Comparing Tardigrada barcode sequence lengths across realms

THOMAS FURTADO

BINF*6210

https://www.nationalgeographic.com/animals/invertebrates/facts/tardigrades-water-bears

Tardigrada

- Tardigrades are microscopic animals (<1mm) that are capable of withstanding extremely harsh environmental conditions
- Tardigrada, commonly referred to as the “water bear”, occupy a wide variety of aquatic environments across the planet.
- Despite their small size, they exhibit remarkable biodiversity and evolutionary adaptations.

(Morek et al., 2021)

Their genetic diversity is therefore challenging to explore with morphology alone, *DNA barcoding* provides us with the opportunity to do so

DNA barcoding and COI-5P

DNA barcoding:

- A method that allows us to manage and analyze taxonomic differences which are not apparent from morphology alone
- The DNA “barcode” is a biological identifier using the COI gene

COI-5P

- COI-5P is the 5-prime end of COI, a mitochondrial gene which is used for DNA barcoding
- It is conserved enough to design primers, but varies enough to distinguish species
- The **sequence lengths** vary, revealing different species, populations or evolutionary patterns (Morek et al., 2021).

Examining COI sequence lengths can reveal evolutionary similarities and differences between two populations. It can be used as a simple yet effective way to explore evolutionary relationships and potentially spark further inquiry

Exploring COI-5P sequence length across realms

Data acquired from BOLD contained samples collected from a variety of realms. The two most data-plentiful realms are:

- **Nearctic** (North America)
- **Palearctic** (Europe and North Africa)

Comparing COI sequence lengths helps uncover evolutionary relationships between populations and can inspire deeper investigation

Therefore, I ask the question:

Do COI-5P sequence lengths in Tardigrada differ across Nearctic and Palearctic realms, and what patterns, if any, emerge between them?

This question is interesting because the two realms have distinct geographical and ecological characteristics, and by examining sequence length I could reveal if this gene is maintained or mutated across populations under the differing ecological conditions.

Methods: visualizing workflow

1. Data Acquisition

- Searched for suitable data
- Collected meta-data from BOLD

2. Data Import & Inspection

- Imported meta-data into R
- Explored realm & base count data
- Identified caveats of data

3. Data Refinement

- Removed outliers: less than 200 bp, more than 900 bp (Guo et al., 2022)
- Then filtered base counts by select *realms* and *marker code*

(Nearctic & Palearctic) **(COI-5P)**

4. Data Exploration & Analysis

- Summarized: mean, median *etc.* sequence length by select realm
- Visualized & compared sequence length distributions with violin plot

4. Interpretation of Results

- Further investigated nuances in the distribution data using a ridgeline plot.
- Performed statistical analysis of variance in COI sequence length among realms (F-test)
- Acknowledged sample size limitations

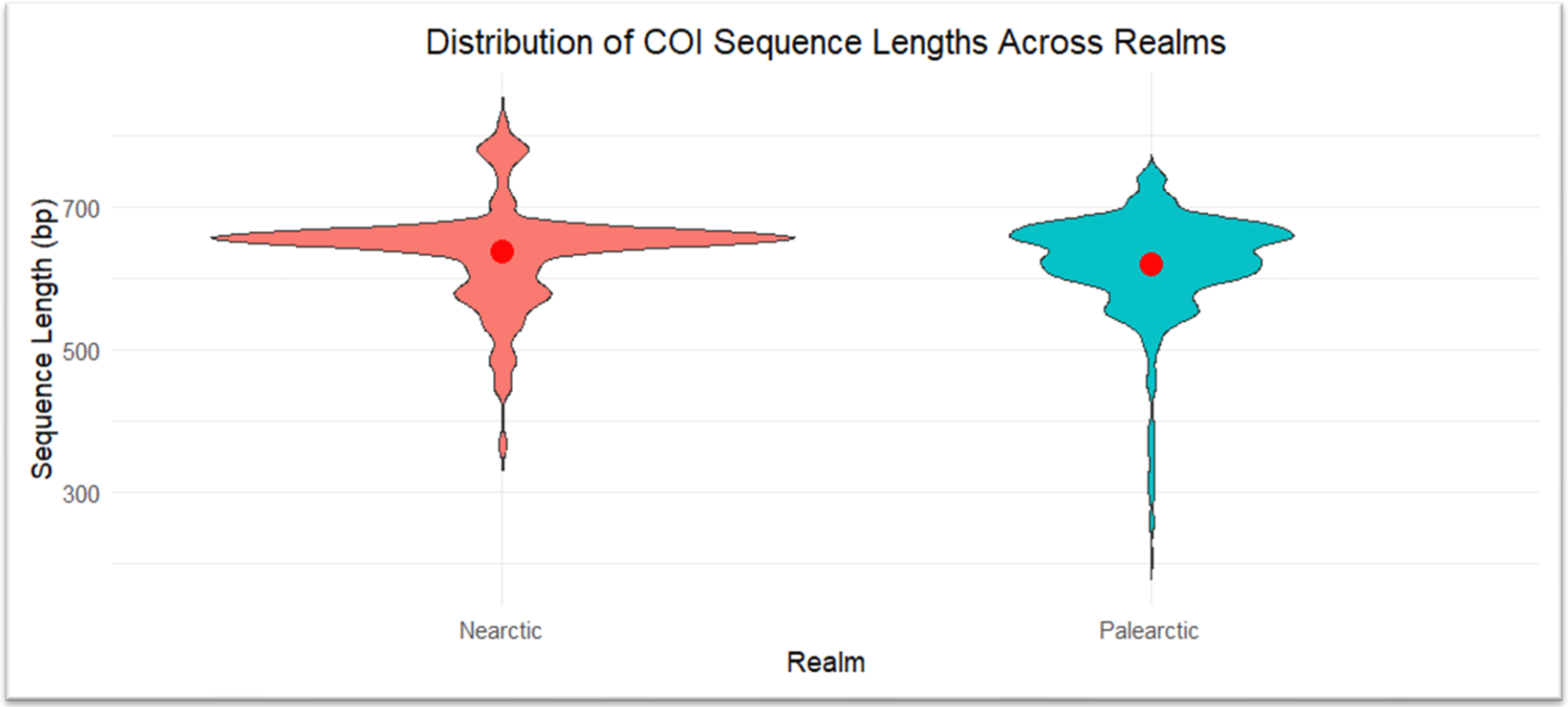
COI-5P sequence lengths are broadly similar, with Nearctic sequences slightly longer on average

