DEVELOPMENT OF A BIOINFORMATIC ANALYSIS PACKAGE TO TEST GLOBAL PHYLOGEOGRAPHIC RELATIONSHIPS OF SPECIES BY USING GEOTAGGED DNA SEQUENCES FROM GENBANK

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

CANER AKTAŞ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
BIOLOGY

JANUARY 2024

DEVELOPMENT OF A BIOINFORMATIC ANALYSIS PACKAGE TO TEST GLOBAL PHYLOGEOGRAPHIC RELATIONSHIPS OF SPECIES BY USING GEOTAGGED DNA SEQUENCES FROM GENBANK

submitted by **CANER AKTAŞ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Biology, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, **Graduate School of Natural and Applied Sciences**

Prof. Dr. Mesut Muyan
Head of the Department, **Biology**

Prof. Dr. Sertaç Önde
Supervisor, **Biology, METU**

Examining Committee Members:

Prof. Dr. Musa Doğan
Biology, METU

Prof. Dr. Sertaç Önde
Biology, METU

Prof. Dr. İrfan Kandemir
Biology, Ankara University

Assoc. Prof. Dr. Ceyhun Kayıhan
Mol. Biol. and Genetics, Başkent Üniversity

Asst. Prof. Dr. Emre Aksoy
Biology, METU

Date: 25.01.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name : Caner Aktaş

Signature :

ABSTRACT

DEVELOPMENT OF A BIOINFORMATIC ANALYSIS PACKAGE TO TEST GLOBAL PHYLOGEOGRAPHIC RELATIONSHIPS OF SPECIES BY USING GEOTAGGED DNA SEQUENCES FROM GENBANK

Aktaş, Caner
Doctor of Philosophy, Biology
Supervisor : Prof. Dr. Sertaç Önde

January 2024, 141 pages

In this thesis, I introduce “PhyloGeoTagging”, a novel R package and Shiny web application specifically designed to enhance phylogeographic research from a bioinformatics perspective. This package eases key challenges in phylogeographic studies using GenBank sequences, such as selecting, downloading, processing, filtering, and analyzing. It can handle and analyze complete datasets from GenBank searches, offering versatile data filtering options before and after downloading sequences. PhyloGeoTagging enables the elimination of the present problem of lacking geographical coordinate information in GenBank by utilizing Nominatim API for geocoding. The package automates key bioinformatic approaches, including clustering homologous sequences and sequence alignment, as well as conducting advanced phylogeographic analyses, such as haplotype detection and the construction of haplotype networks, genetic barrier analysis, Isolation by Distance, Analysis of Molecular Variance, and investigations of diversity and differentiation parameters. Additionally, two new genetic richness measures, Weighted Haplotype Richness and Cross-Country Weighted Haplotype Endemism, are introduced by me, specifically designed to deal with large and complex datasets. The practical application and effectiveness of this package are demonstrated by using specific case studies and examples, such as the entire organelle DNA data available from GenBank for the family Fagaceae, and the genera *Salvia* and *Apis*. The user-friendly interface of this package, equipped with dynamic maps and infographics, makes complex datasets more accessible and interpretable, supporting both research and educational purposes. By integrating advanced computational tools with phylogeographic research, PhyloGeoTagging enriches our understanding of global biodiversity patterns and provides a foundation for future discoveries in the realm of phylogeography.

Keywords: Phylogeographic Analysis Tools, Automated Phylogeography, GenBank, Geocoding, Shiny Web Application

ÖZ

COĞRAFİ ETİKETLİ GENBANK DNA DİZİLERİ KULLANILARAK TÜRLERİN KÜRESEL FİLOCOĞRAFİK İLİŞKİLERİNİ TEST ETMEK İÇİN BİR BİYOİNFORMATİK ANALİZ PAKETİ GELİŞTİRİLMESİ

Aktaş, Caner
Doktora, Biyoloji
Tez Yöneticisi: Prof. Dr. Sertaç Önde

Ocak 2024, 141 sayfa

Bu tez, filocoğrafik araştırmaları biyoinformatik bir bakış açısıyla geliştirmek üzere tasarlanmış bir R paketi ve Shiny web uygulaması olan "PhyloGeoTagging"i tanıtmaktadır. Bu paket, GenBank veritabanını kullanarak yapılan filocoğrafik veri analizi çalışmalarında karşılaşılan, veri seçimi, indirme, işleme, filtreleme ve analiz gibi zorlukları kolaylaştırmaktadır. Büyük veri kümelerini işleme, analiz etme ve dizileri indirmeden önce ve sonra çeşitli veri filtreleme seçenekleri sunmaktadır. PhyloGeoTagging, GenBank'ta coğrafi koordinat bilgilerinin eksikliği gibi mevcut sorunların üstesinden gelmeyi sağlayan Nominatim API'sini kullanarak coğrafi kodlama yapabilir. Paket, homolog dizilerin kümelemesi ve dizi hizalaması gibi temel biyoinformatik yaklaşımları, haplotip tespiti ve haplotip ağlarının oluşturulması, genetik bariyer analizi, Mesafe ile İzolasyon, Moleküler Varyans Analizi ve çeşitlilik ile farklılaşma parametrelerinin incelenmesi gibi ileri düzey filocoğrafik analizleri otomatikleştirir. Ayrıca, büyük ve karmaşık veri kümeleriyle başa çıkmak için özel olarak tasarlanmış olan Ağırlıklı Haplotip Zenginliği ve Ülke Çapında Ağırlıklı Haplotip Endemizmi gibi iki yeni genetik zenginlik ölçütü, tarafımdan tanıtılmaktadır. Bu paketin pratik uygulaması ve etkinliği, Fagaceae ailesi, *Salvia* ve *Apis* cinsleri gibi örnekler kullanılarak gösterilmiştir. Paketin kullanıcı dostu arayüzü, dinamik haritalar ve infografiklerle donatılmış olup, karmaşık veri kümelerini daha erişilebilir ve yorumlanabilir hale getirir, bu da hem araştırma hem de eğitim çalışmaları için destek sağlar. Gelişmiş hesaplama araçlarını filocoğrafik araştırma ile bütünleştiren PhyloGeoTagging, küresel biyoçeşitlilik desenleri hakkındaki anlayışımızı zenginleştirmekte ve filocoğrafya alanında gelecekte yapılabilecek bazı keşifler için bir temel oluşturmaktadır.

Anahtar Kelimeler: Filocoğrafik Analiz Araçları, Otomatik Filocoğrafi, GenBank, Coğrafi Kodlama, Shiny Web Uygulaması

Dedicated to my family for their support and love. To İlayda Başaran, who somehow started this thesis journey and whose memory continues inspiring many.

ACKNOWLEDGMENTS

I extend my deepest gratitude to my supervisor, Prof. Dr. Sertaç Önde, for his unwavering support and invaluable guidance throughout the entirety of this research journey. His expertise and insights have not only shaped this thesis but also profoundly influenced my personal and professional growth.

I am also immensely thankful to my former co-supervisor, Prof. Dr. Tolga Can, his keen observations and constructive feedback have been pivotal in refining my research work. His dedication and commitment to excellence have been a constant source of motivation.

My heartfelt appreciation goes to the members of my thesis committee. Each of you, have contributed significantly to both the substance and the spirit of this work through your thoughtful critiques and encouraging words.

This thesis is not only a reflection of my efforts but also a testament to the collective support and wisdom of all of you. I am deeply indebted for your invaluable contributions to my academic journey.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiii

LIST OF FIGURES xiv

LIST OF ABBREVIATIONS xviii

CHAPTERS

1 INTRODUCTION 1

1.1. Introduction and Literature Review 1

1.1.1. Historical Background of Biogeography 1

1.1.2. The Emergence of Phylogeography 2

1.1.3. Methods in Phylogeography 3

1.1.4. Sequence Databases in Phylogeography: Challenges and Opportunities 6

1.2. Contributions of the Study 10

2 MATERIAL AND METHODS 13

2.1. The Search Interface - Navigating GenBank Database 14

2.2. ESearch Results 17

2.3. Obtain Sequences and Metadata – Parsing GenBank Data 19

2.4. Enhancing Geospatial Information – Geocoding of GenBank Data 24

2.5. Overview of Geospatial Data - Visualization of GenBank 29

2.5.1. Infographics 30

2.5.2. Distribution Map 31

2.6. GeoGenetic Synthesis - Phylogeography of GenBank Data 34

2.6.1 Cluster & Align 36

2.6.2. Haplotype Identification 43

2.6.3. Haplotype Richness 47

2.6.4. ΦST Barriers 51

2.6.5. Isolation by Distance (IBD) 56

2.6.6. Analysis of Molecular Variance (AMOVA) 59

2.6.7. Diversity and Differentiation Parameters 64

3 RESULTS and DISCUSSON 69

3.1. Examples 69

3.1.1. Comprehensive Example: Family Fagaceae 69

3.1.2. Example: Genus *Salvia* 90

3.1.3. Mitochondrial DNA Example: Genus *Apis* 104

3.2. Limitations and Future Prospects 115

4 CONCLUSION 119

REFERENCES 123

CURRICULUM VITAE 141

LIST OF TABLES

TABLES

Table 1. Merging the results of the Mantel test from IBD, AMOVA, and Hapstep for 41 clusters 84
Table 2. AMOVA results for the *Salvia* dataset 95
Table 3. Filtered geocoding results showcasing data from Turkey 101
Table 4. Summary of Hapstep analysis for 59 clusters in *Apis* species and subspecies 110

LIST OF FIGURES

FIGURES

Figure 1. GenBank and WGS Growth Trends, 1982-2023 8

Figure 2. The Search Tab of the Shiny site, showcasing the primary interface for querying the GenBank Nucleotide database using the ESearch utility. 14

Figure 3. The dropdown panel for modifying optional ESearch parameters related to sequence retrieval. 16

Figure 4. Overview of the ESeach Tab showing the output of search process. 18

Figure 5. Overview of the Parser Tab showing the dropdown panel with selectable fields for customizing the GenBank data parsing process...................................... 20

Figure 6. Illustration of the Parser interface, showcasing the Feature table............ 22

Figure 7. The Parser Log window displaying a record of excluded entries due to incomplete data, with options for downloading the detailed log............................ 24

Figure 8. Overview of the Geocoder Tab, available after Parser Feature table is ready 25

Figure 9. Overview of the Geocoder table, available after geocoding. 29

Figure 10. Donut chart in the Infographics Tab illustrating the distribution of a selected feature from the Geocoder table. 30

Figure 11. Interactive map display on the Distribution Map Tab, illustrating the initial view with options for selecting coordinate types and viewing geospatial data from the Geocoder table 32

Figure 12. Jitter points feature on distribution map for overlapping markers and geometry viewing on click. 33

Figure 13. Customization options on the distribution map, demonstrating the selection of factors for color-coding and labeling, and the color-picker tool. 34

Figure 14. The Cluster & Align Tab, showcasing options for customizing clustering parameters.. 38

Figure 15. Interactive circular plot displaying clustered sequence data, with a highlighted example showing detailed taxonomic information upon user interaction. 41
Figure 16. Interactive circular plot displaying clustered and aligned sequence data, with a highlighted example showing detailed taxonomic information upon user interaction, alignment status shown. 43
Figure 17. Interactive map with pie-chart markers indicating haplotype distribution across geographic locations, with options to adjust cluster visibility and color settings. 45
Figure 18. Detailed view of the parsimony network and distribution map highlighting the genetic relationships among haplotypes, with user options for haplotype color customization and network display. 47
Figure 19. The color-coded map that illustrates the weighted haplotype richness across countries available. 51
Figure 20. An interactive raster map from the ΦST Barriers tab, showcasing the spatial distribution of the selected statistic 54
Figure 21. The ΦST Barriers visualization emphasizes local barriers using adjusted weight settings 55
Figure 22. Interactive IBD plot displaying the genetic versus geographic distances among clusters 58
Figure 23. The AMOVA Tab showcasing results with the presence of multiple genera and species 62
Figure 24. Overview of the AMOVA Tab showing the assessment of genetic variation within species and between populations, with a focus on a selected level, *Quercus aucheri*. 63
Figure 25. Interface of the Hapstep Tab in the Application. 67
Figure 26. Overview of the ESeach Tab showing the output of search for Chloroplast DNA sequences (cpDNA) of Fagaceae. 70
Figure 27. Graphical representation of Organism feature of the GenBank cpDNA records for Fagaceae. 71

