Open Access | https://doi.org/10.25080/NVYF1037

Ecological and Spatial Influences on the Genetics of Cumacea (Crustacea: Peracarida) in the Northern North Atlantic

Justin Gagnon[1], and **Nadia Tahiri**[2]

[1]Department of Biology, University of Sherbrooke, 2500, boul. de l'Université, Sherbrooke, Quebec, J1K 2R1 Canada, [2]Department of Computer Science, University of Sherbrooke, 2500, boul. de l'Université, Sherbrooke, Quebec, J1K 2R1 Canada

SciPy 2024

July 8 - July 14, 2024

Proceedings of the 23rd
Python in Science Conference
ISSN: 2575-9752

Published Jul 10, 2024

Correspondence to
Justin Gagnon
Justin.Gagnon2@USherbrooke.ca

Abstract

The peracarid taxon Cumacea is an essential indicator of benthic quality in marine ecosystems. This study investigated the influence of environmental (i.e., biological or ecosystemic), climatic (i.e., meteorological or atmospheric), and spatial (i.e., geographic or regional) variables on their genetic variability and adaptability in the Northern North Atlantic, focusing on Icelandic waters. We analyzed partial sequences of the 16S rRNA mitochondrial gene from 62 Cumacea specimens. Using the *aPhyloGeo* software, we compared these sequences with relevant variables such as latitude (decimal degree) at the end of sampling, wind speed (m/s) at the start of sampling, O_2 concentration (mg/L), and depth (m) at the start of sampling.

Our analyses revealed variability in spatial and biological variables, reflecting the diversity of ecological requirements and benthic habitats. The most common Cumacea families, Diastylidae and Leuconidae, suggest adaptations to various marine environments. Phylogeographic analysis showed a divergence between specific genetic sequences and two habitat variables: wind speed (m/s) at the start of sampling and O_2 concentration (mg/L). This observation may indicate the possibility of varying local adaptations in response to these fluctuating conditions.

These results reinforce the importance of further research into the relationship between Cumacea genetics and global environmental variables to interpret the evolutionary dynamics and adaptation of these deep-sea organisms. This study sheds much-needed light on the acclimatization of invertebrates to climate change, anthropogenic pressures, and marine habitat management, potentially contributing to the evolution of more effective conservation strategies and policies to protect these vulnerable ecosystems.

The *aPhyloGeo* Python package is freely and publicly available on GitHub and PyPi, providing an invaluable tool for future research.

Keywords Adaptation, Atlantic, Bioinformatics, Biology, Cumacea, Iceland, Phylogeography

1. Introduction

The North Atlantic and Subarctic regions, particularly the Icelandic waters, are of ecological interest due to their diverse water masses and unique oceanographic features [1], [2], [3]. These areas form vital benthic habitats[1] [4] and enhance our understanding of deep-sea ecosystems and biodiversity patterns [3], [5], [6]. The IceAGE project and its predecessors,

[1] Areas at the bottom of oceans, lakes, or rivers, including sediments and organisms that live in them.

BIOFAR and BIOICE, provide invaluable data for studying the impacts of climate change and seabed mining, especially in the Greenland, Iceland, and Norwegian (GIN) seas [7].

Cumacea, a crustacean taxon within Peracarida, are major indicators of marine ecosystem health due to their sensitivity to environmental fluctuations [8] and their contribution to benthic food webs [9]. Despite their ecological importance, the evolutionary history of deep-sea benthic invertebrates remains uncharted, notably in the North Atlantic [10]. Analyzing the genetic and distribution patterns of these deep-sea organisms is crucial for predicting their responses to climate change [10] and anthropogenic pressures [7], while also advancing our understanding of their adaptive mechanisms within deep-sea ecosystems.

Given the urgency of the aforementioned factors, this study aims to analyze the effects of ecological (climatic and environmental) and spatial variables on the genetic variation and adaptation of Cumacea in the Northern North Atlantic. More specifically, we will examine whether there is a genetic adaptation between the genetic structure of a region represented by a partial sequence of the 16S rRNA mitochondrial gene of the Cumacea species included in our analyses and their habitat variables. If so, we will identify which variables show the greatest divergence from a specific segment (i.e., window) of this partial sequence and further investigate the potential associated protein using bioinformatics tools to interpret its biological relevance. Our approach includes confirming various phylogeographic models[2] and updating a Python package (currently in beta), *aPhyloGeo*, to facilitate these analyses.

This paper is organized as follows: Section 2 reviews pertinent studies on the biodiversity and biogeography of deep-sea benthic invertebrates; Section 3 summarizes the aims and contributions of this study, highlighting aspects relating to the conservation and adaptation of marine invertebrates to climate change; Section 4 describes data collection, preprocessing and phylogeographic analyses of partial genetic sequence and habitat variables; Section 4.7 describes the metrics used to evaluate the phylogeographic models; Section 5 presents the results; finally, Section 6 discusses their implications for future research and conservation efforts.

2. Related Works

Assessing and quantifying the biodiversity of deep-sea benthic invertebrates has become increasingly important since it was discovered that their species richness may be underestimated [11]. Subsequent research has highlighted the need for large-scale distribution models to interpret the diversity of these organisms across their ecological and evolutionary contexts [12]. Consequently, recent efforts have focused on mapping, managing, and studying the seabed [13]. Advanced technologies, such as acoustic detection are improving our knowledge of benthic ecosystem complexity [13]. Integrating genetic and habitat variables provides a better insight into how ecosystemic, meteorological, and spatial variables influence the genetic variation, distribution, biodiversity, and resilience of deep-sea benthic organisms [14].

However, the relationship between genetics and the environment is complex, involving gene-environment interactions and factors related to natural selection, which makes it difficult to identify clear causal relationships [15]. Additionally, the distinction between the direct and indirect effects of the environment on genetics presents further challenges [16], [17]. The limitations of current methods for measuring genetic and ecological variables, combined with logistical constraints, often limit the scope of such studies [16], [18]. This complexity may explain why the relationship between the environment and genetics of

[2] Phylogeographic models are computational tools that analyze relationships between the genetic structures of populations and their geographic distributions. In our case, by incorporating regional, biological, and atmospheric variables, we can analyze and interpret their impact on the genetic adaptation and spatial patterns of Cumacea species.

Cumacea has been less studied, despite their importance for understanding how deep-sea invertebrates adapt to fluctuating environmental conditions.

3. Our Contribution

Our study focuses on the genetic fluctuation of a partial sequence of the 16S rRNA mitochondrial gene in Cumacea communities in response to variations in their habitat, a topic that has been little explored in previous studies [11], [19]. We aim to refine the natural selection hypothesis by identifying specific divergent genetic regions and the potentially associated proteins using bioinformatics tools, such as protein structure modeling and functional annotation databases, to reveal the potential functions that these proteins may have in the adaptation of Cumacea to habitat fluctuations. By linking this partial sequence to habitat variables using robust analytical methods, such as dissimilarity calculations and phylogenetic reconstructions, we can better interpret the selection effects at the molecular level of this Cumacea sequence, which could confer survival advantages in the harsh environments of the Northern North Atlantic. This represents a major advance over previous research, which has often struggled to integrate genetic and biological data in the context of deep-sea invertebrates [14], [20] or has faced difficulties in linking genetics and environment [21], [22].

Furthermore, our genetic and environmental data highlight habitats of high conservation interest that can be considered for establishing marine protected areas [4]. These results are essential for developing informed conservation strategies in the context of climate change. Finally, our study paves the way for further research on other invertebrate species across different geographic regions. By extending this research to diverse environments and taxonomic groups, scientists will be better able to assess the adaptation and resilience of marine invertebrates to changing conditions.

4. Materials and Methods

This section describes our data and introduces the main data preprocessing steps, the *aPhyloGeo* software, the distance metrics used, and how the figures were created. A flow chart, constructed with the diagram software draw.io, summarizes this section (see Figure 1).