Realm	Count	Mean Length	Median Length	Min Length	Max Length
Nearctic	114	637.8333	658	366	818
Palearctic	635	620.9181	631	206	742

Figure 1. Summary statistics of COI-5P sequence lengths by realm

Note the sample imbalance!

Overall distributions are broadly similar but differ in shape, Palearctic sequences are more variable near the mean



Note: This violin plot measures *density,* not amount of raw data points

Figure 2. Distribution of COI-5P sequence lengths by realm

Multiple peaks suggest clusters of seemingly conserved COI-5P lengths, with Nearctic sequences appearing slightly longer on average

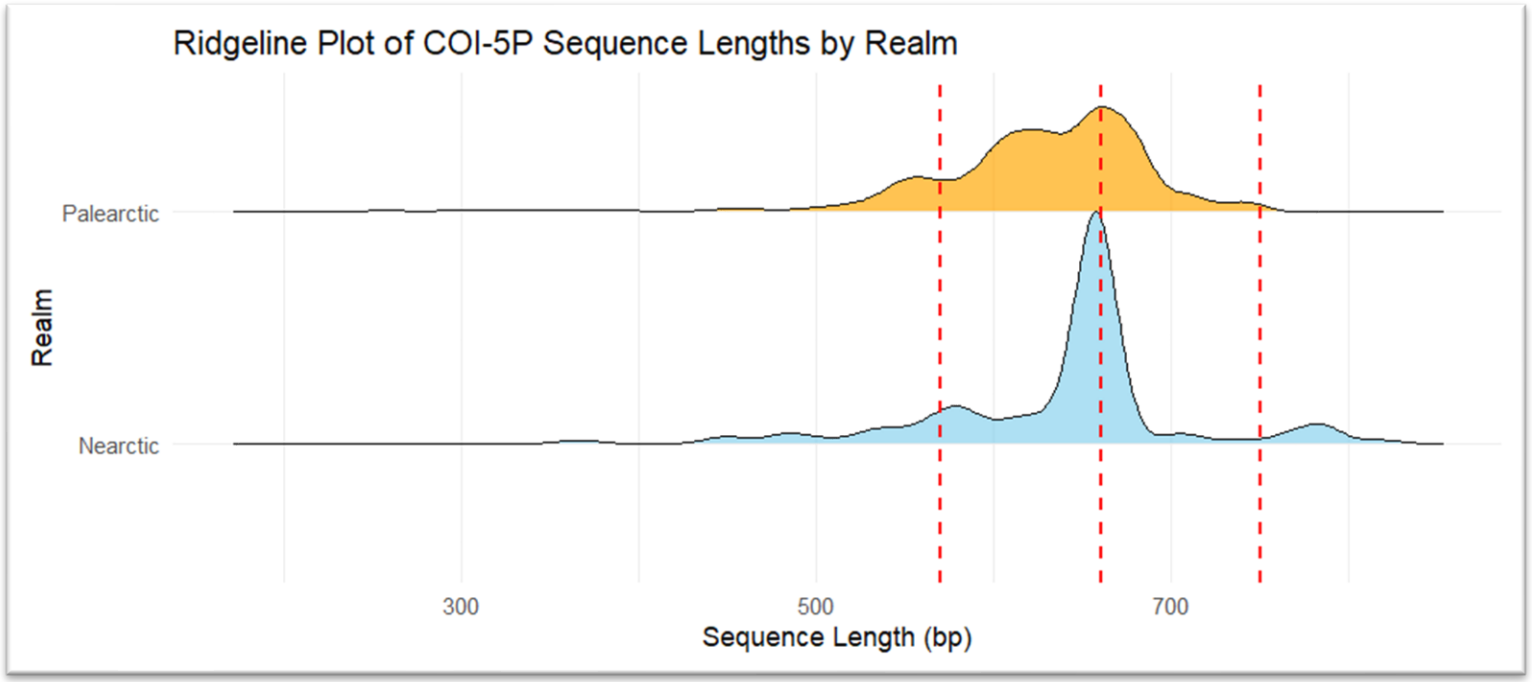


Figure 3. Ridgeline plot annotated with conserved sequence lengths among the Nearctic and Palearctic realms

F-test concludes that the variance is **NOT** statistically significant (p = 0.64)

Main result: there are no significant differences in COI sequence length, but sequence length patterns may be present

Key findings and interpretation: patterns in COI sequence length between realms

Main Results

- Mean sequence lengths were similar between realms (Figure 1)
- Palearctic sequence lengths are more variable around the mean, while Nearctic sequence appear more clustered (Figure 2)
- Both realms have approximately 3 distinguishable clusters of COI lengths, with Nearctic sequence lengths appearing slightly longer on average (Figure 3)
- Despite this, an F-test proved the difference between the two is **not statistically significant** ($P > 0.05$)

Interpretation

- The peaks may indicate **distinct haplotypes** with varying COI length, *or* it could be due **to sequencing inconsistencies**
- The ridge plot (Figure 3) revealed multiple relatively consistent peaks (570, 660, and 750 bp) which could suggest there are multiple distinct lengths that the COI gene exists at.
- The clusters shown in Figure3 may allude to a novel idea: ~3 conserved sequence length clusters, which is a pattern not currently found discussed in the literature
- This length variation in Nearctic sequences *could* have biological meaning, alluding to a potential species-level divergence

Limitations and considerations

Unexpected findings:

- despite having a **smaller sample size**, Nearctic samples showed pronounced clustering around the mean, contrary to "expected" higher variance levels.
- This alludes to a possibility of a significant difference in COI lengths

Limitations and Caveats:

- Sample imbalance:
- **n = 635** (Palearctic) and **n = 114** (Nearctic) likely had an impact on visualization and *statistical power*