Figure 28. Graphical representation of Organism feature of the GenBank cpDNA records for Fagaceae. 72
Figure 29. Overview of the ESeach Tab showing the output of search for Chloroplast DNA sequences (cpDNA) of Fagaceae. 73
Figure 30. Weighted Haplotype Richness analysis results for Fagaceae.. 75
Figure 31. Mantel test results for Cluster-253 77
Figure 32. Haplotype map and parsimony network for the Cluster-253. 78
Figure 33. Detailed AMOVA tables for Cluster-136, in which the taxonomic heterogeneity contributed to the variance. 80
Figure 34. Haplotype map and parsimony network for the Cluster-159, illustrating moderate level of haplotype sharing between *Quercus coccifera*, *Quercus alnifolia*, and *Quercus ilex*. 81
Figure 35. ΦST Barrier analysis for *Quercus serrata* in Cluster-1336 and comparative view across all clusters 85
Figure 36. Haplotype distribution and parsimony network for *Quercus coccifera* in Cluster-32. 87
Figure 37. ΦST Barrier analysis in the Mediterranean showing the effects of chloroplast capture 89
Figure 38. Global ΦST Barriers for the Fagaceae Family 90
Figure 39. Donut chart representation of unique organism names in the *Salvia* dataset. 91
Figure 40. Donut chart of cpDNA sequence sampling by country 92
Figure 41. Visualization of cluster locations with jittered points indicating the central coordinates of China, highlighting the challenge of missing fine level geographic data for phylogeographic analysis. 94
Figure 42. Haplotype distributions and parsimony network for Cluster-2, reflecting the phylogeographic patterns of *Salvia officinalis*. 99
Figure 43. Manual integration of locations and haplotype distributions. 103
Figure 44. Donut chart representation of unique organism names in the *Apis* dataset. 104

Figure 45. Donut chart representation of the sequence region information from the “Product” Feature in the Associated Records' Metadata. 105
Figure 46. Weighted haplotype richness map for countries based on the *Apis* mitochondrion dataset. 106
Figure 47. Cross-Country Weighted Haplotype Endemism map for countries based on the *Apis* mitochondrion dataset. 107
Figure 48. Colored dot plot of population differentiation ΦST across taxa 108
Figure 49. Haplotype differentiation in Cluster-95.. 111
Figure 50. Haplotypes map and parsimony network for Cluster-241.................. 113
Figure 51. Global ΦST Barriers in the *Apis* dataset... 114

LIST OF ABBREVIATIONS

ABBREVIATIONS	
PCR	Polymerase Chain Reaction
mtDNA	Mitochondrial DNA
cpDNA	Chloroplast DNA
RFLP	Restriction Fragment Length Polymorphism
NCBI	National Center for Biotechnology Information
BOLD	Barcode of Life Data System
GBIF	Global Biodiversity Information Facility
UIDs	Unique Identifiers
GI	GenInfo Identifiers
XML	Extensible Markup Language
IBD	Isolation by Distance
AMOVA	Analysis of Molecular Variance
MSA	Multiple Sequence Alignment
CA	Cluster & Align
WHR	Weighted Haplotype Richness
CCWHE	Cross-Country Weighted Haplotype Endemism
DW	Distance Weighting
IDW	Inverse Distance Weighting
D	Nei's Raw Distance

CHAPTER 1

INTRODUCTION

1.1. Introduction and Literature Review

1.1.1. Historical Background of Biogeography

The impact of Charles Darwin's groundbreaking work "On the Origin of Species" in the field of biological sciences is significant (Darwin, 1964). During his famous journey on the HMS Beagle, Darwin's detailed observations, especially in the Galápagos Islands, were key in collecting the real evidence that would later become the core of his evolutionary theories (Losos & Ricklefs, 2009; Johnson & Baarli, 2015). While the full synthesis of his ideas into the theory of evolution by natural selection occurred after his return, the geographical observations made during the voyage, particularly the variation in species he noted across different islands, played an important role in shaping his understanding of natural selection and the diversification of species (Sulloway, 1984). Over time, he recognized that both geographical isolation and biological variation play significant roles in species evolution. This new understanding led to a more detailed view of evolution, emphasizing the complex interaction of various factors, including geographical and biological elements, in species diversification (Briggs, 2009).

Alfred Russel Wallace's work, especially in “The Geographical Distribution of Animals” was of great significance in establishing fundamental principles of natural history and species distribution (Wallace, 2011). His systematic approach involved not only identifying key biogeographic boundaries like Wallace's Line, which defines the faunal regions of Asia and Australia, but also cataloging species distributions and correlating them with geological and climatic shifts, thereby providing comprehensive insights into how historical processes have influenced the present-day distribution of species (Riddle & Hafner, 2010; Costa, 2013).

These early concepts, though not yet termed phylogeography, established a fundamental understanding of how geographic factors influence species evolution. This groundwork, rooted in initial geographical observations of species variations, set the stage for future phylogeographic exploration. The advent of genetic studies, particularly the development of molecular techniques in the late 20th century, enabled scientists to delve deeper into these geographical patterns. The integration of genetics into biogeography and evolutionary studies has not only led to a more detailed and comprehensive understanding of species' historical dispersion and evolution in diverse environments but also, through the combination of genetic data with geographical information, provided profound insights into speciation, migration, and adaptation processes (Riddle et al., 2008; Hickerson, 2016).

1.1.2. The Emergence of Phylogeography

Building on knowledge in biogeography, genetics, and our understanding of biodiversity and evolutionary history, the field of phylogeography emerged as a distinct area of study. The term phylogeography, first introduced in Avise et al. using mitochondrial DNA (mtDNA) in 1987, is chiefly concerned with understanding how past and present events affect the current geographical distribution of genes, populations, and closely related species. Advanced from conventional population genetics and phylogenetics, phylogeography was developed to investigate genetic differentiation within and among taxa across different geographical areas, focusing particularly on the biogeographical background of the group of interest (Avise, 2000;

Arbogast, 2001). It also acts as a link between population genetics and phylogenetics, utilizing coalescent theory to merge population-level phenomena like gene flow, genetic drift, and changes in population size into practical research (Jackson et al., 2017).

Historical biogeography, an interdisciplinary field, explores the influence of historical, geological, climatic, and ecological factors on the past and present distribution of species. Within this field, the study of geographical and evolutionary relationships among organisms has long been a focus. However, two pivotal advancements during the 1960s and 1970s significantly shaped the foundation of modern phylogeography. Primarily, it was the widespread adoption of cladistic thought that brought a new perspective to evolutionary relationships. The emergence of plate tectonics theory, which revolutionized our understanding of Earth's geological and geographic history, was equally crucial (Zimmerer,1994; Arbogast, 2001, de Queiroz, 2005; Tiffney, 2008; Trewick, 2017).

In the mid-1970s, population genetic analyses increasingly turned their attention to mitochondrial DNA sequences (Avise, 1998; Avise et al., 2016). This shift coincided with the emergence of the polymerase chain reaction (PCR) technology, a landmark development in the evolution of phylogeography. PCR, enabling the replication of millions of copies of a DNA segment, significantly improved the accessibility of genetic data, advancing the field forward (Arbogast, 2001; Bowen & Karl, 2007).

1.1.3. Methods in Phylogeography

The advent of PCR in the 1980s initiated significant changes in phylogeographic research. The widespread accessibility of DNA sequencing, enhanced by the integration of molecular techniques with traditional biogeographic methods, has led to a deeper understanding of the evolutionary dynamics and spatial genetic diversity.

This approach has provided insights into how historical and geographical factors shape species distribution and variation (Riddle et al., 2008; Kholodova, 2009; Mantooth & Riddle, 2011; Cutter, 2013; Held, 2014).

This breakthrough in PCR and DNA sequencing technology marked a turning point in genetic research, as it significantly enhanced the accessibility of information encoded in DNA sequences. This was not just a major advancement in laboratory techniques, but also a significant progress in computational methods that maximized the utility of genetic data. These advancements in both laboratory and computational domains have collectively advanced phylogeographic research forward (Avise,1998; 2000).

Mitochondrial DNA (mtDNA) has been a cornerstone in the evolution of phylogeographic studies, offering unique insights into the genetic history of species. Its distinct characteristics, such as maternal inheritance, absence of recombination, and a relatively rapid mutation rate, make mtDNA an invaluable tool for tracing evolutionary lineages and understanding species' historical migrations (Zhong et al., 2020; Doan et al., 2021; Ghanavi et al., 2022; Rusin, 2023). These properties facilitate the mapping of genetic lineages over temporal scales, thereby offering researchers a valuable perspective into the historical migratory patterns, evolutionary processes, and DNA barcoding studies of a diverse array of organisms, including various animal and plant species, as well as other eukaryotic lineages (Gill et al., 1993; Fedorov, 1999; Weider, 1999; Petit & Vendramin, 2007; Vidya et al., 2008; Semerikova, 2014; Benn Torres, 2016; Zheng et al., 2018; Torroni et al., 2020; Zhong et al., 2020; da Silva Ferrette et al., 2021; Al-Jumaili et. al, 2022; P. Lumogdang et al., 2022). The choice to utilize mtDNA in the nascent stages of phylogeography was driven by its ability to offer a more straightforward and detailed view of genetic variation across generations and geographical spaces. This focus on mtDNA was instrumental in the early breakthroughs of the field, allowing scientists to uncover the complex interplay between genetics, geography, and evolutionary history (Avise et al., 1987).

In parallel with the study of mtDNA in eukaryotic organisms, research in phylogeography has also turned its attention to chloroplast DNA (cpDNA). Similar to mtDNA, cpDNA is maternally inherited in most angiosperm species, albeit with some exceptions, offering another valuable genetic marker for studying evolutionary patterns and lineages (Petit & Vendramin, 2007; McCauley et al., 2007). Its non-Mendelian, uniparental transmission without recombination, relatively small genome size, great abundance in cells, and the conserved nature of its sequence and structural forms allow cpDNA to be effectively used in plant phylogeographic studies (Rieseberg & Soltis, 1991). These studies include investigating migration patterns from glacial periods to the present (Petit et al., 2002a; Cotrell et al., 2005; López de Heredia et al., 2007; Nevill et al., 2010; Douda et al., 2014; Shepherd et al., 2017; Gömöry et al., 2020), examining population structures (Petit et al., 2002b; Pettenkofer et al., 2019; Wöhrmann et al., 2020), exploring intra- and interspecific hybridizations (Coyne & Orr, 2004; Curtu et al., 2007; Zhang et al., 2015; Tamaki & Okada, 2014; Tekpinar et al., 2021), and conducting DNA barcoding studies (Simeone et al., 2013; Bi et al., 2018; Amandita et al., 2019).

Initial investigations into organelle DNA (mtDNA and cpDNA) employed indirect methods such as restriction fragment length polymorphism (RFLP), rather than direct sequencing. However, this method of analysis was soon supplemented and largely overtaken by DNA sequencing methods, which offered more detailed and comprehensive genetic information. By the late 1990s, DNA sequencing had become the standard approach within a short time.

Advances in DNA sequencing and analysis methods have enabled the examination of phylogeographic relationships across various taxonomic groups simultaneously. This approach, known as comparative phylogeography, investigates biogeographical regions and the mechanisms responsible for observed phylogenetic relationships. Species with similar ecologies and distributions often tend to have similar phylogeographical structures, highlighting the interconnectedness of ecological and geographical factors in shaping genetic diversity. By comparing multiple taxa, phylogeographers can elucidate shared biogeographical histories and understand how past events have influenced present-day biodiversity.

1.1.4. Sequence Databases in Phylogeography: Challenges and Opportunities

The proliferation of DNA sequencing has led to the creation of extensive sequence databases, which have become crucial tools across various fields of phylogenetic research. While these databases serve a wide range of scientific inquiries, they are particularly valuable for phylogeographers, who can access a vast array of genetic data from diverse species. Sequence databases, such as the NCBI GenBank, play a pivotal role in modern biological research (Benson et al., 1993; Benson et al., 2017; Leray et al., 2019; Sayers et al., 2022a). These databases are repositories of genetic sequences collected from numerous organisms across the globe. NCBI GenBank is one of the largest and most widely used public databases for nucleotide sequences (Sayers et al., 2022a; 2022b). It serves as a critical resource for researchers, providing access to a vast collection of DNA sequences submitted by scientists worldwide.

The rapid growth in the number of DNA records submitted to the NCBI nucleotide database (GenBank) in recent years is a remarkable trend. With many sequences available in the GenBank database, researchers now have more chances to do in-depth molecular studies using in silico methods to conduct secondary data analysis (Ficetola et al., 2010; Azara & Yakubu, 2014; Leray et al., 2019). This easy access reduces the need for extensive wet-lab work, as scientists can study, analyze, and use these sequences remotely. This development not only streamlines the research process but also opens new possibilities for studies that were previously limited by the availability of physical samples. It heralds a new era in molecular biology where data accessibility plays a crucial role in advancing our understanding of genetics and its applications.

Phylogeographic studies, inherently linked with geographic coordinate data, are essential for analyzing the collection sequences. While mtDNA and cpDNA sequence data is extensively utilized for phylogeographic analyses of angiosperms, these investigations have predominantly focused on a local scale, encompassing a limited range of species. In contrast, global-scale phylogeographic analyses, as documented in the literature, primarily feature microorganisms and viruses, with notable examples being the studies by Gagneux & Small (2007), Albery et al. (2020), and Edwards et al. (2019). Research using mitochondrial DNA (mtDNA) for phylogeographic studies is less common but has been conducted on insects (Zahiri et al., 2019), reptiles (Jensen et al., 2019), and mammals (Kraft et al., 2020). Plant studies in this domain, such as those by Der et al. (2009), are relatively scarce. A notable demonstration of the value of data reanalysis is evident in the study by Miraldo et al. (2016). This research meticulously compiled mitochondrial DNA (mtDNA) sequences from almost 2000 terrestrial mammal and amphibian species, revealing a trend of higher genetic diversity in tropical areas and lower diversity in regions with dense human populations. The extensive effort involved in this analysis was substantial, requiring the manual downloading of GenBank and BOLD accessions that included geographic coordinates, as well as direct communication with researchers for additional data. The manual process used in the Miraldo et al. (2016) study naturally imposed a limit on the number of species that could be included in their analysis.