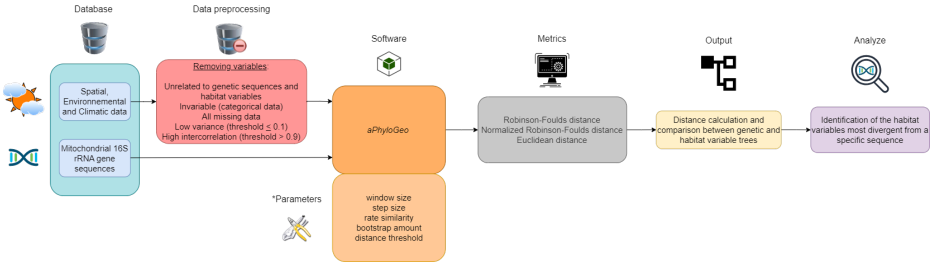
Database
Data preprocessing
Software
Metrics
Output
Analyze
Spatial, Environnemental and Climatic data
Mitochondrial 16S rRNA gene sequences
Removing variables:
Unrelated to genetic sequences and habitat variables
Invariable (categorical data)
All missing data
Low variance (threshold ≤ 0.1)
High intercorrelation (threshold > 0.9)
aPhyloGeo
*Parameters
window size
step size
rate similarity
bootstrap amount
distance threshold
Robinson-Foulds distance
Normalized Robinson-Foulds distance
Euclidean distance
Distance calculation and comparison between genetic and habitat variable trees
Identification of the habitat variables most divergent from a specific sequence

Figure 1. *Flow chart summarizing the Materials and Methods section workflow. Six different colors highlight the blocks. The first block (blue) represents our database. The second block (in red) is data preprocessing, which consists of deleting certain variables. The third and fourth blocks (orange) implement the* aPhyloGeo *software and its parameters for our phylogeographic analyses. The fifth block (gray) applies distance metrics to the genetic and habitat variable trees produced. The sixth block (yellow) calculates and compares distance metrics between genetic and habitat variable trees. The seventh block (purple) identifies the most divergent habitat variables of a specific region of the partial sequence of the 16S rRNA mitochondrial gene based on the results of tree comparisons. *See YAML files on GitHub for more details on these parameters.*

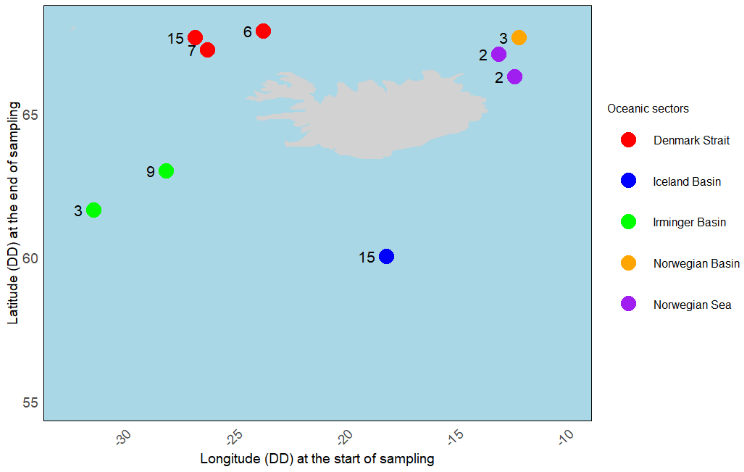

Figure 2. *Distribution map of Cumacea specimens included in our analyses according to the oceanic sector where they were sampled. The grey area represents Iceland, and the number next to the point is the number of specimens found at that sampling point.*

4.1. Description of the data

The study area is located in the Northern part of the North Atlantic, including the Denmark Strait, the Iceland Basin, the Irminger Basin, and the Norwegian Basin and Sea (see Figure 2). The specimens analyzed were collected as part of the IceAGE project (Icelandic marine Animals: Genetic and Ecology; Cruise ship M85/3 in 2011), which focused on the deep continental slopes and abyssal waters around Iceland [7]. The sampling period for the included specimens was from August 30 to September 22, 2011, and they were collected at depths ranging from 315.9 m to 2567.7 m. Detailed protocols concerning the sampling plan, sample processing, DNA extraction steps, PCR amplification, sequencing, and aligned DNA sequences are available in [3].

4.2. Data preprocessing

We used data from the article [3], the IceAGE project, and related data from the BOLDSystem database, as described in [3]. Given these databases' enormous variety of variables, we applied a selective reduction procedure. Variables with no variability (categorical data) were excluded from our study, for which all data were missing and were not linked to genetic sequences or spatial, environmental, and climatic variables. Out of the 495 available in the IceAGE dataset, we considered 62 specimens for which partial 16S rRNA mitochondrial gene sequences were available.

Next, we calculated the variance (S^2) using the $var()$ function in RStudio Desktop 4.3.2 for each of the selected variables (numerical and categorical). This step aimed to eliminate variables with low variation, as they are unlikely to provide essential data for analysis. We set a variance threshold of ≤ 0.1 to exclude uninformative variables. The latter retains variables whose variability is reasonably sufficient for our analyses while rejecting those with little variation. Only water salinity was eliminated based on this criterion ($S^2_{\text{Salinity}} = 0.02146629$). The formula (see Equation Equation 1) and code (Program 1) used to calculate the variance of our final variables, available in the data file on GitHub, are provided below:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - | \, x)^2}{n-1} \tag{1}$$

where S^2 is the variance of the variable, x_i represents each value of the variable, $| \, x$ is the average of the values for this variable, and n is the number of values for this variable in the dataset.

```r
# Import data from the CSV file
Data <- read.csv(file="Final_Data_Article.csv", header=TRUE, sep=";")

# Define a function to calculate entropy for categorical variables
calculate_entropy <- function(x) {
  # Calculate the frequency of each category
  freq_table <- table(x)

  # Calculate probabilities
  probabilities <- freq_table / sum(freq_table)

  # Calculate entropy using the probabilities
  entropy_value <- -sum(probabilities*log(probabilities), na.rm=TRUE)

  return(entropy_value)
}

# Calculate variance
variances <- sapply(Data, function(x) {
  # Check if the column is numeric
  if (is.numeric(x)) {
    # Compute variance, excluding NA values
    var(x, na.rm=TRUE)
  } else if (is.factor(x) || is.character(x)) {
    # If the column is categorical, compute entropy
    calculate_entropy(x)
  } else {
    NA  # Return NA for other types of columns
  }
})

# Display variances/entropies
print(variances)
```

Program 1. *RStudio script to calculate the variance of each numerical and categorical variables in our final dataset.*

We calculated the Pearson correlation (r) between variables using the $cor()$ function in RStudio Desktop 4.3.2. Variables (numerical) exhibiting strong correlations with each other (threshold > 0.90) were removed to avoid repetition and guarantee variable independence. We considered the threshold of > 0.90 to be an adequate compromise between preserving properties for our analyses and eliminating the repetition of information in our data. Since we have three missing data for O_2 concentration (mg/L), we have used the “pairwise.complete.obs method”. This method calculates the Pearson correlation matrix using all accessible pairs of observations, even if some data are missing. Using the above threshold, four variables were discarded: latitude (DD) at the start of sampling (Lat_start_end: $r = 0.9996658$), longitude (DD) at the end of sampling (Long_start_end: $r = 0.9999979$), depth (m) at the end of sampling (Depth_start_end: $r = 0.9998579$) and wind direction at the start of sampling (WindD_start_end: $r = 0.9752331$). The decision to remove these variables was based on their variance (S^2) value: $S^2_{\text{Lat\_start}} = 10.03077$ and $S^2_{\text{Lat\_end}} = 10.71335$; $S^2_{\text{Long\_start}} = 30.47940$ and $S^2_{\text{Long\_end}} = 30.47574$; $S^2_{\text{Depth\_start}} = 776437.1$ and $S^2_{\text{Depth\_end}} = 775394.7$; $S^2_{\text{WindD\_start}} = 2.405077$ and $S^2_{\text{WindD\_end}} = 4.482285$. The formula (see Equation Equation 2) and code (Program 2) used to calculate the Pearson correlation coefficient between our final numerical variables are shown below:

$$r = \frac{\sum_{i=1}^{n}(x_i - \mid x)(y_i - \mid y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mid x)^2 \sum_{i=1}^{n}(y_i - \mid y)^2}} \tag{2}$$

where r is the Pearson correlation coefficient between two variables, x_i are the values of the variable x, y_i are the values of the variable y, $\mid x$ and $\mid y$ are respectively the averages of the two variables, and n is the number of values of the two variables in the dataset.

This selection of variables and data resulted in a table containing 62 rows ($n = 62$) and 16 columns (number of variables).

```
# Import data
Data <- read.csv(file="Final_Data_Article.csv", header=TRUE, sep=";")

# Select numeric columns only from the dataset
numeric_Data <- Data[sapply(Data, is.numeric)]

# Calculate Pearson correlation matrix
correlation_matrix <- cor(numeric_Data, use="pairwise.complete.obs")

# Display correlation matrix
print(correlation_matrix)
```

Program 2. *RStudio script to calculate the Pearson correlation coefficient between all the numerical variables in our final dataset.*

4.3. *Selected variables in the IceAGE database*

4.3.1. *Spatial data:*

- The latitude at the end of sampling (see Figure 3a) and longitude at the start of sampling (see Figure 3b), both in decimal degrees (DD), as they are intimately linked to the environmental gradients and historical mechanisms modeling genetic heterogeneity [23].
- The five oceanic sectors across the seas around Iceland (see Figure 2): the Denmark Strait ($n = 28$), the Iceland Basin ($n = 15$), the Irminger Basin ($n = 12$), the Norwegian Sea ($n = 4$), and the Norwegian Basin ($n = 3$).

4.3.2. *Environmental data:*

- Depth (m) at the start of sampling (see Figure 3c), as well as water temperature (°C) (see Figure 3e), and O_2 concentration (mg/L) (see Figure 3f), as these are vital elements of the marine ecosystem that have an impact on the distribution and evolutionary acclimatization of marine species [24], [25].
- Since the sedimentary characteristics directly influence the distribution of Cumacea [3], they were included in our data. They are divided into six ecological niche categories: mud ($n = 30$), sandy mud ($n = 15$), sand ($n = 9$), forams ($n = 3$), muddy sand ($n = 3$), and gravel ($n = 2$).