- Though the data presented in Figure1 and Figure3 seems to begin revealing differences, a higher sample size is required to provide a more statistically accurate confirmation or denial of this notion
- COI-5P length might not vary strongly among taxa, so the observed differences might only reveal inconsistent data handling, and not true evolutionary differences

Lack of COI variance does not rule out biological differences!

COI-5P consistency across realms aligns with Kayastha et al. (2023), showing that COI variation often remains low even when morphology or geography differ.

This suggests that COI lengths are strongly conserved, and subtle biological changes may *still occur* without drastic change to the marker

Future directions and broader implications

Next steps

1. Expand dataset to encapsulate other realms, to conduct global comparison
2. Ensure sufficient and consistent data collection to accurately represent all realms of tardigrade
3. Incorporate use of phylogenetic data and investigate the idea of conserved COI length clusters in tardigrade

This could provide insight into the epigenetic trends of tardigrade, or other taxa spread out across global realms!

Broader significance

- Understanding the COI length variation could help further streamline and standardize the use of DNA barcodes
- This will also allow for elucidating the true biodiversity of tradigrada and many other highly variable taxa
- This work can provide a future foundation for understanding the evolution of DNA sequence lengths on a molecular basis, and how it relates to biodiversity

References

Guo, M., Yuan, C., Tao, L., Cai, Y., & Zhang, W. (2022). Life barcoded by DNA barcodes. *Conservation Genetics Resources*, *14*(4), 351–365. https://doi.org/10.1007/s12686-022-01291-2

Kayastha, P., Szydło, W., Mioduchowska, M., & Kaczmarek, Ł. (2023). Morphological and genetic variability in cosmopolitan tardigrade species—Paramacrobiotus Fairbanksi Schill, Förster, Dandekar & Wolf, 2010. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-42653-6

Morek, W., Surmacz, B., López, A., & Michalczyk, L. (2021). 'everything is not everywhere': Time-calibrated phylogeography of the genus milnesium (tardigrada). *Molecular Ecology*, *30*(14), 3590–3609. https://doi.org/10.22541/au.161073927.77224986/v1

yuzaR Data Science (2023, Nov. 24). *Master Box-Violin Plots in {ggplot2} and Discover 10 Reasons Why They Are Useful* [Video]. YouTube. https://www.youtube.com/watch?v=rvm94zcoKT0