Quantifying the spread of genetic variation geographically is crucial for understanding evolutionary dynamics and current biodiversity. This is a fundamental aspect of landscape genetics and phylogeographic research. The NCBI GenBank's vast database, hosting over two hundred million DNA sequences, is pivotal in this area (Figure 1). However, a significant challenge is the lack of essential metadata, particularly the collection locations of organisms. Most of these sequences lack the necessary locality data, which limits their potential use (Sidlauskas et al., 2009). It is found that merely 7% of GenBank's barcoding gene entries provide latitude and longitude data (Marques et al., 2013). Peterson et al. (2018) estimated that around

90% of biodiversity data are underutilized due to missing geographical information. As of November 2023, a basic search in GenBank's Nucleotide database using the ESearch utility reveals that the availability of latitude-longitude data remains limited. For the Angiosperm records, only a fraction of the total entries include geographical coordinates, representing a mere 0.015% of all records. Focusing specifically on chloroplast DNA (cpDNA) records, the percentage of entries with geographic data is 0.11%, while for mitochondrial DNA (mtDNA) records, this figure stands at 0.05%. In the case of animal mtDNA, the proportion of sequences with geographical coordinate data is higher, at 0.42%. This data underscores the ongoing challenge of integrating spatial metadata in genetic databases.

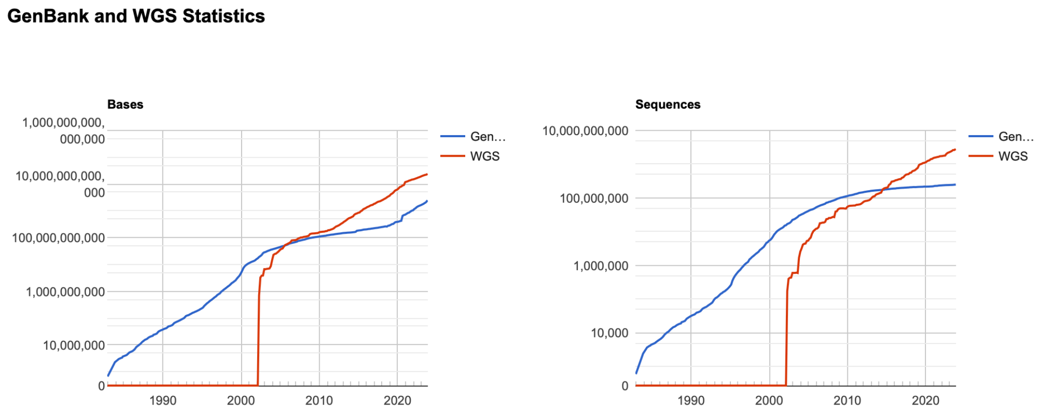


Figure 1. GenBank and WGS Growth Trends, 1982-2023. The left graph shows base count increases, and the right graph sequence record growth. Source: https://www.ncbi.nlm.nih.gov/GenBank/statistics/

Building on the foundation of comparative phylogeography, which examines phylogeographic relationships across various groups (Bermingham & Moritz, 1998; Kholodova, 2009), a significant opportunity arises when analyzing a large number of species. This approach can reveal phylogeographic patterns on both a global scale and at the community level. GenBank's Database, with its geotagged or georeferenced sequences, provides a valuable resource for such studies. While the number of geotagged chloroplast DNA (cpDNA) accessions in GenBank is considerably lower than the total number of accessions, these records are vital for conducting phylogeographic analysis at a global scale. So far, there have been a few studies that explore the meta-analysis possibilities of non-plant phylogeographic data previously collected in GenBank and suggest methods for accessing this data (Gratton et al., 2016; Porter & Hajibabae, 2018). However, these studies primarily focus on assessing the suitability of GenBank sequences for automated phylogeographic analysis and have not yet delved deeply into applying these data in comprehensive comparative phylogeographic research. This underscores the untapped potential of GenBank’s geotagged sequences in enhancing our understanding of global biodiversity patterns and the evolutionary processes that shape them.

Concurrent with this work presented in this thesis, Pelletier et al. (2022) introduced **phylogatR**, an innovative database and toolkit designed for aggregating and repurposing phylogeographic data. **PhylogatR** compiles genetic sequences from major databases like GenBank, GBIF (https://www.gbif.org), and BOLD (http://www.boldsystems.org/index.php), establishing a comprehensive platform for phylogeographic studies. Equipped with R scripts, it aids in data curation, analysis, and educational applications, thus enhancing global research into genetic diversity and population structure. This tool aims to improve data accessibility for researchers and educators, aligning with open-data science principles. It provides insights into the effects of sampling on genetic diversity and includes robust protocols for data cleaning, standardization, and quality checks, such as verifying geographic coordinates and screening for misidentified sequences or alignment errors. **PhylogatR**'s suite of analytical tools includes functions for calculating nucleotide diversity and conducting regression analysis to explore relationships between geographic sampling and genetic diversity, applying corrections for multiple testing.

It also identifies data outliers and inconsistencies, enhancing database accuracy. The tool extends its utility to educational settings by integrating real research experiences into classroom instruction, facilitating exploration of evolutionary processes, and encouraging novel phylogeographic research. Offering its capabilities under a Creative Commons Attribution License, **phylogatR** potentially represents a significant advancement in phylogeography. It not only enables large-scale meta-analyses but also contributes to a deeper understanding of global biodiversity patterns, marking a substantial step forward in the field of phylogeography and open-access scientific resources.

1.2. Contributions of the Study

The study presented in this thesis makes significant contributions to the field of phylogeography, particularly in the realm of automated analysis and global scale research. Historically, phylogeography studies have often been constrained by the focus on small taxonomic groups or geographic scales, typically involving well-designed sampling procedures with robust statistical approaches. Comparative phylogeography, on the other hand, has largely been limited to descriptive review articles or case studies using GenBank metadata (Carstens et al., 2018; Riddle, 2016; Gratton et al., 2016). This thesis work addresses these limitations by employing a comprehensive bioinformatics approach for global phylogeographic analysis of records available in GenBank.

The core of this thesis is the development of **PhyloGeoTagging**, a specialized R package (R Core Team, 2022) and interactive website designed to streamline documentation, downloading, geocoding, and phylogeographic analyses of GenBank sequences, thereby accommodating a wide range of geographical and taxonomic scales. The scripting of R functions for bulk-downloading and extracting various features from GenBank records, such as accession numbers, addresses, latitude-longitude data, organism names, DNA sequences, and more, is an integral part of this advancement. Additionally, the ability to geocode addresses where latitude-longitude coordinate metadata is lacking enhances the dataset's utility for phylogeographic analysis.

A significant element of this thesis is the development of a user-friendly graphical interface using the Shiny package (Chang et al., 2022), resulting in an interactive website. This platform provides interactive phylogeographic analysis methods using geotagged GenBank sequences, contributing to the field's advancement. The website serves as both a practical tool for researchers and an educational resource, integrating real data and computational methods into educational settings, in line with the study’s goal of facilitating more accessible and accurate global phylogeographic analysis.

Beyond merely facilitating data gathering and initial processing, this study has bridged a critical gap in the field by incorporating highly optimized and well-established phylogeographic analysis methods. These advanced techniques offer a more refined and comprehensive understanding of the data, overcoming the limitations of existing tools. Thus, this thesis not only adds to methodological developments in phylogeography but also significantly expands the range and depth of research that can be conducted using GenBank's extensive datasets. This marks an improvement over existing methods, which may not fully exploit the potential of large-scale data from sources like GenBank. The use of these advanced methods enables more detailed and nuanced insights into phylogeographic patterns, contributing to a deeper understanding of species evolution and biogeographic distributions.

In conclusion, this thesis represents a valuable step in automated phylogeography, both theoretically and practically. It offers a novel perspective on conducting large-scale analyses and enhances our understanding of global biodiversity patterns. The inclusive nature of this package, offering an array of analysis methods in its initial version, sets the foundation for future enhancements. The completion and launch of the website, along with the detailed technical documentation in the thesis, will contribute to the phylogeography community and open-data science.

CHAPTER 2

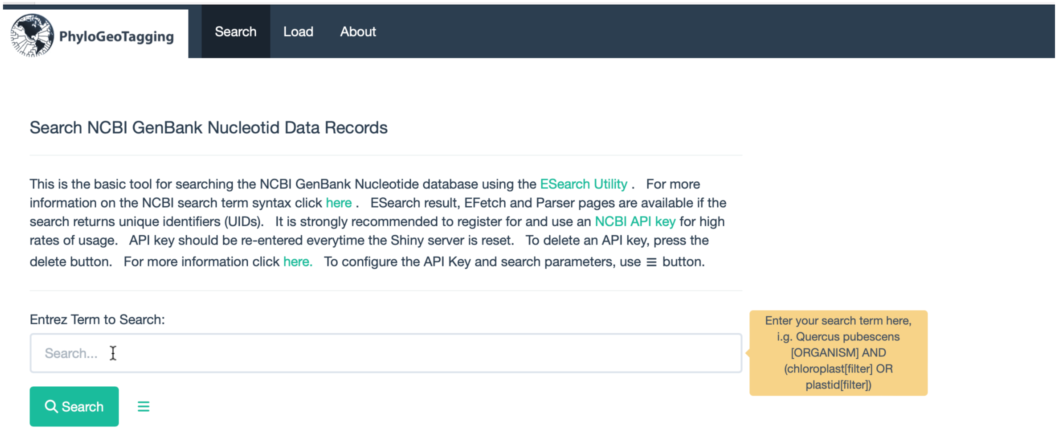
MATERIAL AND METHODS

This thesis introduces a comprehensive Shiny web application and R package, which are collectively referred to as **PhyloGeoTagging**. Designed to facilitate phylogeographic studies, this package allows users to select and download diverse sequences from the GenBank database, tailored to their specific research requirements. It enables efficient extraction and processing of sequence data along with associated metadata, particularly focusing on geographical coordinates. The package and application are equipped with geocoding capabilities to supplement datasets lacking detailed latitude-longitude information. They automate crucial tasks such as clustering homologous sequences and performing sequence alignment, thus streamlining the preparation of data for targeted analysis for the next step. The culmination of these functionalities is the ability to conduct comprehensive phylogeographic analyses of sequences, offering an integrated solution for researchers in the field.

With the comprehensive suite of the Shiny web application and R package, referred to as **PhyloGeoTagging**, a wide array of phylogeographic research tasks becomes more accessible and efficient. Each element of this package is expertly designed to enhance the entire spectrum of phylogeographic research, from data retrieval to in-depth analysis. The following sections will delve into the key functionalities that this package offers. The Shiny site is equipped with helpful tooltips on most buttons, input fields, and graphical elements to assist users in navigating the application and to provide on-the-spot guidance.

2.1. The Search Interface - Navigating GenBank Database

Upon accessing the Shiny site, users first encounter the Search tab, which serves as the portal to the GenBank Nucleotide database through the ESearch utility (Figure 2). After conducting a search, the ESearch Tab appears and is selected, indicating the search is done. If the search yield any unique identifiers (UIDs), the Parser Tab also become available, unlocking further functionality for data retrieval and processing.