4.3.3. *Climatic data:*

Wind speed (m/s) at the start (see Figure 3d) and end of sampling and wind direction at the end of sampling were also included, giving the contribution of wind to benthic ecosystem dynamics and the restructuring of species distribution by wind currents and sediment transport [26], [27], [28]. The wind direction at the end of sampling comprises eight orientations: south (S, $n = 15$), southwest (SW, $n = 15$), northeast (NE, $n = 9$), west-southwest (WSW, $n = 7$), southeast (SE, $n = 6$), north-northwest (NNW, $n = 5$), south-southeast (SSE, $n = 3$), and east (E, $n = 2$).

4.4. *Selected variables in the BOLDSystem database*

4.4.1. *Taxonomic data:*

The family, genus, and scientific name of the Cumacea were integrated into our data to study evolutionary relationships and genetic variation to habitat and acclimatization variables among the specimens. These comprise seven families (see Figure 4): Diastylidae ($n = 21$), Lampropidae ($n = 13$), Leuconidae ($n = 12$), Astacidae ($n = 7$), Bodotriidae ($n = 4$), Ceratocumatidae ($n = 3$), and Pseudocumatidae ($n = 2$). A total of 20 Cumacea species were included in our dataset (see Figure 4). We have also included the sample identity (Sampleid) so that each specimen remains unique. Some specimens were only identified to family ($n = 1$) or genus ($n = 4$).

4.5. Selected variables from article C. Uhlir et al. [3]

4.5.1. *Other environmental data:*

The habitat and water mass of the sampling points were the only environmental variables taken directly from Table 1 of [3], as they can provide insight into how they may affect Cumacea genetic diversity and the acclimatization of these species in the GIN seas around Iceland. Thus, the water masses definitions, as described in [3], were used as a reference: Arctic Polar Water (APW, $n = 15$), Iceland Sea Overflow Water (ISOW, $n = 15$), North Atlantic Water (NAW, $n = 9$), warm Norwegian Sea Deep Water (NSDWw, $n = 8$), Arctic Polar Water/ Norwegian Sea Arctic Intermediate Water (APW/NSAIW, $n = 7$), Labrador Sea Water (LSW, $n = 3$), cold Norwegian Sea Deep Water (NSDWc, $n = 3$), and Norwegian Sea Arctic Intermediate Water (NSAIW, $n = 2$) (see Figure 5). In terms of habitat, we considered the three categories used in [3]: Deep Sea ($n = 38$), Shelf ($n = 15$), and Slope ($n = 9$) (see Figure 6).

4.5.2. *Genetic data:*

The aligned partial DNA sequence of the 16S rRNA mitochondrial gene was included, as this region is standard in phylogeny and phylogeography studies [29] and sufficiently conserved over time to guarantee exact alignments between different species [30]. We examined 61 of the 306 aligned DNA sequences used for phylogeographic analyses by [3]. As some specimens have their DNA sequence duplicated, or even quadruplicated with a difference of one or two nucleotides, the longest-aligned DNA sequence of each specimen was retained. The “ICE1-Dia004” specimen is the only one whose sequence (not aligned) was taken from the BoldSystem database, as it was absent from the [3] aligned DNA database.

4.6. aPhyloGeo *software*

We used the cross-platform Python software *aPhyloGeo*, developed by the Tahiri Lab team, for our phylogeographic analyses. This software is designed to analyze phylogenetic trees using ecological and spatial variables (Program 3) to interpret the evolution of species under different environmental conditions [31], [32], [33].

This software was selected for our analysis as it is the first phylogeographic tool capable of establishing similarity or dissimilarity between the genetics of species and environmental, climatic, and spatial variables [31], [32], [33], which is precisely the objective of our study. The *aPhyloGeo* software offers several key functionalities:

1. **Phylogenetic tree evaluation**: The software identifies the evolutionary relationships among species based on their genetic sequences [31], [32], [33], which is essential for interpreting phylogeographic models that connect species evolution to their spatial distribution and biological and meteorological contexts.
2. **Ecological and regional dissimilarity analysis**: The software highlights divergence and convergence between genetic sequences and habitat variables [31], [32], [33], enabling the assessment of the influence of these variables on genetic fluctuations and the evolutionary history of Cumacea species.
3. **Evaluation of genetic diversity**: The software quantifies genetic heterogeneity, facilitating the identification of potential evolutionary processes (e.g., mutation, speciation, and genetic drift) and local adaptations.

The *aPhyloGeo* Python package is freely and publicly available on GitHub, and is also available on PyPi, to facilitate complex phylogeographic analyses. The software process has three main stages:

```
if __name__ == "__main__":

    # Load parameters
    Params.load_from_file(Params.reference_yaml_filepath)

    # Load the sequence file
    sequence_file = utils.loadSequenceFile(
                    Params.reference_gene_filepath)

    # Create an AlignSequences object
    align_sequence = AlignSequences(sequence_file)

    # Load variable data
    variable_data = pd.read_csv(Params.file_name)

    # Perform the alignment of sequences
    alignments = align_sequence.align()

    # Generate genetic trees based on aligned sequences
    geneticTrees = utils.geneticPipeline(alignments.msa)

    # Create a GeneticTrees object
    trees = GeneticTrees(trees_dict=geneticTrees,
                        format="newick")

    # Generate variable trees
    variableTrees = utils.climaticPipeline(variable_data)

    # Filter the results based on the generated trees
    utils.filterResults(variableTrees,
                       geneticTrees,
                       variable_data)
```

Program 3. *Main script for tutorial using the aPhyloGeo package.*

1. **The first step** was to collect DNA sequences from Cumacea of sufficient quality for the needs of our results [31], [32], [33]. In this study, 62 Cumacea specimens were selected to represent 62 partial sequences of the 16S rRNA mitochondrial gene. We then included, from our database, two climatic variables, namely wind speed (m/s) at the start and end of the sampling; three environmental variables, such as depth (m) at the start of sampling, water temperature (°C), and O_2 concentration (mg/L); and two geographic variables, latitude (DD) at the end of sampling and longitude (DD) at the start of sampling.
2. **The second step** was to generate trees from genetic, biological, spatial, and meteorological data. For spatial variables, the Neighbor-Joining method[3] was applied between each pair of Cumacea from distinct spatial conditions to produce a symmetrical square matrix and build the spatial tree from this matrix [33]. Each geographic variable generates a distinct phylogenetic tree. If there are m windows from the genetic sequences, there will be m geographic trees. The same approach was applied to biological, meteorological, and genetic data.

For the genetic data, phylogenetic reconstruction was repeated to build genetic trees based on 62 partial sequences of the 16S rRNA mitochondrial gene, considering only data within a window that progresses along the alignment [31], [32], [33]. Each window in the alignment will give a genetic tree. If there are n windows from the sequences, there will be n phylogenetic trees. This displacement can vary according to the steps and the size of the window defined by the user (their length is determined by the number of base pairs (bp)) [31], [32], [33].

In our case, we set up the *aPhyloGeo* software as follows: $pairwiseAligner$ for sequence alignment; $\mathrm{Hamming\ distance}$ to measure simple dissimilarities between sequences; $\mathrm{Wider\ Fit\ by\ elongating\ with\ Gap(starAlignment)}$ algorithm takes alignment gaps into ac-

[3] It is a method used to construct phylogenetic trees using distance matrices.

count, which is often mandatory in the case of major deletions or insertions in the sequences; windows_size: 10 nucleotide (nt); and finally, step_size: 1 nt. The last two configurations imply that for each 10 nt window, a phylogenetic tree is produced using the 10 nt sequence of each Cumacea. Next, the window is moved by 1 nt, creating a new tree with the next 10 nt, and so on until the end of the alignment. Genetic trees will be stored in an object called T_1, while spatial and ecological trees will be stored in another object called T_2.

1. **The third step** is to compare the genetic trees constructed in each sliding window with the ecosystemic, atmospheric, and regional trees using the Robinson-Foulds distance [34], normalized Robinson-Foulds distance and Euclidean distance. These contribute to understanding the correspondence between Cumacea genetic sequences and their habitat variables. The approach also takes bootstrapping into account [31], [32], [33]. The results of these metrics were obtained using the functions $robinson_foulds(tree1, tree2)$ and $euclidean_dist(tree1, tree2)$ from the ***aPhyloGeo*** software and were organized by the main function (Program 3). Those for the normalized Robinson-Foulds distance were obtained with the function $robinson_foulds(tree1, tree2)$ (see the last line of code in Program 4). The result of the metrics indicates which variables show the greatest genetic divergence according to the magnitude of the metric distances (see figures Figure 7 and Figure 8).

A sliding-window approach enables the precise location of subtle sequences with high rates of genetic divergence [31], [32], [33]. This method involves moving a fixed-size window over the alignment of genetic sequences. This allows genetic trees to be built for each part of the aligned sequences, depending on the size of the window and the step size. It therefore makes it possible to recognize changes in evolutionary relationships along the partial sequence region of the 16S rRNA mitochondrial gene of Cumacea species. This method is essential to determine whether this region of the Cumacea genome can be affected by certain ecological or spatial variables in their habitat (see Figure 7 and Figure 8).

