidyverse.org/reference/scale_viridis.html



REFERENCES

- Piacentini LN, Ramírez MJ. Hunting the wolf: A molecular phylogeny of the wolf spiders (Araneae, Lycosidae). *Mol Phylogenet Evol.* 2019 Jul;136:227-240. doi: 10.1016/j.ympev.2019.04.004. Epub 2019 Apr 4. PMID: 30953780.
- Aly SM. Reliability of long vs short COI markers in identification of forensically important flies. *Croat Med J.* 2014 Feb;55(1):19-26. doi: 10.3325/cmj.2014.55.19. PMID: 24577823; PMCID: PMC3944415.
- Zhao LL, Feng SJ, Tian JY, Wei AZ, Yang TX. Internal transcribed spacer 2 (ITS2) barcodes: A useful tool for identifying Chinese *Zanthoxylum*. *Appl Plant Sci.* 2018 Jun 15;6(6):e01157. doi: 10.1002/aps3.1157. PMID: 30131899; PMCID: PMC6025816.
- Rix, M. G., Harvey, M. S., & Roberts, J. D. (2008). Molecular phylogenetics of the spider family micropholcommatidae (Arachnida: Araneae) using nuclear rrna genes (18s and 28s). *Molecular Phylogenetics and Evolution*, 46(3), 1031-1048.
<https://doi.org/10.1016/j.ympev.2007.11.001>
- Kress WJ, Erickson DL. DNA barcodes: genes, genomics, and bioinformatics. *Proc Natl Acad Sci U S A.* 2008 Feb 26;105(8):2761-2. doi: 10.1073/pnas.0800476105. Epub 2008 Feb 19. PMID: 18287050; PMCID: PMC2268532.