Search NCBI GenBank Nucleotid Data Records

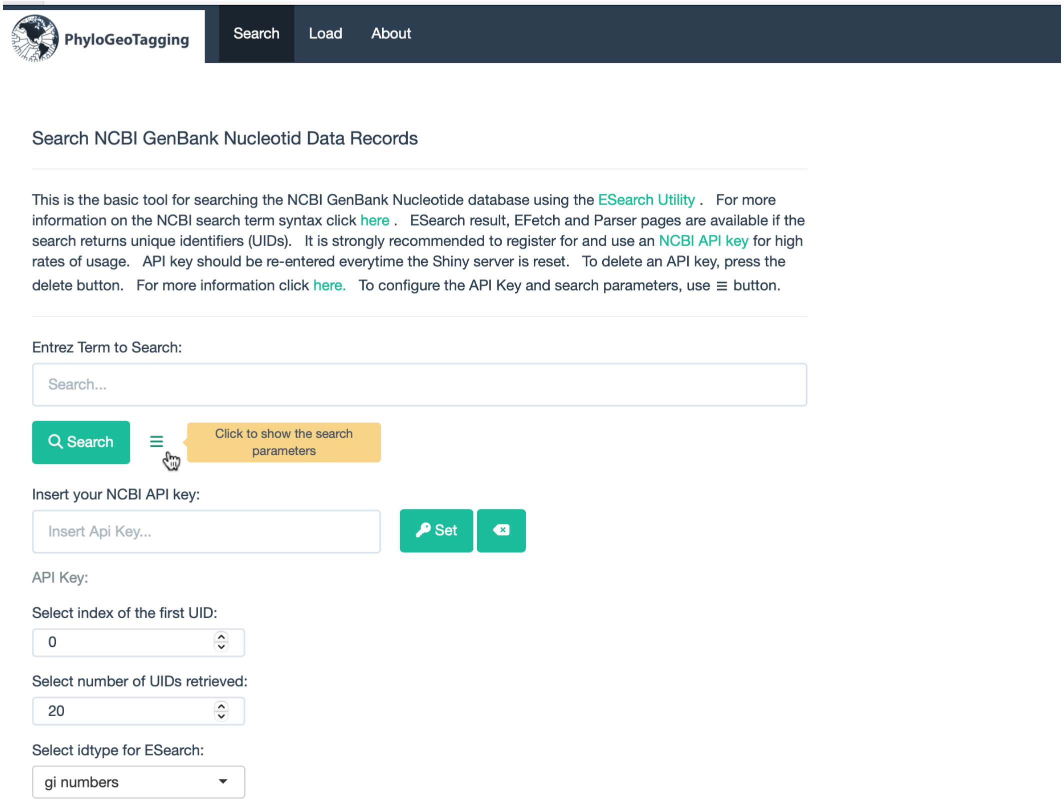
This is the basic tool for searching the NCBI GenBank Nucleotide database using the ESearch Utility . For more information on the NCBI search term syntax click here . ESearch result, EFetch and Parser pages are available if the search returns unique identifiers (UIDs). It is strongly recommended to register for and use an NCBI API key for high rates of usage. API key should be re-entered everytime the Shiny server is reset. To delete an API key, press the delete button. For more information click here. To configure the API Key and search parameters, use ≡ button.

Enter your search term here, i.g. Quercus pubescens [ORGANISM] AND (chloroplast[filter] OR plastid[filter])

Figure 2. The Search Tab of the Shiny site, showcasing the primary interface for querying the GenBank Nucleotide database using the ESearch utility.

The search utility is designed to accept a range of search terms, from text-based matching to search using specific field names. For example, entering Quercus pubescens by itself retrieves records wherever this term appears in the database, equivalent to using Quercus pubescens[WORD], Quercus pubescens[ALL] or Quercus pubescens[ALL FIELDS]. By specifying the search with Quercus pubescens[ORGANISM], the search targets the organism attribute of the records, returning sequences exclusively from the *Quercus pubescens* species. Users have the option to use abbreviated or lowercase terms and field names, such as quercus pubescens[orgn] for organism. To extend the search to broader taxonomic categories, terms for genus (e.g., Quercus[ORGN]), family (e.g., Fagaceae[ORGN]), and higher classifications can be used. Refinement of the search is further enhanced by applying filters for sequence type; for instance, Quercus pubescens [ORGN] AND (chloroplast[FILTER] OR plastid[FILTER]) specifically searches for chloroplast DNA sequences of *Quercus pubescens*. To locate animal mitochondrial DNA sequences with geographic coordinates, the search term Metazoa [ORGN] AND mitochondrion[FILTER] AND src lat lon[PROP] will yield all relevant animal mitochondrial DNA sequences that include latitude and longitude information from GenBank submissions. For updated and detailed guidelines on the NCBI search term syntax and the Entrez Programming Utilities, users can refer to the hyperlink included in the explanatory text under this tab. All available fields for the nucleotide database can be viewed dynamically through R by using the command entrez_db_searchable("nucleotide") from the **rentrez** package (Winter, 2017).

The ESearch utilitys default parameters, *db* and *use_history* are pre-configured to *nucleotide* and *TRUE*, respectively, to optimize search results for the purpose of this study. These parameters are utilized by the entrez_search function of the **rentrez** R package, with the user's input in the search box serving as the *term* parameter. The search action utilizes “https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi” as its base URL and combines the entered search terms to form the query. This returns a web history object that is stored within the R environment during the Shiny session and used in subsequent calls to NCBI. Some optional ESearch parameters related to sequence retrieval can be modified using the dropdown panel accessible via the ☰ button (Figure 3).



Search NCBI GenBank Nucleotid Data Records

This is the basic tool for searching the NCBI GenBank Nucleotide database using the ESearch Utility . For more information on the NCBI search term syntax click here . ESearch result, EFetch and Parser pages are available if the search returns unique identifiers (UIDs). It is strongly recommended to register for and use an NCBI API key for high rates of usage. API key should be re-entered everytime the Shiny server is reset. To delete an API key, press the delete button. For more information click here. To configure the API Key and search parameters, use ≡ button.

Entrez Term to Search:

Search...

Insert your NCBI API key:

Insert Api Key...

API Key:

Select index of the first UID: 0

Select number of UIDs retrieved: 20

Select idtype for ESearch: gi numbers

Figure 3. The dropdown panel for modifying optional ESearch parameters related to sequence retrieval.

For enhanced performance, particularly under high usage conditions, it is recommended that users insert an NCBI API key within the parameters panel. Once the API key is configured, it will persist across all NCBI database queries within the session. It should be noted, however, that the API key may need to be re-entered if the R session ends for any reason, as the key is not stored permanently. On a local machine, the API key will remain until the R session is explicitly reset. In a server environment, ending the Shiny session might also reset the R session, necessitating the re-entry of the API key. The API key can be re-entered directly in the provided field or deleted using the “Delete” button.

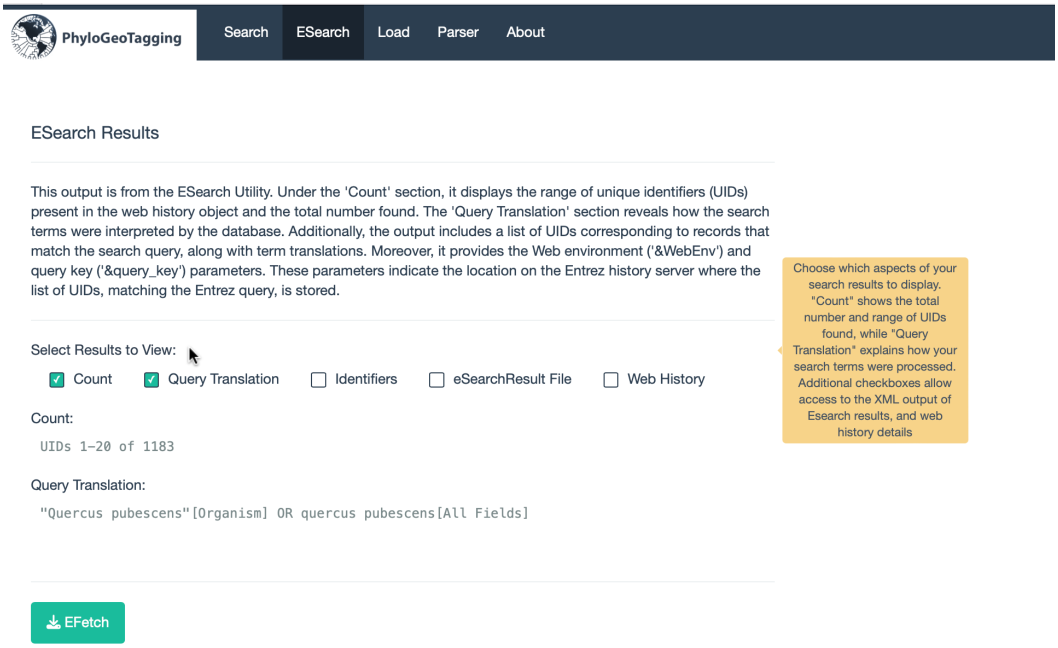
The parameter panel also allows users to adjust the index of the first UID to be displayed, corresponding to Entrez's *retstart* parameter. *retstart* sets the starting point for the XML output display and can be anywhere from 0, indicating the first record, to the total number of records. If a specified index exceeds the available range, it will automatically adjust to fall within the valid range.

Another parameter that users can control is *retmax*, which determines the maximum number of records retreived from the GenBank database in a single request. This can be adjusted using the numeric input labeled “Select number of UIDs retrieved”. The *retmax* parameter works in conjunction with *retstart*; if the sum of *retstart* and *retmax* exceeds the total available records, only the remaining records from *retstart* to the end of the dataset are returned. By default, *retmax* is set to 20, the same as the Entrez standard setting, but users have the option to increase this number to retrieve more records at once. However, it's important to note that a higher *retmax* value may result in slower response times due to the increased volume of data being processed. Therefore, for basic searches aimed at determining the number of available sequences matching the search terms in the database without fetching the sequences, it is recommended to use the default parameters.

The last selectable parameter determines the identifier type (*idtype*) that will be returned from the nucleotide database. By default, the system is set to return GenInfo Identifiers (GI numbers). However, users have the option to select *acc* if they prefer to retrieve identifiers as accession numbers with versions.

2.2. ESearch Results

After a search is executed, the interface transitions to the ESearch Tab to display the search outcomes (Figure 4). This tab provides a comprehensive overview of the search results, confirming the system has successfully processed the query.



ESearch Results

This output is from the ESearch Utility. Under the 'Count' section, it displays the range of unique identifiers (UIDs) present in the web history object and the total number found. The 'Query Translation' section reveals how the search terms were interpreted by the database. Additionally, the output includes a list of UIDs corresponding to records that match the search query, along with term translations. Moreover, it provides the Web environment ('&WebEnv') and query key ('&query_key') parameters. These parameters indicate the location on the Entrez history server where the list of UIDs, matching the Entrez query, is stored.

Select Results to View:

☑ Count ☑ Query Translation ☐ Identifiers ☐ eSearchResult File ☐ Web History

Count:

```
UIDs 1-20 of 1183
```

Query Translation:

```
"Quercus pubescens"[Organism] OR quercus pubescens[All Fields]
```

Choose which aspects of your search results to display. "Count" shows the total number and range of UIDs found, while "Query Translation" explains how your search terms were processed. Additional checkboxes allow access to the XML output of Esearch results, and web history details

Figure 4. Overview of the ESeach Tab showing the output of search process.

By default, the “Count” and “Query Translation” outputs are visible. In the “Count” section, users can see the range of UIDs present in the web history object and the total number found. The “Query Translation” section reveals how the search terms were interpreted by the database. Additional checkboxes offer a more detailed view of the search output, including a list of unique identifiers (UIDs) that correspond to records matching the entered search criteria. Depending on the user's choice in the parameters panel of the Search tab, these identifiers may be GI numbers or “accession.version” identifiers. Selecting the *eSearchResult* checkbox displays the XML output of ESearch results. “Web History” details the web history associated with the search session.

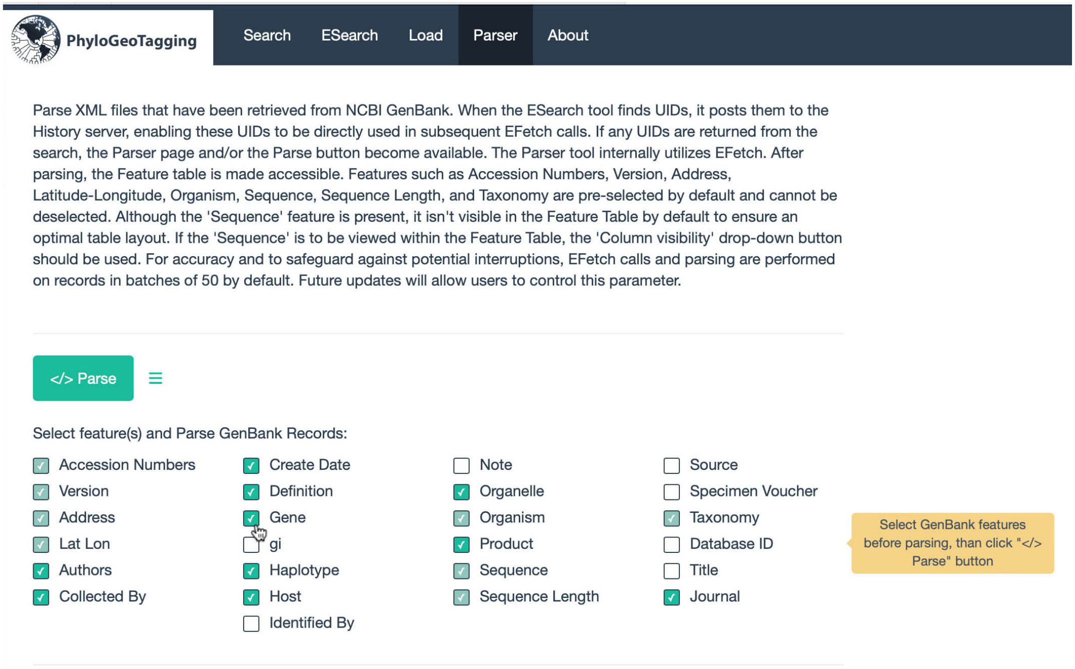
At the bottom of the ESearch tab, the EFetch button becomes available once the search returns UIDs. EFetch enables users to retrieve GenBank records in standardized XML format directly from the NCBI database. It fetches data records that correspond to the list of IDs contained in the web history object. While EFetch allows for the downloading of GenBank records, it is important for users to note that direct loading of these XML files for further analysis within this package is not available. The functionality is intended as a convenience for users who require sequence data for analyses performed outside of this website. However, alternative mechanisms for sequence retrieval and loading are provided through various functionalities across different tabs within this application.