4.7. Metrics

In our phylogeographic analysis, we employed three distance metrics to quantify differences between phylogenetic trees and habitat trees, as well as to evaluate dissimilarities between genetic sequences and the associated environmental variables. This approach allowed for a detailed examination of the evolutionary patterns of Cumacea communities across varying ecological conditions.

The following section provides a detailed description of the three distance functions referenced in the second and third steps of Section 4.6, offering a more rigorous examination of their role in the analysis.

4.7.1. *Robinson-Foulds distance:*

The Robinson-Foulds (RF) distance [34] calculates the distance between genetic trees built in each sliding window (T_1) and the variable trees (T_2) (see the list in the first step of the Section 4.6) [35], [36]. This metric is used to evaluate the topological differences between the two sets of trees by measuring the minimum number of elementary operations (merging and splitting nodes) required to transform one tree (genetic) into another (variable habitat) (see Equation Equation 3 and Program 4). A high distance of a specific window in RF distance analysis may imply that the habitat variable has little to no impact on the evolution of this particular DNA sequence and that the fluctuation of this variable might not explain the genetic divergence observed.

$$\mathrm{RF}(T_1, T_2) = \mid \Sigma(T_1) \Delta \Sigma(T_2) \mid \tag{3}$$

where $\mathrm{RF}(T_1, T_2)$ is the Robinson-Foulds distance between the two sets of trees, $\Sigma(T_1)$ and $\Sigma(T_2)$ are the sets of divisions in trees T_1 and T_2 and Δ, the difference between these two sets.

```python
def robinson_foulds(tree1, tree2):

    # Initialize the Robinson-Foulds distance
    rf = 0

    # Convert trees from Newick format to ete3.Tree objects
    tree1_newick = ete3.Tree(tree1.format("newick"), format=1)
    tree2_newick = ete3.Tree(tree2.format("newick"), format=1)

    # Calculate the Robinson-Foulds distance
    rf, rf_max, common_leaves = tree1_newick.robinson_foulds(
                                tree2_newick,
                                unrooted_trees=True)

    # If there are no common leaves, set the RF distance to 0
    if len(common_leaves) == 0:
        rf = 0

    # Return the RF distance and its normalized value
    return rf, rf / rf_max
```

Program 4. *Python script for calculating the Robinson-Foulds Distance using the ete3 package in the aPhyloGeo package. The Newick format represents the phylogenetic and variable trees in text form.*

4.7.2. *Normalized Robinson-Foulds distance:*

The normalized Robinson-Foulds (nRF) distance scales the RF distance to account for the size variations in the trees (number of clades; i.e., a group of species with a common origin), allowing a more equitable comparison. It scales the distance to a range between 0 and 1. In our context, the distance has been normalized by $2n - 6$, where n represents the number of taxa (see Equation Equation 4 and the last line of code in Program 4).

Since the size of environmental trees constructed with O_2 concentration data (mg/L) differs from that of other variables due to missing data, this nRF distance allows its dissimilarity with genetic trees to be compared more fairly [33], [36]. It reveals the relative influence of O_2 concentration (mg/L) on Cumacea phylogenetic relationships, independent of tree size [33], [36]. A high distance of a specific window in the nRF distance analysis suggests that we cannot conclude that there is a correlation between this DNA sequence and the variable. It may indicate a topological dissimilarity between the habitat variable trees and the genetic trees at that position in the DNA sequence alignments.

$$\mathrm{RF}_{\mathrm{norm}}(T_1, T_2) = \frac{\mid \Sigma(T_1)\Delta\Sigma(T_2) \mid}{\mid \Sigma(T_1) \mid + \mid \Sigma(T_2) \mid} \tag{4}$$

where $\mathrm{RF}_{\mathrm{norm}}(T_1, T_2)$ is the normalized Robinson-Foulds distance between the two sets of trees, $\Sigma(T_1)$ and $\Sigma(T_2)$ are the sets of divisions in trees T_1 and T_2 and Δ, the difference between these two sets.

4.7.3. *Euclidean distance:*

The Euclidean distance calculates the straight-line distance between two sets of points in a multidimensional space, which designates the length divisions of the two sets of trees (T_1 and T_2). It is used to evaluate the degree of divergence or similarity of topologies between two respective sets of trees (see Equation Equation 5 and Program 5). A high distance of a specific window in the Euclidean distance analysis suggests evolutionary divergences between members of the Cumacea communities at the level of this DNA sequence and the variation of the habitat variable (see Figure 7d and Figure 8d). In other words, the habitat variable may not have a dominant contribution to the evolution of this specific sequence of Cumacea communities.

```python
def euclidean_dist(tree1, tree2):

    # Initialize the Euclidean distance
    ed = 0

    # Create a TaxonNamespace object to handle taxon information
    tns = dendropy.TaxonNamespace()

    # Load the first tree into a dendropy Tree object
    tree1_tc = dendropy.Tree.get(data=tree1.format("newick"),
                                 schema="newick",
                                 taxon_namespace=tns)

    # Load the second tree into a dendropy Tree object
    tree2_tc = dendropy.Tree.get(data=tree2.format("newick"),
                                 schema="newick",
                                 taxon_namespace=tns)

    # Encode the bipartitions of both trees
    tree1_tc.encode_bipartitions()
    tree2_tc.encode_bipartitions()

    # Calculate the Euclidean distance
    ed = dendropy.calculate.treecompare.euclidean_distance(
                                                tree1_tc,
                                                tree2_tc)

    return ed
```

Program 5. *Python script for calculating the Euclidean distance using the ete3 and the dendropy packages in the aPhyloGeo package. The Newick format represents the phylogenetic and variable trees in text form.*

$$d_{\text{Euclidean}}(T_1, T_2) = \sqrt{\sum_{i=1}^{n} (T1_i - T2_i)^2} \tag{5}$$

where $d_{\text{Euclidean}}(T_1, T_2)$ is the Euclidean distance between the two sets of trees, and $T1_i$ and $T2_i$ represent the respective divisions of trees T_1 and T_2 for each i-th division.

Interestingly, Euclidean distance is more sensitive to the subtle tree topology, making it suitable for identifying detailed correlations between genetic fluctuations and those of habitat variables [37]. It can therefore be used to study fine divergences between trees, enabling nuanced identification of the effects of habitat variables on the genetic structure of species [37]. As for the Robinson-Foulds distance (normalized or not), although widely applied in evolutionary biology, it is less sensitive to slight topological dissimilarities, making it less accurate for identifying fine correlations between genetics and habitat variables due to its structural nature [38], [39].

4.8. *Creating Figures*

Figure 3, Figure 4, Figure 7 and Figure 8 were made with Python 3.11, while Figure 2, Figure 5 and Figure 6 were made with RStudio Desktop 4.3.2.

5. Results

The violin diagrams shown in Figure 3 are used to display summary statistics similar to box plots, showing medians (white lines), interquartile ranges (thickened black bars), and the rest of the distributions (thin black lines), except the “extreme” points. Wider areas indicate a greater probability of the variables taking a given value. They summarize the distribution of spatial (latitude at the end of sampling and longitude at the start of sampling, both in DD), atmospheric (wind speed (m/s) at the start of sampling), and ecosystemic (depth (m) at the start of sampling, water temperature (°C), and O_2 concentration (mg/L)) data. These

Table 1. *Table summarizing key statistics such as mean, median, standard deviation (Std Dev), 1st quartile (Q1) and 3rd quartile (Q3) of biological (depth (m) at the start of sampling, water temperature (°C), and O_2 concentration (mg/L)), spatial (latitude (DD) at the end of sampling and longitude (DD) at the start of sampling) and atmospheric (wind speed (m/s) at the start of sampling) variables for our phylogeographic analyses.*

Attributes	Statistics	Mean	Std Dev	Q1	Median	Q3	Min	Max
Lat end (DD)		64.75	3.27	61.65	67.15	67.63	60.05	67.86
Long start (DD)		-23.12	5.52	-26.77	-26.21	-18.14	-31.35	-12.16
Depth start (m)		1412.57	881.16	579.10	1574.70	2504.70	315.90	2567.70
Watertemp ground (°C)		1.45	1.73	0.07	0.71	2.65	0.85	4.28
O2 saturation ground (mg/L)		271.88	18.11	258.39	278.77	290.90	245.53	292.97
Windspeed start (m/s)		6.26	2.16	5.25	6.00	7.00	2.00	11.00

diagrams are essential for understanding habitat conditions and highlighting the variables that can potentially influence genetic fluctuation and adaptability in Cumacea. In Table 1 and Figure 3, the variables are designated by their names from the IceAGE database, except for latitude (DD) at the end of sampling and longitude (DD) at the start of sampling, for which the term “dec” has been removed at the end to avoid confusion.