2.3. Obtain Sequences and Metadata – Parsing GenBank Data

After UIDs are retrieved from a successful search, they are sent to the NCBI's History server, enabling their use in subsequent EFetch calls. Concurrently, the Parser Tab is activated. Within this tab, the parsing algorithm utilizes EFetch to process the data. It is configured to download the data in manageable batches of 50 records at a time, a strategy designed to minimize errors and maintain efficient operation with the NCBI server.

Between each EFetch call, XML files are downloaded to the temporary directory of R. These files are then parsed using the xmlEventParse function from the R package **XML** (Lang, 2022). This function employs an event-driven parsing method, which is notably memory-efficient. It allows for the extraction of specific tags and data from the XML files without the need to load the entire file into R's memory. This approach is particularly advantageous for handling large XML files, as it minimizes memory usage while efficiently retrieving the necessary information.

The parsing process utilizes specialized code to systematically read through the XML files. This code is adept at recognizing and extracting a broad range of specific information, including fields like Accession Numbers, Version, Address, Latitude-Longitude, Authors, Collected By, Create Date, Definition, Gene, gi, Haplotype, Host, Identified By, Note, Organelle, Organism, Product, Sequence, Sequence Length, Source, Specimen Voucher, Taxonomy, Database ID, Title, and Journal. Each field represents a unique aspect of the record, facilitating a thorough extraction of pertinent information for analysis. Users can customize the fields for parsing through the dropdown panel, which is accessible via the ☰ button prior to initiating the parsing process. Essential fields necessary for geocoding and phylogeographic analysis, such as Accession Numbers, Version, Address, Latitude-Longitude, Organism, Sequence, Sequence Length, and Taxonomy, are pre-selected and locked by default (Figure 5).



Parse XML files that have been retrieved from NCBI GenBank. When the ESearch tool finds UIDs, it posts them to the History server, enabling these UIDs to be directly used in subsequent EFetch calls. If any UIDs are returned from the search, the Parser page and/or the Parse button become available. The Parser tool internally utilizes EFetch. After parsing, the Feature table is made accessible. Features such as Accession Numbers, Version, Address, Latitude-Longitude, Organism, Sequence, Sequence Length, and Taxonomy are pre-selected by default and cannot be deselected. Although the 'Sequence' feature is present, it isn't visible in the Feature Table by default to ensure an optimal table layout. If the 'Sequence' is to be viewed within the Feature Table, the 'Column visibility' drop-down button should be used. For accuracy and to safeguard against potential interruptions, EFetch calls and parsing are performed on records in batches of 50 by default. Future updates will allow users to control this parameter.

Select feature(s) and Parse GenBank Records:

- ☑ Accession Numbers
- ☑ Version
- ☑ Address
- ☑ Lat Lon
- ☑ Authors
- ☑ Collected By
- ☑ Create Date
- ☑ Definition
- ☑ Gene
- ☐ gi
- ☑ Haplotype
- ☑ Host
- ☐ Identified By
- ☐ Note
- ☑ Organelle
- ☑ Organism
- ☑ Product
- ☑ Sequence
- ☑ Sequence Length
- ☐ Source
- ☐ Specimen Voucher
- ☑ Taxonomy
- ☐ Database ID
- ☐ Title
- ☑ Journal

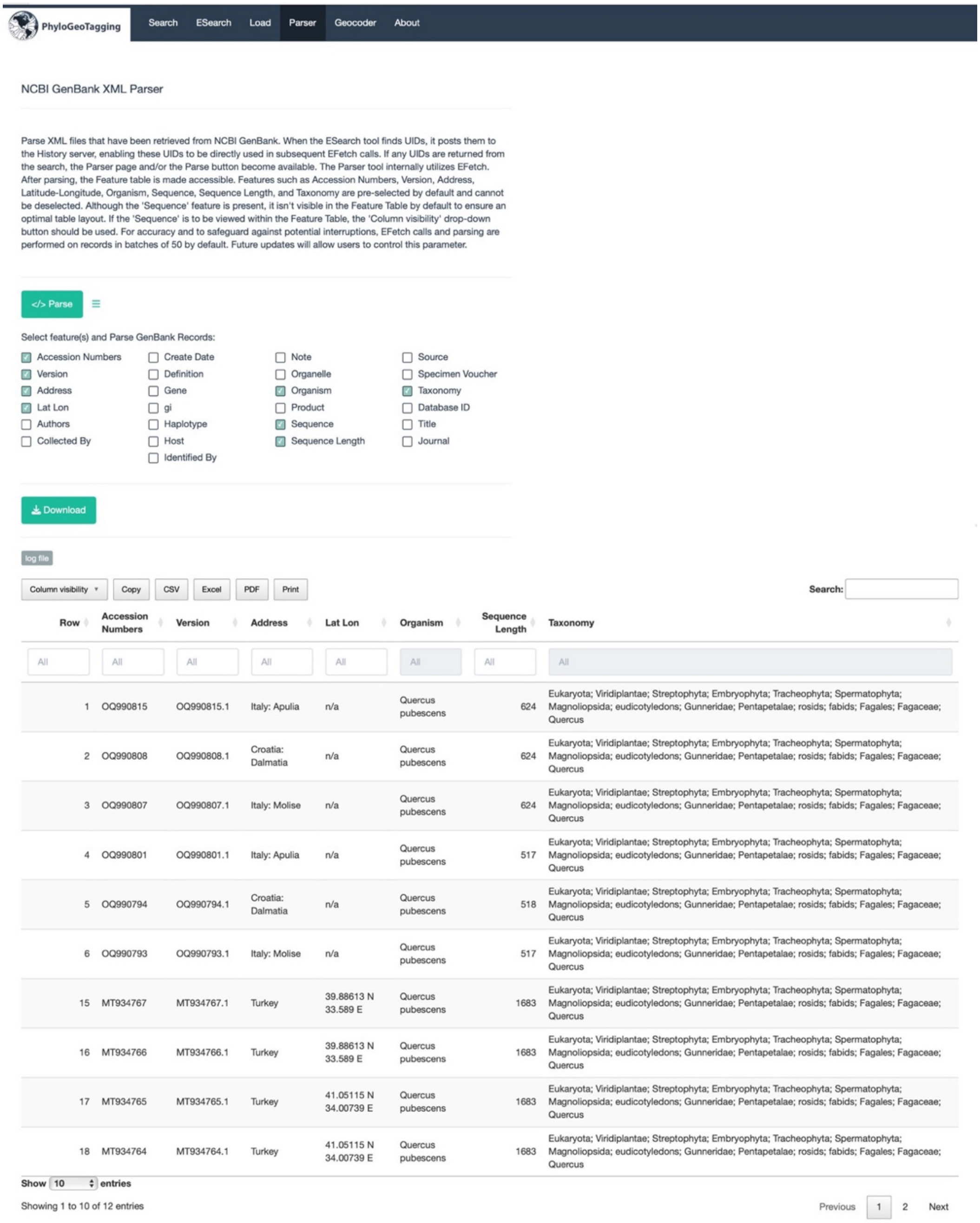
Select GenBank features before parsing, than click "</> Parse" button

Figure 5. Overview of the Parser Tab showing the dropdown panel with selectable fields for customizing the GenBank data parsing process.

The parsing process is powered by a specialized piece of code that systematically reads through the XML files. This code is structured to recognize and extract specific pieces of information. Each piece of data is meticulously identified and placed into a corresponding table, which will then be available for users to review and analyze. For instance, when the code encounters the term “GBSeq_locus”, it knows to extract and save that information under “Accession Numbers” in the data table. Similarly, it recognizes “GBSeq_length” as the cue to save sequence length details. For standard fields with unique identifiers, the algorithm straightforwardly matches the tag, like “GBSeq_locus”, with the corresponding table column, such as “Accession Numbers”. When it encounters “GBSeq_length”, it similarly extracts the data and places it under “Sequence Length” in the table. This methodical recognition ensures that each defined element, tagged distinctly within the XML, is accurately reflected in the parsed output. However, certain data points, such as geographic coordinates indicated by “lat_lon”, present a more complex challenge due to their non-unique or contextual tags within the XML hierarchy. The algorithm is crafted to handle such intricacies by employing a contextual tracking system. It does not solely rely on the current tag but also considers the preceding tags to determine the correct course of action. When the parser reads a “GBQualifier_name” tag and finds the content to be lat_lon, it then anticipates a “GBQualifier_value” tag to follow. Once this value tag is reached, the algorithm systematically records the geographic coordinates and includes them in the “Lat Lon” column of the dataset.

After parsing is complete, the resultant data is presented in the Feature table within the Parser Tab (Figure 6). This table is a core element of the application, organizing and displaying the parsed metadata in a structured and accessible format. Essential information such as Accession Numbers, Version, Address, Latitude-Longitude coordinates, Organism, Sequence Length, Taxonomy, and other customized features are systematically arranged for immediate review and analysis.

The Sequence data is inherently part of the parsed output, yet it is not visible by default in the Feature table to preserve the table's readability and to avoid overwhelming the page load. Users requiring detailed sequence information can easily make it visible through the “Column visibility” controls, thereby customizing the data presentation to suit their specific needs.



NCBI GenBank XML Parser

Parse XML files that have been retrieved from NCBI GenBank. When the ESearch tool finds UIDs, it posts them to the History server, enabling these UIDs to be directly used in subsequent EFetch calls. If any UIDs are returned from the search, the Parser page and/or the Parse button become available. The Parser tool internally utilizes EFetch. After parsing, the Feature table is made accessible. Features such as Accession Numbers, Version, Address, Latitude-Longitude, Organism, Sequence, Sequence Length, and Taxonomy are pre-selected by default and cannot be deselected. Although the 'Sequence' feature is present, it isn't visible in the Feature Table by default to ensure an optimal table layout. If the 'Sequence' is to be viewed within the Feature Table, the 'Column visibility' drop-down button should be used. For accuracy and to safeguard against potential interruptions, EFetch calls and parsing are performed on records in batches of 50 by default. Future updates will allow users to control this parameter.

Select feature(s) and Parse GenBank Records:

☑ Accession Numbers
☑ Version
☑ Address
☑ Lat Lon
☐ Authors
☐ Collected By
☐ Create Date
☐ Definition
☐ Gene
☐ gi
☐ Haplotype
☐ Host
☐ Identified By
☐ Note
☐ Organelle
☑ Organism
☐ Product
☑ Sequence
☑ Sequence Length
☐ Source
☐ Specimen Voucher
☑ Taxonomy
☐ Database ID
☐ Title
☐ Journal

Row	Accession Numbers	Version	Address	Lat Lon	Organism	Sequence Length	Taxonomy
All	All	All	All	All	All	All	All
1	OQ990815	OQ990815.1	Italy: Apulia	n/a	Quercus pubescens	624	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fagales; Fagaceae; Quercus
2	OQ990808	OQ990808.1	Croatia: Dalmatia	n/a	Quercus pubescens	624	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fagales; Fagaceae; Quercus
3	OQ990807	OQ990807.1	Italy: Molise	n/a	Quercus pubescens	624	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fagales; Fagaceae; Quercus
4	OQ990801	OQ990801.1	Italy: Apulia	n/a	Quercus pubescens	517	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fagales; Fagaceae; Quercus
5	OQ990794	OQ990794.1	Croatia: Dalmatia	n/a	Quercus pubescens	518	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fagales; Fagaceae; Quercus
6	OQ990793	OQ990793.1	Italy: Molise	n/a	Quercus pubescens	517	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fagales; Fagaceae; Quercus
15	MT934767	MT934767.1	Turkey	39.88613 N 33.589 E	Quercus pubescens	1683	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fagales; Fagaceae; Quercus
16	MT934766	MT934766.1	Turkey	39.88613 N 33.589 E	Quercus pubescens	1683	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fagales; Fagaceae; Quercus
17	MT934765	MT934765.1	Turkey	41.05115 N 34.00739 E	Quercus pubescens	1683	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fagales; Fagaceae; Quercus
18	MT934764	MT934764.1	Turkey	41.05115 N 34.00739 E	Quercus pubescens	1683	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fagales; Fagaceae; Quercus

Show 10 entries

Showing 1 to 10 of 12 entries

Figure 6. Illustration of the Parser interface, showcasing the Feature table.

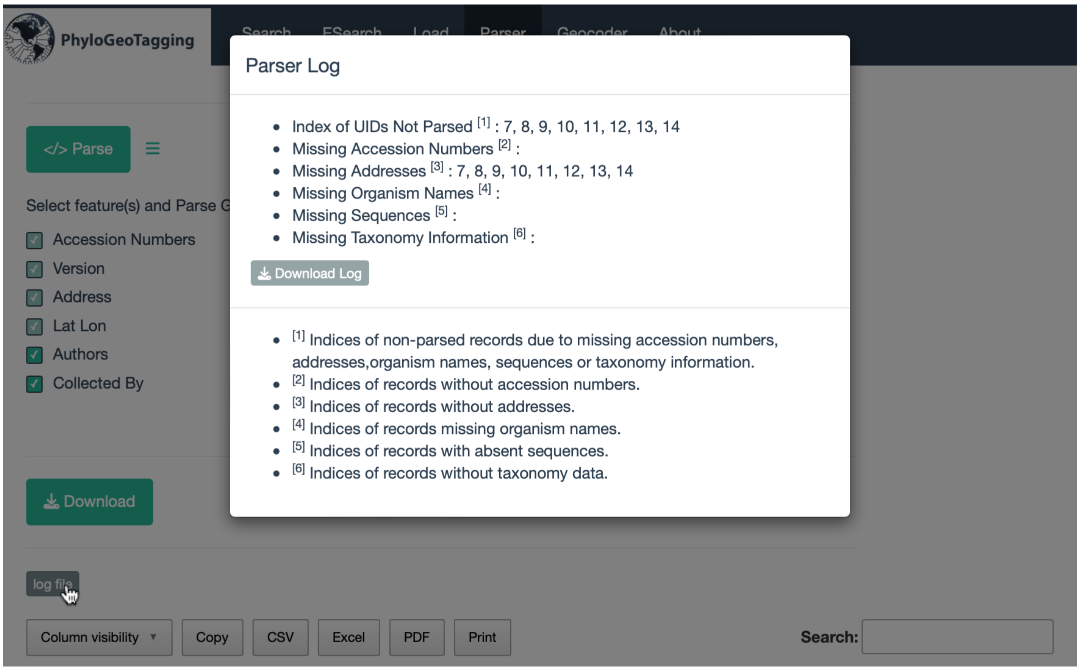
The table facilitates navigation through parsed records by presenting them in default sets of 10, while offering users the option to view 25, 50, 100, 1000, 10^4, 10^5, or 10^6 records at a time, thus enhancing the manageability of data review.

Individual columns in the table come with filters, allowing users to refine the data according to specific criteria. Most columns support multi-factor level filtering, while columns such as “Accession Numbers”, “Version”, “gi”, and “Sequence” allow for text input, offering precise control. Numerical columns such as “Row” and “Sequence Length” include sliders for range selection, enabling users to dynamically filter data by numerical values. The table also offers a keyword search functionality, conveniently located at the top-right, allowing for efficient data retrieval within the table itself. These filters are not only instrumental for on-the-spot analysis but also carry forward in the workflow, remaining active during subsequent geocoding processes. Users are advised to clear filters to revert to the full dataset before proceeding with further analysis steps within the site.

The table rendered using the R function renderDataTable from the **DT** package (Xie et al.,2023) offers several interactive options: users can copy the table data, save it in formats such as “csv”, “Excel”, or “Pdf”, or print it directly.

Above the Feature table, a download button activates once the table is populated. This allows users to download the feature table as a data.frame object, along with a log file of the parsing process. This feature is particularly valuable as it provides users with the ability to save their progress. The downloaded “.Rdata” file can later be reloaded via the Load tab, enabling users to resume their work seamlessly from the point of download. However, it's important to note that loading the downloaded “.Rdata” file is only available in the local version of the tool due to security reasons. While encryption might be implemented in the future to enhance security, tests during the development of the tool have shown that it adds a substantial amount of computation, which could significantly affect performance.

The Parser Log window serves as a crucial audit tool within the parsing process, ensuring transparency and accountability. It documents missing data instances in categories such as Accession Numbers, Addresses, Organism Names, Sequences, and Taxonomy Information. By listing the indices of records lacking this information, it aids in identifying and addressing gaps in the dataset, enhancing its robustness and reliability for analysis. This feature is not only valuable for troubleshooting but also for tracking the parsing history, as users can download the log for a detailed review. The log's diagnostic capability is integral to the process, highlighting incomplete records and the reasons for their exclusion, thus preserving the integrity and accuracy of the overall analysis. The log can also be downloaded as a text file (Figure 7).



Select feature(s) and Parse G

☑ Accession Numbers
☑ Version
☑ Address
☑ Lat Lon
☑ Authors
☑ Collected By

Parser Log

- Index of UIDs Not Parsed [1] : 7, 8, 9, 10, 11, 12, 13, 14
- Missing Accession Numbers [2] :
- Missing Addresses [3] : 7, 8, 9, 10, 11, 12, 13, 14
- Missing Organism Names [4] :
- Missing Sequences [5] :
- Missing Taxonomy Information [6] :

- [1] Indices of non-parsed records due to missing accession numbers, addresses,organism names, sequences or taxonomy information.
- [2] Indices of records without accession numbers.
- [3] Indices of records without addresses.
- [4] Indices of records missing organism names.
- [5] Indices of records with absent sequences.
- [6] Indices of records without taxonomy data.

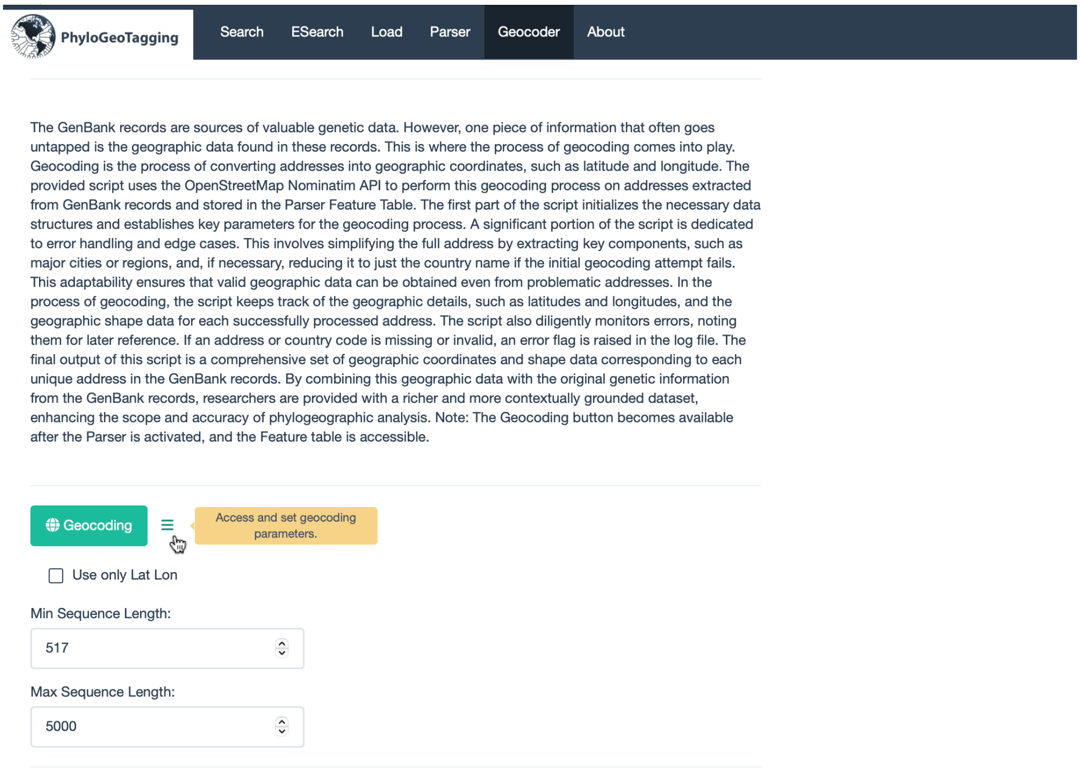
Figure 7. The Parser Log window displaying a record of excluded entries due to incomplete data, with options for downloading the detailed log.

2.4. Enhancing Geospatial Information – Geocoding of GenBank Data

The GenBank records are sources of valuable genetic data. However, one piece of information that often goes untapped is the geographic data found in these records.

This is where the process of geocoding comes into play. Geocoding is the process of converting addresses into geographic coordinates, such as latitude and longitude.

The Geocoding Tab in this site is designed to become available as a natural progression from the Parser Tab once the Parser Feature table is ready, guiding users through the intended workflow (Figure 8). Although this transition is automated to direct users efficiently, it may be unexpected for those who wish to remain on the Parser tab. If users need to review or continue working within the Parser tab, they can easily re-select it from the menu.



The GenBank records are sources of valuable genetic data. However, one piece of information that often goes untapped is the geographic data found in these records. This is where the process of geocoding comes into play. Geocoding is the process of converting addresses into geographic coordinates, such as latitude and longitude. The provided script uses the OpenStreetMap Nominatim API to perform this geocoding process on addresses extracted from GenBank records and stored in the Parser Feature Table. The first part of the script initializes the necessary data structures and establishes key parameters for the geocoding process. A significant portion of the script is dedicated to error handling and edge cases. This involves simplifying the full address by extracting key components, such as major cities or regions, and, if necessary, reducing it to just the country name if the initial geocoding attempt fails. This adaptability ensures that valid geographic data can be obtained even from problematic addresses. In the process of geocoding, the script keeps track of the geographic details, such as latitudes and longitudes, and the geographic shape data for each successfully processed address. The script also diligently monitors errors, noting them for later reference. If an address or country code is missing or invalid, an error flag is raised in the log file. The final output of this script is a comprehensive set of geographic coordinates and shape data corresponding to each unique address in the GenBank records. By combining this geographic data with the original genetic information from the GenBank records, researchers are provided with a richer and more contextually grounded dataset, enhancing the scope and accuracy of phylogeographic analysis. Note: The Geocoding button becomes available after the Parser is activated, and the Feature table is accessible.

☐ Use only Lat Lon

Min Sequence Length: 517

Max Sequence Length: 5000

Figure 8. Overview of the Geocoder Tab, available after Parser Feature table is ready. Several parameters can be adjusted using drop-down parameters panel.

The provided script uses the OpenStreetMap Nominatim API to perform this geocoding process on addresses extracted from GenBank records and stored in the Parser Feature table. OpenStreetMap Nominatim API is a popular, open-source geocoding tool that offers several advantages. It is freely accessible and provides global coverage, making it a great option for projects of any scale. Another notable benefit of this API is the absence of daily query limits, unlike other services. The data, sourced and updated by a worldwide community, tends to be diverse and frequently updated. This API also allows a high degree of customization for specific queries and is equally effective for reverse geocoding. However, one notable limitation is its strict usage policy, which includes restrictions on the number of requests per second to prevent server overload. This limitation makes the geocoding the slowest step in this package. However, the process is somewhat mitigated by only geocoding unique addresses found in parsing, which improves the overall time consumed in the geocoding process. To access the Nominatim API, the R package **nominatimlite** is utilized (Hernangómez 2023).

The first part of the script in this package written for geocoding initially encodes addresses for web compatibility, a necessary step given the constraints of URL formats which cannot include spaces or certain special characters. This encoding facilitates effective interaction with the geocoding API. To address the challenges of geocoding, the function is equipped with error-handling mechanisms. If an address does not yield results initially, it is modified to increase the likelihood of successful geocoding. This involves simplifying the address by extracting key components such as major cities or regions, and if necessary, reducing it to just the country name. This adaptability ensures that valid geographic data can be obtained even from problematic addresses. The core of the function involves interactions with a geocoding API. URLs are constructed for these API requests, incorporating the encoded address and various parameters, such as format specifications and limits. Separate API calls are made for point-of-interest (POI) data and shape data. POI data provides exact location coordinates, whereas shape data offers geographical boundary details. The responses from the API are then converted into an sf (simple features) format for analysis within R using the **sf** package (Pebesma, 2018).

In the final stage, geocoded data is extracted and stored. Successful geocoding results in the extraction of longitude and latitude coordinates from the POI data and the storage of shape data, which may include complex geographical boundaries. A progress indicator is included in the function, providing real-time updates on the geocoding process. In the event of errors, these are logged to enable the identification and potential rectification of issues with specific addresses.

The development of this geocoding function significantly advances the phylogeography tool set, as it adeptly converts textual addresses into valuable geographic data, enhancing the dataset by finding coordinates and cross-verifying already available latitude-longitude data with address information. The dual functionality of both finding coordinates and cross-verifying already available latitude-longitude data with address information significantly enhances the accuracy and reliability of the geographic data in the dataset. This aspect is particularly crucial for augmenting data in cases where latitude-longitude information is not provided in GenBank, thus ensuring a more comprehensive and reliable dataset for phylogeographic analysis. The function also incorporates sophisticated error handling mechanisms to accurately identify and resolve issues with address data. Data logging, while a more auxiliary feature, supports the process by ensuring transparency and traceability. Collectively, these capabilities are essential in phylogeography, enabling comprehensive studies into the geographic distribution and evolution of genetic traits by providing an enriched geographical context to genetic data.