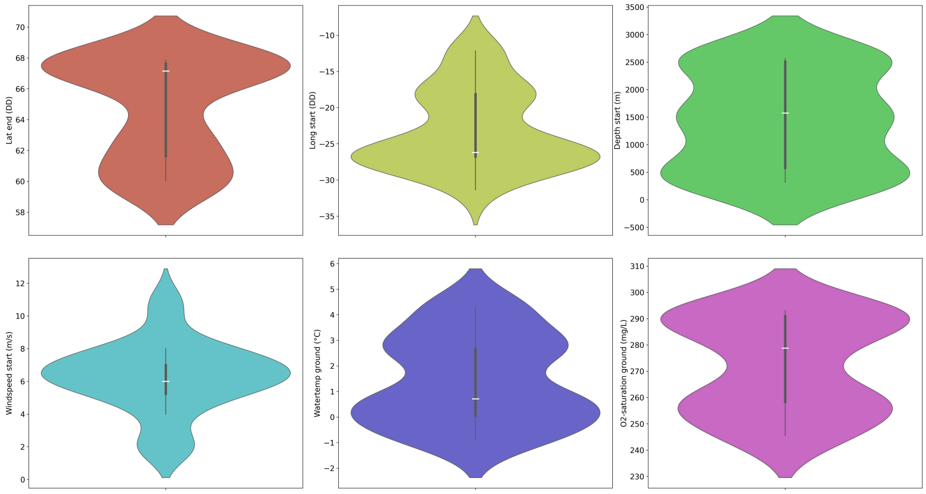

Figure 3. *Violin diagrams of two regional, one atmospheric, and three ecosystemic variables that provide essential information about the ecological and meteorological conditions of Cumacea habitats. a) Latitude (DD) at the end of sampling (red) suggest that the specimens come from two dominant latitudinal (DD) regions (around 61.65 DD and 67.63 DD); b) Longitude (DD) at the start of sampling (yellow) implies that the specimens come from two dominant longitudinal (DD) regions (around -26.77 DD and -18.14 DD); c) Depth (m) at the start of sampling (green) suggest that the specimens were mainly collected and concentrated at three different depths (m) (around 500 m, 1500 m and 2500 m); d) Wind speed (m/s) at start of sampling (light blue) indicate stable the wind conditions (m/s) at the start of sampling (around 6.00 m/s); e) Water temperature (°C) (dark blue) suggest that the specimens were mostly collected and concentrated at two different water temperatures (°C) (around 0.07 °C and 2.66 °C); f) O_2 concentration (mg/L) (pink) implies that the specimens were primarily collected and concentrated at two different O_2 concentrations (mg/L) (around 258.39 mg/L and 290.90 mg/L).*

Our results revealed variability in most habitat variables, as shown in Figure 3. For instance, the median of the latitude at the end of sampling (67.15 DD; Table 1) is higher than the mean (64.75 DD; Table 1), showing an asymmetric distribution skewed towards lower values. This trend is also observed for depth (m) at the start of sampling (Median: 1574.70 m; Mean: 1412.57 m; see Figure 3c and Table 1) and O_2 concentration (mg/L) (Median: 278.77 mg/L; Mean: 271.88 mg/L; see Figure 3f and Table 1). The bimodal shape of the latitude distribution curve suggests that the specimens came from two dominant latitudinal regions at the end of sampling (around 61.65 DD and 67.63 DD; see Figure 3a and Table 1). This bimodality is also observed in longitude (DD) at the start of sampling (around -26.77 DD and -18.14 DD; see Figure 3b and Table 1), as well as for water temperature (°C) (around 0.07 °C and 2.66 °C; see Figure 3e and Table 1), and O_2 concentration (mg/L) (around 258.39 mg/L and 290.90 mg/L; see Figure 3f and Table 1).

The median of the longitude (DD) at the start of sampling (-26.21 DD; Table 1) is lower than the mean (-23.12 DD; Table 1), indicating asymmetry on the higher sides (see Figure 3b), as does the water temperature (°C) (Mean: 1.45 °C; Median: 0.71 °C; see Figure 3e and Table 1). Unlike all the other diagrams in Figure 3, the curve of the depth (m) at the start of sampling (see Figure 3c) has a multimodal shape with three prominent peaks, suggesting that the specimens were mainly collected and concentrated at three different depths (around 500 m, 1500 m and 2500 m; see Figure 3c).

The mean (6.26 m/s; Table 1) and median of wind speed (m/s) at the start of sampling are fairly similar, with a high density of data around the median (6.00 m/s; see Figure 3d and Table 1). This suggests stable wind conditions (m/s) at the start of sampling. The key statistics and the figure for the wind speed (m/s) at the end of sampling are available in the img file on GitHub. The standard deviation of water temperature (°C) is relatively high (1.73 °C; Table 1) compared to the mean (1.45 °C; Table 1), suggesting acclimatization of Cumacea to a variety of habitat temperatures (-0.85 °C – 4.28 °C; see Figure 3e and Table 1). The range of data for O_2 concentration (mg/L) shows some variability (245.53 mg/L – 292.97 mg/L; see Figure 3f and Table 1) in the environmental conditions. This reflects a diversity of requirements in terms of O_2 concentration (mg/L), with Cumacea potentially affected by the heterogeneity of biogeochemical cycles, such as photosynthesis, respiration, and organic decomposition, which affect depth-dependent dissolved O_2 concentration (mg/L).

The distribution and diversity of the various Cumacea species and family found are shown in Figure 4. It shows that the most represented species are *Leptostylis ampullacea* (14.1%) and *Leucon pallidus* (12.5%). In contrast, species like *Bathycuma brevirostre* and *Styloptocuma gracillimum* are less represented (1.6%), implying that some species may have restricted ecological niches or face ecological forces that limit their distribution. The dominance of certain species (such as *Leptostylis ampullacea* and *Leucon pallidus*) suggests that they may have adaptive traits that enable them to make the most of the accessible resources, resist interspecific competition, or survive in fluctuating ecosystemic conditions, aligns with our study's aim of relating genetic adaptation to habitat characteristics.

The figure above supports the objective of our study by showing the distribution of the different Cumacea families in the various water masses (see Figure 5). The Diastylidae family, for example, is the most common in all water masses (turquoise color in Figure 5), testifying to its resilience and ecological adaptability to a wide variety of habitat conditions, reminiscent of the dominance of *Leptostylis ampullacea* which belongs to the Diastylidae family (see Figure 4, 14.1%).

The distribution of the different Cumacea families according to the type of habitat where they were collected during sampling is shown in Figure 6. The deep-sea habitats show the greatest diversity of families, mainly Diastylidae and Lampropidae, suggesting they are well acclimatized to deep-sea conditions. In contrast, the slope has the lowest diversity,

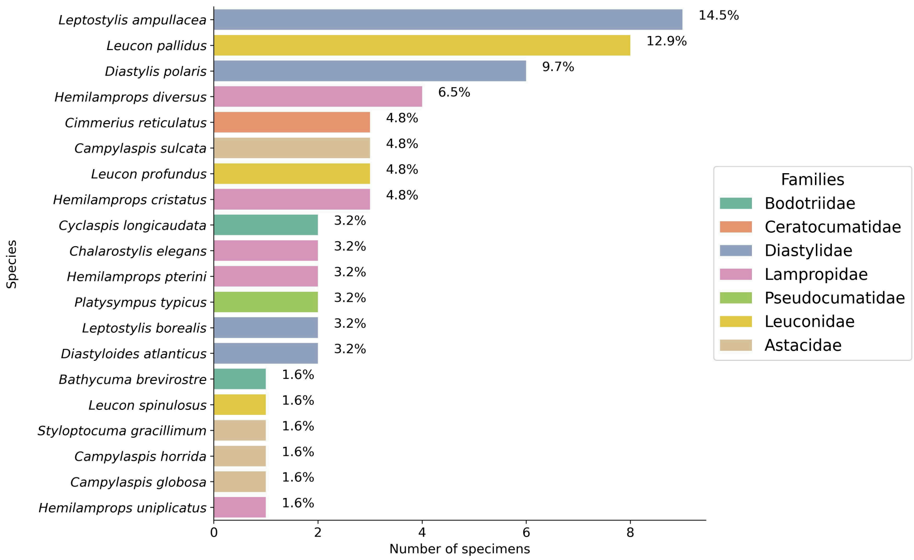

Figure 4. *Cumacea frequency distribution by species and family. The percentages (%) displayed above the bars indicate the relative abundance of each species in our dataset. Unlike less common species, those that are abundant (such as* Leptostylis ampullacea *and* Leucon pallidus*) may have adaptive characteristics that enable them to exploit resources more easily, resist interspecific competition or withstand changing biological conditions.*

with Diastylidae again the most dominant, implying that some Cumacea species have fewer ecological niches or are less adapted to this habitat. Although less diverse than the deep sea, the shelf is dominated by Leuconidae, indicating that this family may be specifically well-acclimated to this habitat. These patterns imply that certain Cumacea families, such as the Diastylidae, Lampropidae, Leuconidae, Pseudocumatidae, and Astacidae, have developed distinct adaptations (physiological, behavioral, or morphological) to remain in particular ecological niches, reflecting the impact of habitat conditions on the genetic distribution of Cumacea.

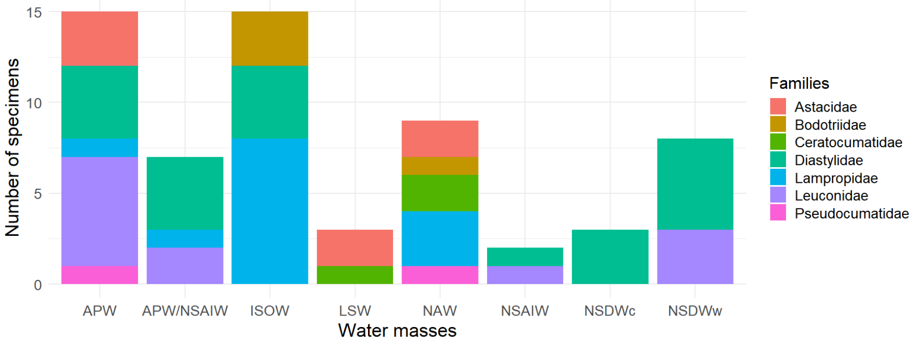

Figure 5. *Distribution of Cumacea families by water mass. This histogram represents the frequency of occurrence of the different Cumacea families, classified according to the water mass in which they were collected. Eight water mass categories are represented: Arctic Polar Water (APW), Arctic Polar Water/ North Sub-Arctic Intermediate Water (APW/NSAIW), Iceland Scotland Overflow Water (ISOW), Labrador Sea Water (LSW), North Atlantic Water (NAW), North Sub-Arctic Intermediate Water (NSAIW), cold North Sub-Atlantic Deep Water (NSDWc), and warm North Sub-Atlantic Deep Water (NSDWw). The presence of the Diastylidae (turquoise) family in the majority of water bodies (APW, APW/NSAIW, ISOW, NSAIW, NSDWc, and NSDWw) accentuates the resilience and ecological acclimatization of this family to various ecological niches and conditions.*

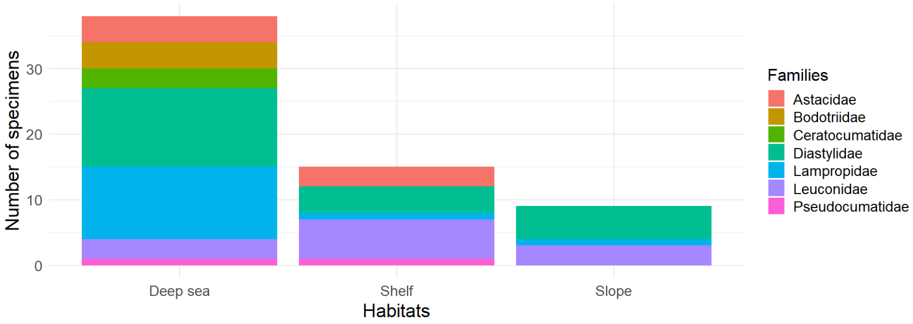

Figure 6. *Distribution of Cumacea families by habitat. This histogram represents the frequency of occurrence of the different Cumacea families, classified according to the habitat in which they were collected. Three habitat categories are represented: Deep Sea, Shelf, and Slope. The presence of Cumacea families in more than one habitat, such as Diastylidae (turquoise), Lampropidae (blue), Leuconidae (purple), Pseudocumatidae (pink), and Astacidae (red), may indicate the development of adaptations, whether morphological, physiological or behavioral, that could favor their persistence in these habitats.*

The divergence between specific genetic sequences and two variables, one climatic (wind speed (m/s) at the start of sampling) and the other environmental (O_2 concentration (mg/L)) is presented in Figure 7 and Figure 8. All the variables given in the first step of the Section 4.6 were analyzed and the configuration parameters are available in the $scripts$ Python file on GitHub. However, only these two show the most interesting rate of divergence. Using the three metrics mentioned in the Section 4.7, we noticed that the Euclidean distance is particularly sensitive to our data, manifesting considerable sequence variation at the position in MSA 560-569 amino acids (aa) (Euclidean distance: 0.85; see Figure 7d) and 1210-1219 aa (Euclidean distance: 1.23; see Figure 8d). The fluctuations in wind speed (m/s) at the start of sampling and in O_2 concentration (mg/L) do not appear to explain the variations in these two specific windows. This could indicate the absence of directional selection in these sequences due to these habitat variables, local selective pressures not considered in our analysis, or other evolutionary factors (e.g., genetic drift or biotic interactions) predominate over these two variables concerning these two sequences. On the other hand, this may suggest that these two variables could potentially influence the divergent (i.e., genetic diversification) rather than a convergent adaptation of these Cumacea, reflecting unique evolutionary responses to these specific ecological pressures. These results are consistent with the aim of our study, which is to identify the Cumacea genetic region that diverges most as a function of habitat variables, to determine whether this is due to divergent local adaptation or other evolutionary processes.

These results provide important insight into the genetic adaptation of Cumacea to their environment. These results need to be analyzed in greater depth to certify their involvement, especially in contrast with [3], which investigated similar topics of environmental and climatic effects on Cumacea distribution and genetics. The *aPhyloGeo* package is still in the process of being updated.

6. Conclusion

This study examines the effects of meteorological, regional, and ecosystemic variables on the genetics of Cumacea in the waters surrounding Iceland. Our main objective is to determine whether there is a discrepancy between the genetic informations of the partial 16S rRNA mitochondrial gene sequence (i.e. a window) of Cumacea species and their habitat variables. In addition to data distribution representations (see Figure 3, Figure 4, Figure 5 and Figure 6), DNA sequence analyses, using the *aPhyloGeo* software, have identified specific genetic windows that diverge from atmospheric and biological variables such as

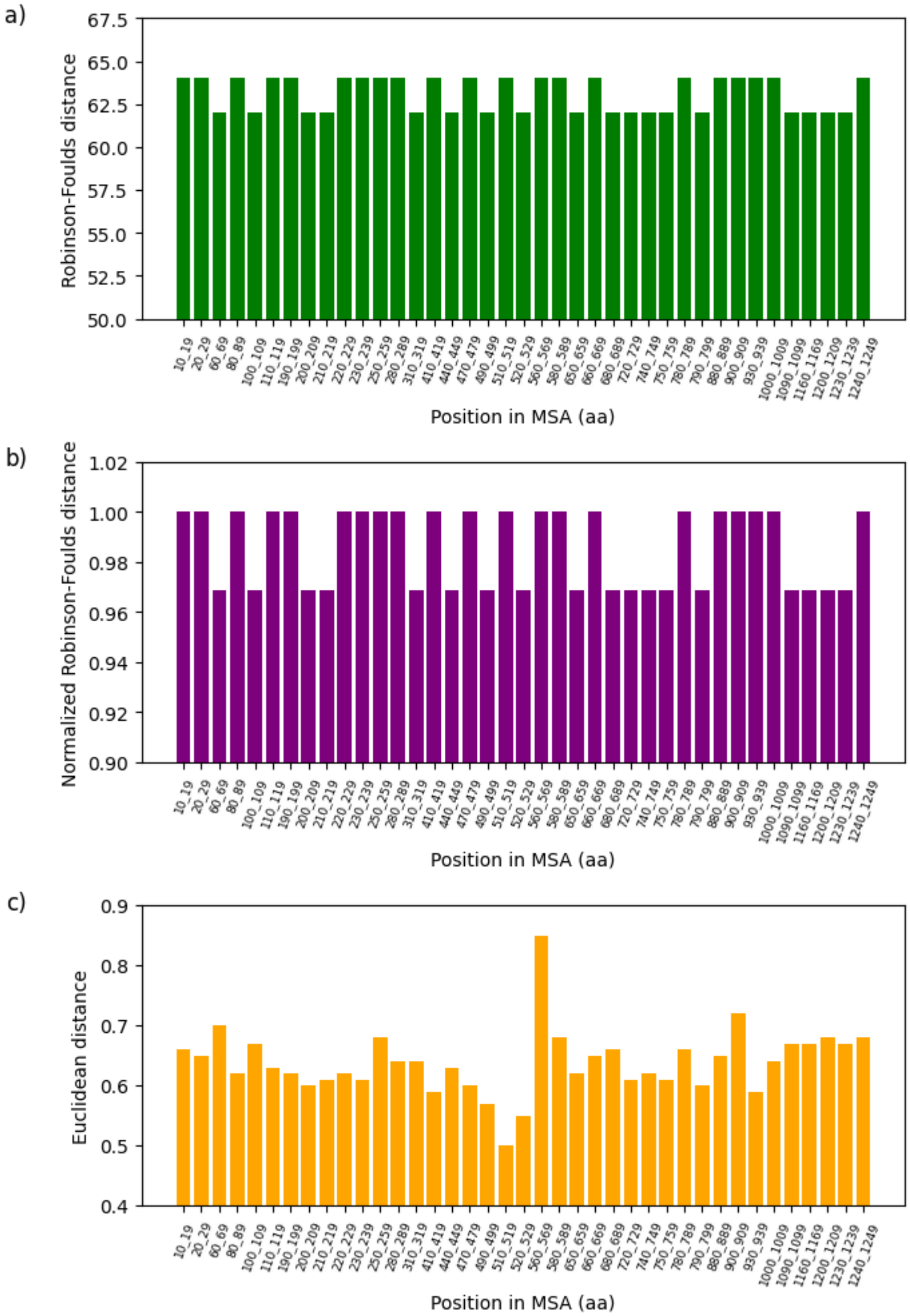

Figure 7. *Analysis of fluctuations in three distance metrics using multiple sequence alignment (MSA): a) Robinson-Foulds distance, b) normalized Robinson-Foulds distance, and c) Euclidean distance. Distance variations are studied to establish the potential dissimilarity between the partial sequence of the 16S rRNA mitochondrial gene of 62 Cumacea specimens and the variability of wind speed (m/s) at the start of sampling.*

wind speed (m/s) at the start of sampling (Position in MSA: 560-569 aa; Euclidean distance: 0.85; see Figure 7d) and O_2 concentration (mg/L) (Position in MSA: 1210-1219 aa; Euclidean distance: 1.23; see Figure 8d). These results could mean that these specimens have been shaped by these unique local environments, resulting in genetic sequences adapted to their particular conditions.