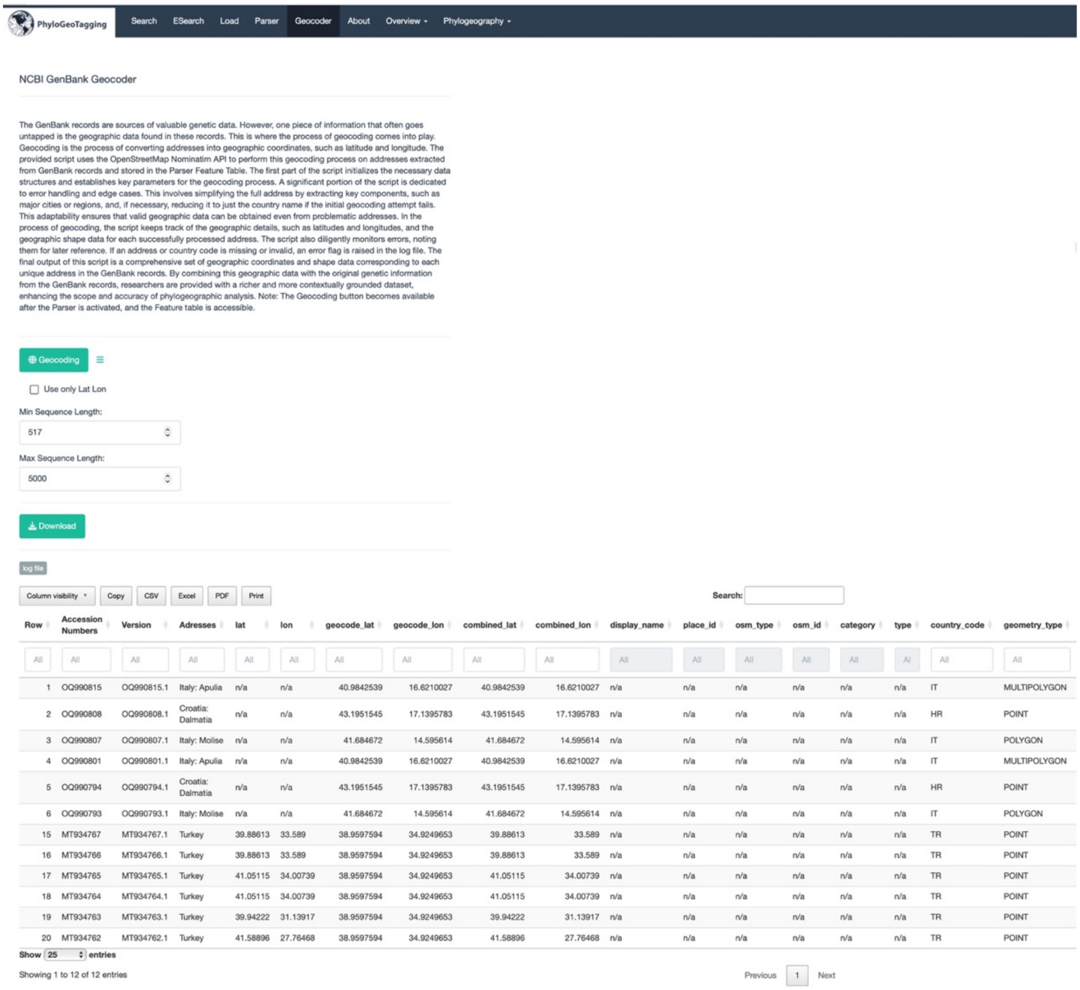
In the geocoding module of the application, users have access to several parameters that can be adjusted to tailor the geocoding process to their specific research needs. These parameters can be accessed via the ☰ button.

When the *Use only Lat Lon* checkbox is selected, the geocoding will be performed exclusively on records that have latitude and longitude (Lat Lon) information available from the Parser table. To ensure accuracy, this Lat Lon data is subjected to reverse geocoding, which checks if the country specified in the address aligns with the country deduced from the Lat Lon information. If there's a match, the Lat Lon data are deemed reliable and used for the coordinates in the combined latitude-longitude column of the resultant Geocoder table.

The *Min Sequence Length* parameter allows users to filter out sequences that are shorter than a specified value from the parsed results. This means that any genetic sequences in the dataset shorter than this minimum threshold will not be included in the geocoding analysis. The default setting for this parameter is determined by the shortest sequence returned during the parsing stage.

Conversely, the *Max Sequence Length* parameter enables the exclusion of sequences that exceed a certain length. The default maximum length is set to 5000, which serves to filter out complete genome sequences. This is based on the observation that most of the sequence data in GenBank are shorter than this limit. Including longer sequences often complicates and slows down the phylogeographic analysis due to the increased data volume and complexity present more challenges in processing and interpreting the information. However, if a research project requires longer sequences to be included, such as complete genomes, this parameter can be adjusted to fit the specific requirements. Together, these settings provide researchers with the flexibility to customize the geocoding process, ensuring that the geographic analysis is as relevant and efficient as possible for their particular line of inquiry.

Following the geocoding process, the Geocoder table becomes available, presenting a comprehensive set of data columns that integrate both the original and newly obtained geographic information (Figure 9). The table is designed to provide an at-a-glance overview of the geocoded data, with columns that serve specific functions and offer insights into the geocoding results.



NCBI GenBank Geocoder

The GenBank records are sources of valuable genetic data. However, one piece of information that often goes untapped is the geographic data found in these records. This is where the process of geocoding comes into play. Geocoding is the process of converting addresses into geographic coordinates, such as latitude and longitude. The provided script uses the OpenStreetMap Nominatim API to perform this geocoding process on addresses extracted from GenBank records and stored in the Parser Feature Table. The first part of the script initializes the necessary data structures and establishes key parameters for the geocoding process. A significant portion of the script is dedicated to error handling and edge cases. This involves simplifying the full address by extracting key components, such as major cities or regions, and, if necessary, reducing it to just the country name if the initial geocoding attempt fails. This adaptability ensures that valid geographic data can be obtained even from problematic addresses. In the process of geocoding, the script keeps track of the geographic details, such as latitudes and longitudes, and the geographic shape data for each successfully processed address. The script also diligently monitors errors, noting them for later reference. If an address or country code is missing or invalid, an error flag is raised in the log file. The final output of this script is a comprehensive set of geographic coordinates and shape data corresponding to each unique address in the GenBank records. By combining this geographic data with the original genetic information from the GenBank records, researchers are provided with a richer and more contextually grounded dataset, enhancing the scope and accuracy of phylogeographic analysis. Note: The Geocoding button becomes available after the Parser is activated, and the Feature table is accessible.

☐ Use only Lat Lon

Min Sequence Length: 517

Max Sequence Length: 5000

Search:

Row	Accession Numbers	Version	Adresses	lat	lon	geocode_lat	geocode_lon	combined_lat	combined_lon	display_name	place_id	osm_type	osm_id	category	type	country_code	geometry_type
All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	Al	All	All
1	OQ990815	OQ990815.1	Italy: Apulia	n/a	n/a	40.9842539	16.6210027	40.9842539	16.6210027	n/a	n/a	n/a	n/a	n/a	n/a	IT	MULTIPOLYGON
2	OQ990808	OQ990808.1	Croatia: Dalmatia	n/a	n/a	43.1951545	17.1395783	43.1951545	17.1395783	n/a	n/a	n/a	n/a	n/a	n/a	HR	POINT
3	OQ990807	OQ990807.1	Italy: Molise	n/a	n/a	41.684672	14.595614	41.684672	14.595614	n/a	n/a	n/a	n/a	n/a	n/a	IT	POLYGON
4	OQ990801	OQ990801.1	Italy: Apulia	n/a	n/a	40.9842539	16.6210027	40.9842539	16.6210027	n/a	n/a	n/a	n/a	n/a	n/a	IT	MULTIPOLYGON
5	OQ990794	OQ990794.1	Croatia: Dalmatia	n/a	n/a	43.1951545	17.1395783	43.1951545	17.1395783	n/a	n/a	n/a	n/a	n/a	n/a	HR	POINT
6	OQ990793	OQ990793.1	Italy: Molise	n/a	n/a	41.684672	14.595614	41.684672	14.595614	n/a	n/a	n/a	n/a	n/a	n/a	IT	POLYGON
15	MT934767	MT934767.1	Turkey	39.88613	33.589	38.9597594	34.9249653	39.88613	33.589	n/a	n/a	n/a	n/a	n/a	n/a	TR	POINT
16	MT934766	MT934766.1	Turkey	39.88613	33.589	38.9597594	34.9249653	39.88613	33.589	n/a	n/a	n/a	n/a	n/a	n/a	TR	POINT
17	MT934765	MT934765.1	Turkey	41.05115	34.00739	38.9597594	34.9249653	41.05115	34.00739	n/a	n/a	n/a	n/a	n/a	n/a	TR	POINT
18	MT934764	MT934764.1	Turkey	41.05115	34.00739	38.9597594	34.9249653	41.05115	34.00739	n/a	n/a	n/a	n/a	n/a	n/a	TR	POINT
19	MT934763	MT934763.1	Turkey	39.94222	31.13917	38.9597594	34.9249653	39.94222	31.13917	n/a	n/a	n/a	n/a	n/a	n/a	TR	POINT
20	MT934762	MT934762.1	Turkey	41.58896	27.76468	38.9597594	34.9249653	41.58896	27.76468	n/a	n/a	n/a	n/a	n/a	n/a	TR	POINT

Show 25 entries

Showing 1 to 12 of 12 entries

Figure 9. Overview of the Geocoder table, available after geocoding.

2.5. Overview of Geospatial Data - Visualization of GenBank

The Geocoder table is a central element of the **PhyloGeoTagging** package, serving as the primary interface for examining the final GenBank data. Its importance extends to the preparation of data for subsequent analysis, where it functions as an essential filter. However, understanding the specific properties of GenBank metadata and the spatial connections within the data requires more than tabular representations. To facilitate this, the package offers two graphical interfaces in the initial release of **PhyloGeoTagging**, under separate tabs accessible from the Overview menu item. These tabs become accessible once the Geocoder table is created.

2.5.1. Infographics

The Infographics Tab serves as a dynamic visual aid for interpreting the geospatial data derived from GenBank records. Upon the creation of the Geocoder table, users gain access to this tab, which offers an immediate graphical representation of the metadata specifics. This interface transforms complex datasets into engaging and comprehensible visual formats, such as donut charts, which are particularly useful for highlighting proportions and relationships within the data (Figure 10).

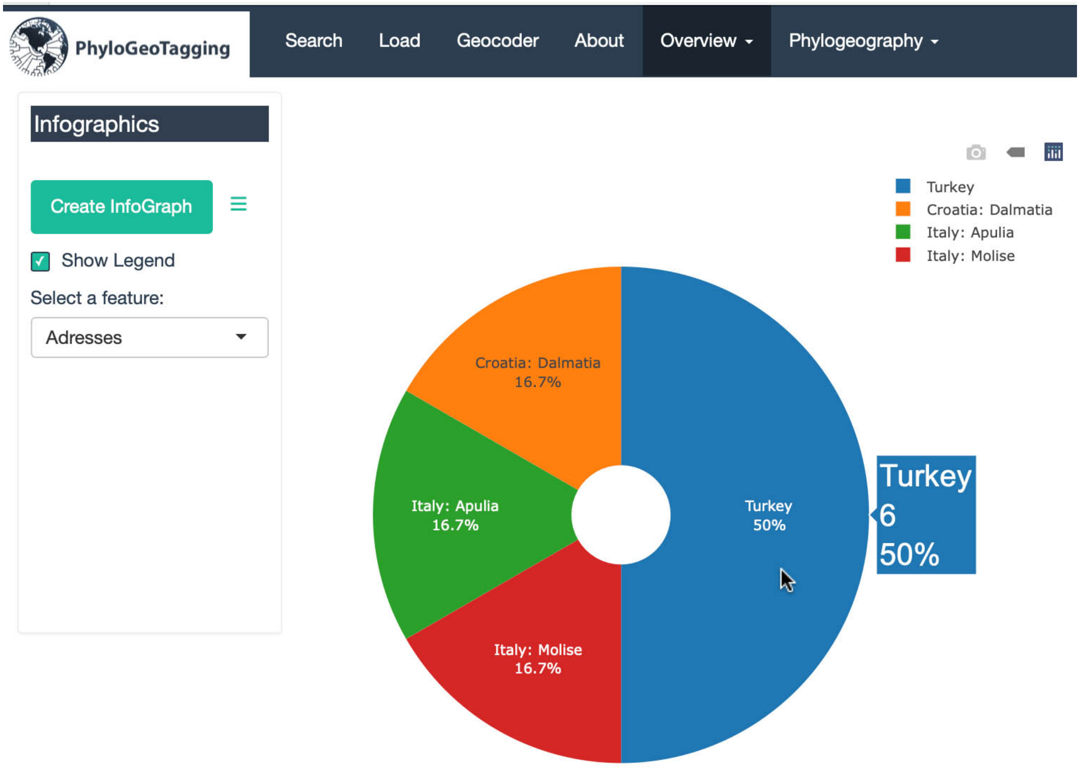
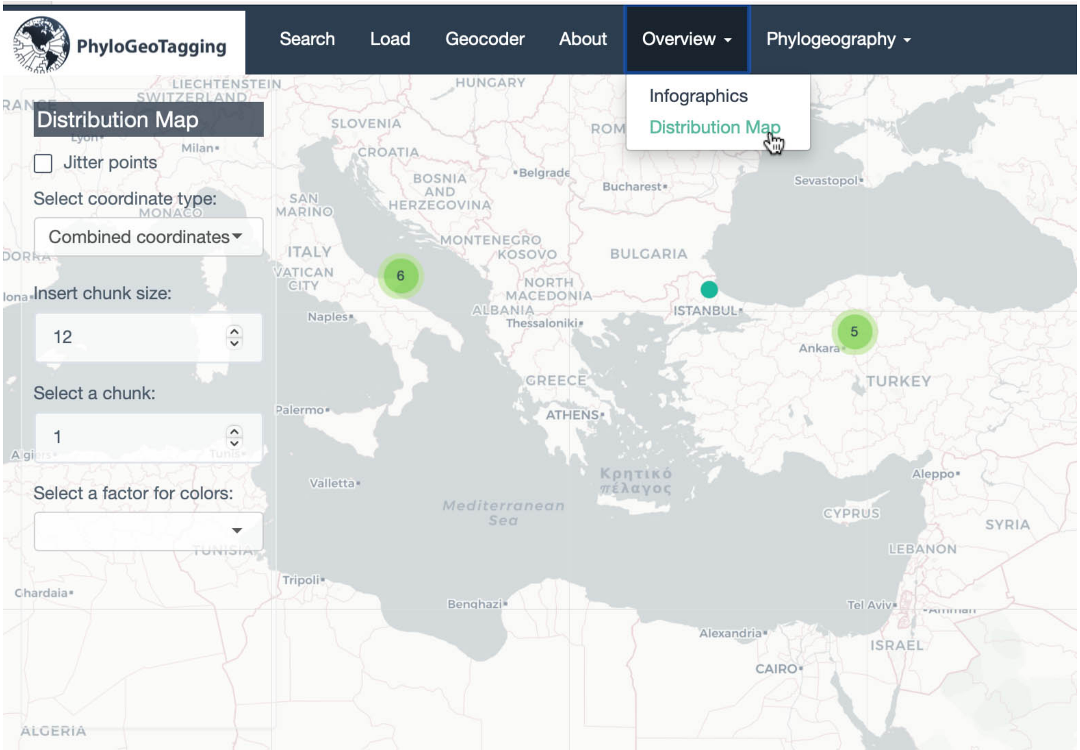