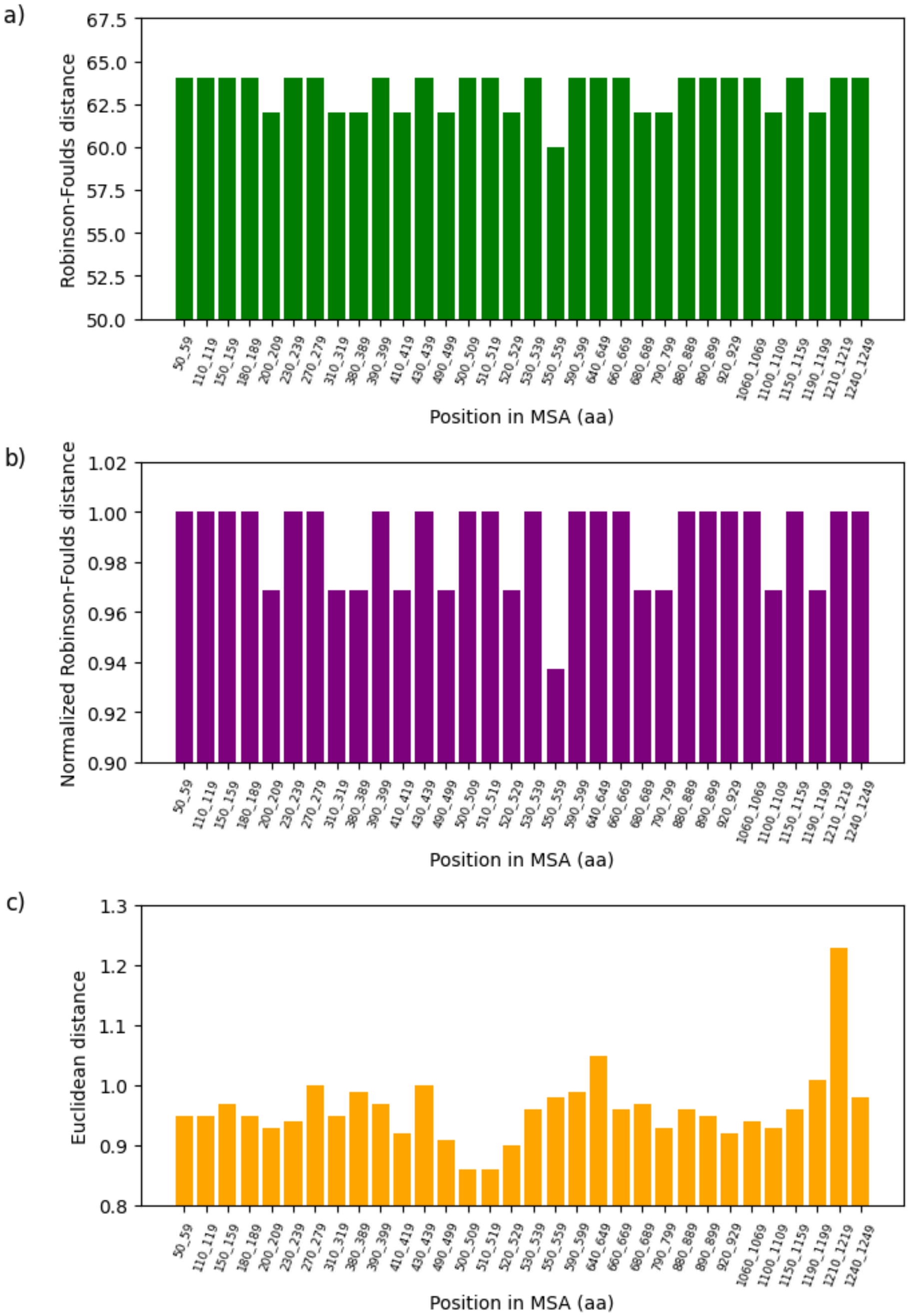

Figure 8. *Analysis of fluctuations in three distance metrics using multiple sequence alignment (MSA): a) Robinson-Foulds distance, b) normalized Robinson-Foulds distance, and c) Euclidean distance. These distances aim to determine the degree of dissimilarity between the partial sequence of the 16S rRNA mitochondrial gene of 62 Cumacea specimens and the variation in O_2 concentration (mg/L) at the sampling sites.*

The novelty in our research lies in the exhaustive divergence between habitat variables and genetic divegence in Cumacea, particularly in identifying genetic windows that diverge from habitat fluctuations, which has not been widely investigated in previous studies [14], [21]. Our integrated method identifies specific genetic regions sensitive to ecosystemic and atmospheric variations. Thus, the eventual identification of proteins linked to one of these variable DNA sequences will make it possible to represent its functional effects in responses to habitat changes. Our future research will focus on verifying the prediction of this protein and assessing its role in the physiological adaptation of Cumacea to fluctuating conditions, adding a link between genetic data and ecological function.

Interpreting how marine invertebrates genetically adapt to variations in their habitat can help predict their response to climate change and advance conservation plans to protect them. Identifying the variables that influence genetic variability in Cumacea can contribute to the designation and supervision of marine protected areas, assuring they include habitats crucial to the survival and acclimatization of these species. Thus, our results can inform the management of fishing and seabed mining companies by revealing ecologically vulnerable areas where these disturbances can seriously affect benthic biodiversity.

Furthermore, our results provide essential knowledge to guide future studies on the genetic adaptation of Cumacea and other invertebrates to ecological and regional variability. Based on these findings, future research should focus on additional ecosystemic and meteorological variables, such as nutrient accessibility, water pH, ocean currents, and the degree of human disturbance, to further improve the interpretation of the complex interactions between genetics and the environment. Extending the scope of application to other marine species, not just marine invertebrates, and various spatial regions would provide a better means of generalizing the results. With this in mind, longitudinal study models on these different species could reflect long-term climatic and biological fluctuations, and improve knowledge of the dynamics of genetic acclimatization.

However, it is important to recognize the limitations of our study. In particular, the three missing data points on O_2 concentration (mg/L) and the relatively small sample size ($n = 62$) may have induced a bias, which could impact the validity of our interpretations and restrict the generalizability of our results. Moreover, these missing data could provide partial insight into the relationship between O_2 concentration (mg/L) and genetic fluctuation in Cumacea, and our sample size may reduce the statistical power of our results. Future studies should address these gaps by incorporating larger sample sizes and more complete datasets to confirm and expand our conclusions. Additionally, as our research focuses solely on the partial sequence of the mitochondrial 16S rRNA gene, utilizing more elaborate genomic methods, such as whole-gene or even whole-genome sequencing, could help better understand marine species' genetic variety and global acclimatization mechanisms. This would provide more comprehensive genetic databases to improve accuracy and knowledge in identifying existing (and new) marine invertebrate species using DNA barcoding (e.g., mitochondrial DNA cytochrome c oxidase I (COX1)). Finally, multidisciplinary collaborations between ecology, genetics, and oceanography would be essential to enhance knowledge sharing and its application in future research.

Acknowledgments

The authors thank the SciPy conference and reviewers for their valuable comments on this paper as well as Mansour Kebe for his technical support and Carolin Uhlir for her clarifications and advice on her study [3]. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Fonds de recherche du Québec - Nature et technologies (FRQNT), the Université de Sherbrooke grant, and the Centre de recherche en écologie de l'Université de Sherbrooke (CREUS).

References

[1] S. Schnurr, A. Brandt, S. Brix, D. Fiorentino, M. Malyutina, and J. Svavarsson, "Composition and distribution of selected munnopsid genera (Crustacea, Isopoda, Asellota) in Icelandic waters," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 84, pp. 142–155, 2014, doi: 10.1016/j.dsr.2013.11.004.

[2] K. Meißner, N. Brenke, and J. Svavarsson, “Benthic habitats around Iceland investigated during the IceAGE expeditions,” *Polish Polar Research*, vol. 35, no. 2, pp. 177–202, 2014, doi: 10.2478/popore-2014-0016.

[3] C. Uhlir *et al.*, “Adding pieces to the puzzle: insights into diversity and distribution patterns of Cumacea (Crustacea: Peracarida) from the deep North Atlantic to the Arctic Ocean,” *PeerJ*, vol. 9, p. e12379, 2021, doi: 10.7717/peerj.12379.

[4] L. A. Levin and P. K. Dayton, “Ecological theory and continental margins: where shallow meets deep,” *Trends in ecology & evolution*, vol. 24, no. 11, pp. 606–617, 2009, doi: 10.1016/j.tree.2009.04.012.

[5] A. D. Rogers, A. Baco, H. Griffiths, T. Hart, and J. M. Hall-Spencer, “Corals on seamounts,” *Seamounts: ecology, fisheries & conservation*, pp. 141–169, 2007, doi: 10.1002/9780470691953.ch8.

[6] R. Danovaro *et al.*, “Exponential decline of deep-sea ecosystem functioning linked to benthic biodiversity loss,” *Current Biology*, vol. 18, no. 1, pp. 1–8, 2008, doi: 10.1016/j.cub.2007.11.056.

[7] K. Meißner, S. Brix, K. M. Halanych, and A. M. Jażdżewska, “Preface—biodiversity of Icelandic waters,” *Marine Biodiversity*, vol. 48, no. 2, pp. 715–718, 2018, doi: 10.1007/s12526-018-0884-7.

[8] B. Stransky and J. Svavarsson, “Diversity and species composition of peracarids (Crustacea: Malacostraca) on the South Greenland shelf: spatial and temporal variation,” *Polar Biology*, vol. 33, no. 2, pp. 125–139, 2010, doi: 10.1007/s00300-009-0691-5.

[9] P. Rehm, “Cumacea (Crustacea; Peracarida) of the Antarctic shelf-diversity, biogeography, and phylogeny= Cumacea (Crustacea; Peracarida) des antarktischen Schelfs-Diversität, Biogeographie und Phylogenie,” *Berichte zur Polar-und Meeresforschung (Reports on Polar and Marine Research)*, vol. 602, 2009, doi: 10.2312/BzPM_0602_2009.

[10] R. M. Jennings and R. J. Etter, “Phylogeographic Estimates of Colonization of The Deep Atlantic by The Protobranch Bivalve Nucula Atacellana,” *Polish Polar Research*, no. 2, pp. 261–278, 2014, doi: 10.2478/popore-2014-0017.

[11] J. F. Grassle and N. J. Maciolek, “Deep-sea species richness: regional and local diversity estimates from quantitative bottom samples,” *The American Naturalist*, vol. 139, no. 2, pp. 313–341, 1992, doi: 10.1086/285329.

[12] M. A. Rex, R. J. Etter, and C. T. Stuart, “Large-scale patterns of species diversity in the deep-sea benthos,” *Marine biodiversity: patterns and processes*, pp. 94–121, 1997, doi: 10.1017/CBO9780511752360.006.

[13] C. J. Brown, S. J. Smith, P. Lawton, and J. T. Anderson, “Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques,” *Estuarine, Coastal and Shelf Science*, vol. 92, no. 3, pp. 502–520, 2011, doi: 10.1016/j.ecss.2011.02.007.

[14] R. C. Vrijenhoek, “Cryptic species, phenotypic plasticity, and complex life histories: assessing deep-sea faunal diversity with molecular markers,” *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 56, no. 19–20, pp. 1713–1723, 2009, doi: 10.1016/j.dsr2.2009.05.016.

[15] N. Balkenhol *et al.*, “Identifying future research needs in landscape genetics: where to from here?,” *Landscape Ecology*, vol. 24, pp. 455–463, 2009, doi: 10.1007/s10980-009-9334-z.

[16] S. Manel *et al.*, “Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field,” *Molecular Ecology*, vol. 19, no. 17, pp. 3760–3772, 2010, doi: 10.1111/j.1365-294X.2010.04717.x.

[17] N. Balkenhol *et al.*, “Landscape genomics: understanding relationships between environmental heterogeneity and genomic characteristics of populations,” *Population genomics: Concepts, approaches and applications*, pp. 261–322, 2019, doi: 10.1007/13836_2017_2.

[18] A. B. Shafer and J. B. Wolf, “Widespread evidence for incipient ecological speciation: a metaanalysis of isolation-byecology,” *Ecology letters*, vol. 16, no. 7, pp. 940–950, 2013, doi: 10.1111/ele.12120.

[19] M. A. Rex, C. T. Stuart, and G. Coyne, “Latitudinal gradients of species richness in the deep-sea benthos of the North Atlantic,” *Proceedings of the National Academy of Sciences*, vol. 97, no. 8, pp. 4082–4085, 2000, doi: 10.1073/pnas.050589497.

[20] R. J. Etter and M. A. Rex, “Population differentiation decreases with depth in deep-sea gastropods,” *Deep Sea Research Part A. Oceanographic Research Papers*, vol. 37, no. 8, pp. 1251–1261, 1990, doi: 10.1016/0198-0149(90)90041-S.

[21] S. Manel, M. K. Schwartz, G. Luikart, and P. Taberlet, “Landscape genetics: combining landscape ecology and population genetics,” *Trends in ecology & evolution*, vol. 18, no. 4, pp. 189–197, 2003, doi: 10.1016/S0169-5347(03)00008-9.

[22] N. Balkenhol, L. P. Waits, and R. J. Dezzani, “Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data,” *Ecography*, vol. 32, no. 5, pp. 818–830, 2009, doi: 10.1111/j.1600-0587.2009.05807.x.

[23] M. R. Gaither and L. A. Rocha, “Origins of species richness in the Indo-Malay-Philippine biodiversity hotspot: Evidence for the centre of overlap hypothesis,” *Journal of biogeography*, vol. 40, no. 9, pp. 1638–1648, 2013, doi: 10.1111/jbi.12126.

[24] M. A. Rex *et al.*, “Global bathymetric patterns of standing stock and body size in the deep-sea benthos,” *Marine Ecology Progress Series*, vol. 317, pp. 1–8, 2006, doi: 10.3354/meps317001.

[25] R. Danovaro, A. Dell'Anno, A. Pusceddu, C. Gambi, I. Heiner, and R. M̸bjerg Kristensen, “The first metazoa living in permanently anoxic conditions,” *BMC biology*, vol. 8, pp. 1–10, 2010, doi: 10.1186/1741-7007-8-30.

[26] S. A. Siedlecki *et al.*, “Experiments with seasonal forecasts of ocean conditions for the northern region of the California Current upwelling system,” *Scientific Reports*, vol. 6, no. 1, p. 27203, 2016, doi: 10.1038/srep27203.

[27] H. Waga, T. Hirawake, and J. M. Grebmeier, “Recent change in benthic macrofaunal community composition in relation to physical forcing in the Pacific Arctic,” *Polar Biology*, vol. 43, no. 4, pp. 285–294, 2020, doi: 10.1007/s00300-020-02632-3.

[28] H. Saeedi, D. Warren, and A. Brandt, “The Environmental Drivers of Benthic Fauna Diversity and Community Composition,” *Frontiers in Marine Science*, vol. 9, 2022, doi: 10.3389/fmars.2022.804019.

[29] P. Hugenholtz, B. M. Goebel, and N. R. Pace, “Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity,” *Journal of bacteriology*, vol. 180, no. 18, pp. 4765–4774, 1998, doi: 10.1128/jb.180.18.4765-4774.1998.

[30] C. Saccone, C. De Giorgi, C. Gissi, G. Pesole, and A. Reyes, “Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system,” *Gene*, vol. 238, no. 1, pp. 195–209, 1999, doi: 10.1016/s0378-1119(99)00270-x.

[31] W. Li and N. Tahiri, “Host–Virus Cophylogenetic Trajectories: Investigating Molecular Relationships between Coronaviruses and Bat Hosts,” *Viruses*, vol. 16, no. 7, p. 1133, 2024, doi: 10.3390/v16071133.

[32] W. Li and N. Tahiri, “aPhyloGeo-Covid: A web interface for reproducible phylogeographic analysis of SARS-CoV-2 variation using Neo4j and Snakemake,” p. 44, 2023, doi: 10.25080/gerudo-f2bc6f59-00f.

[33] A. Koshkarov, W. Li, M.-L. Luu, and N. Tahiri, *Phylogeography: Analysis of genetic and climatic data of SARS-CoV-2*. 2022. doi: 10.25080/majora-212e5952-018.

[34] D. F. Robinson and L. R. Foulds, “Comparison of phylogenetic trees,” *Mathematical Biosciences*, vol. 53, no. 1, pp. 131–147, 1981, doi: 10.1016/0025-5564(81)90043-2.

[35] W. Li, A. Koshkarov, and N. Tahiri, “Comparison of phylogenetic trees defined on different but mutually overlapping sets of taxa: A review,” *Ecology and Evolution*, vol. 14, no. 8, p. e70054, 2024, doi: 10.1002/ece3.70054.

[36] N. Tahiri, M. Willems, and V. Makarenkov, “A new fast method for inferring multiple consensus trees using k-medoids,” *BMC evolutionary biology*, vol. 18, pp. 1–12, 2018, doi: 10.1186/s12862-018-1163-8.

[37] A. Czarna, R. Sanjuán, F. González-Candelas, and B. Wróbel, “Topology testing of phylogenies using least squares methods,” *BMC Evolutionary Biology*, vol. 6, pp. 1–13, 2006, doi: 10.1186/1471-2148-6-105.

[38] M. R. Smith, “Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets,” *Biology letters*, vol. 15, no. 2, p. 20180632, 2019, doi: 10.1098/rsbl.2018.0632.

[39] M. R. Smith, “Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees,” *Bioinformatics*, vol. 36, no. 20, pp. 5007–5013, 2020, doi: 10.1093/bioinformatics/btaa614.