Figure 10. Donut chart in the Infographics Tab illustrating the distribution of a selected feature from the Geocoder table.

Within this tab, the “Create InfoGraph” button initiates the generation of the infographic, while users have access to several parameters via the ☰ button. Users can select specific data features to visualize from a dropdown menu, such as “Organism” (default), “Addresses”, or other relevant fields available in the Geocoder table. The *Show Legend* checkbox toggles the display of a legend, providing an explanatory key to understand the color-coded segments of the infographic. Each segment of the generated donut chart corresponds to a distinct category, with the size of the segment representing its relative proportion within the dataset. Hover labels display the category’s name, total number, and its percentage of the whole record set. This intuitive presentation allows for a quick assessment of the distribution and prevalence of different categories, facilitating insights into patterns that might not be readily apparent in a standard data table. The Infographics Tab is designed not only to provide statistical information but also to serve as an educational tool, enhancing the users' ability to discern and communicate complex data trends in a clear, visually appealing manner.

2.5.2. Distribution Map

The Distribution Map Tab is a sophisticated component of this application, integrating the metadata from the records with their geospatial distribution to create an interactive map using the R **leaflet** package (Cheng et al., 2023). Users can select the type of coordinates they wish to visualize from the control panel on the left, choosing from *Combined coordinates*, *Geocoded coordinates*, or *Lat Lon*, each connecting to the geospatial information available in the Geocoder table (Figure 11). For improved performance and visual clarity, especially with large datasets, markers on the map are clustered dynamically at specific zoom levels.



☐ Jitter points

Select coordinate type: Combined coordinates

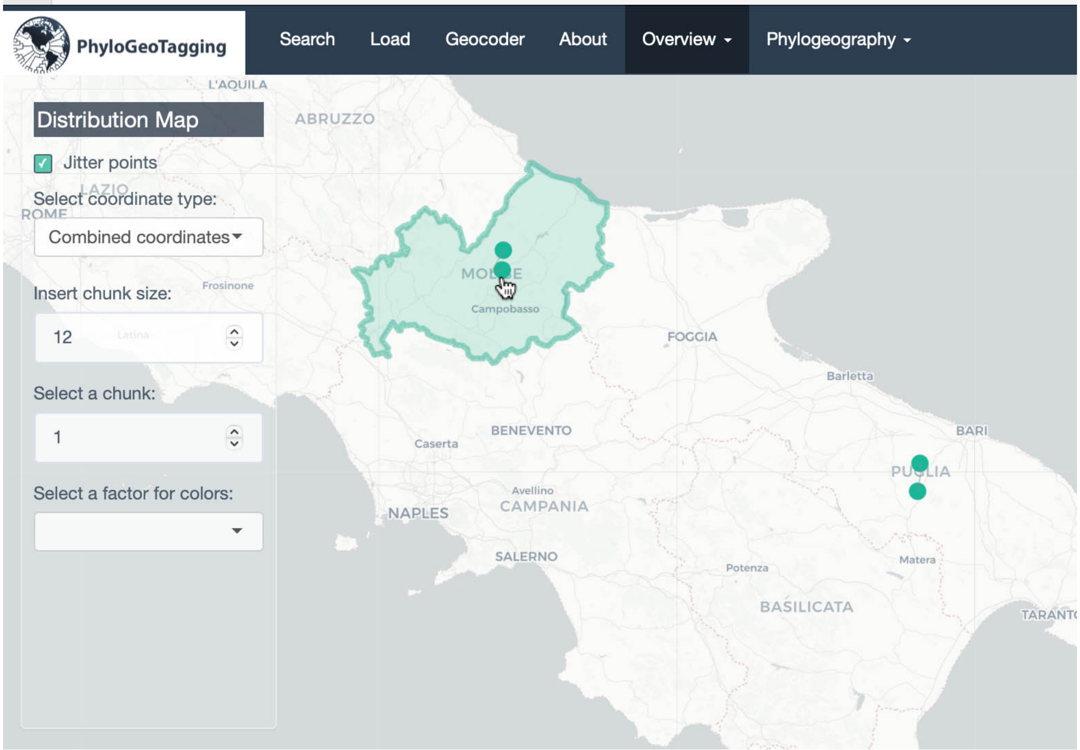
Insert chunk size: 12

Select a chunk: 1

Select a factor for colors:

Figure 11. Interactive map display on the “Distribution Map” Tab, illustrating the initial view with options for selecting coordinate types and viewing geospatial data from the Geocoder table.

When datasets contain overlapping points, the *Jitter points* option can be enabled to disperse markers for clearer visibility; this ensures that each data point is visible and distinct (Figure 12). Additionally, guiding lines are provided to indicate the original coordinates of these jittered markers, maintaining a reference to their actual locations. Mapping is performed in selectable-sized chunks, segmenting the visual rendering to ensure a seamless user experience with large datasets. Markers initially present in a uniform color, with the detailed data attributes obscured until the user selects a distinguishing factor for color-coding via the *Select a factor for colors* dropdown, bringing a spectrum of insights into view (illustrated in Figure 12). Clicking on a marker unfolds the shape files on the map, showcasing geometries such as “LINE”, “POLYGON”, or “MULTIPOLYGON” that result from the geocoding process, except for “POINT” geometries, which are represented as precise singularities on the map.



Distribution Map

☑ Jitter points

Select coordinate type: Combined coordinates

Insert chunk size: 12

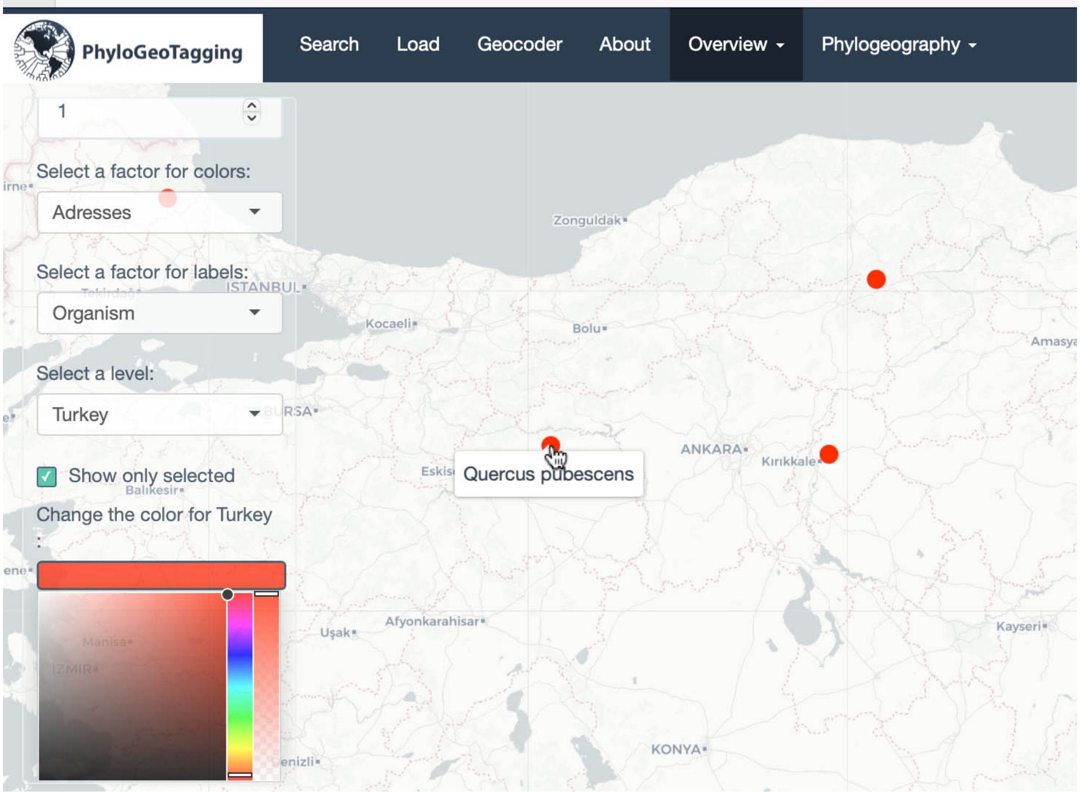
Select a chunk: 1

Select a factor for colors:

Figure 12. Jitter points feature on distribution map for overlapping markers and geometry viewing on click.

Users can select any factor from the Geocoder table to color-code the map markers using the *Select a factor for colors* control (Figure 13). Further customization includes the ability to select factors for both colors and hover labels separately using the *Select a factor for labels* control. This dual-selection process enables a layered visual inspection of the dataset's attributes, directly on the map. The *Select a level* control, or simply clicking on any marker of that level, along with the *Show only selected* checkbox, provides an option for isolating and examining specific data segments within larger datasets (as demonstrated in Figure 13). In the final step of customization, users can alter the color palette of the map through the color-picker.

This tool, intuitive by design, allows for a custom display of data layers, replacing the default rainbow colors with a chosen color that suits the user's visual preferences.



1

Select a factor for colors: Adresses

Select a factor for labels: Organism

Select a level: Turkey

☑ Show only selected

Change the color for Turkey :

Figure 13. Customization options on the distribution map, demonstrating the selection of factors for color-coding and labeling, and the color-picker tool.

2.6. GeoGenetic Synthesis - Phylogeography of GenBank Data

In the quest to unravel the complex web of life's ancestry, phylogeography stands as a pivotal discipline, bridging the spatial distribution of genetic lineages with their evolutionary history. It forms the core of this research due to its analytical nature, which bears significant weight. This section is dedicated to phylogeographic analysis using the comprehensive data from GenBank, representing the central piece that gives depth to the work presented. Traditionally, phylogeography has been challenging due to the vast amount of data and the complexity of the required analysis. With the development of this Shiny application, some of these challenges are addressed. The innovative Shiny application introduced in this thesis is designed to simplify the phylogeographic process. It acts as a gateway for exploring the phylogenetic data in GenBank with ease. The application features several important tabs: Cluster & Align, Haplotype ID (Identification), Haplotype Richness, ΦST Barriers, IBD (Isolation by Distance), AMOVA (Analysis of Molecular Variance), and Hapstep (Analysis of Diversity and Differentiation Parameters).

This package allows users to perform comprehensive phylogeographic analyses with the simplicity of a button click. Notably, some of the analysis results are displayed interactively on a map, providing an engaging and informative user experience. Its design for eventual web-based deployment is a notable feature, promising access from anywhere with internet connectivity. Nevertheless, this represents merely the initial phase of an ongoing development process. The field of software development is dynamic, and future updates promise to integrate new methods and enhance existing ones.

Prior to this thesis, I developed and published the R package **haplotypes** (Aktas C., 2020), which has since undergone significant development for this thesis work. I have rewritten key components and the entire package in the C language to enhance efficiency and upgraded the package to **haplotypes 2.0-beta**. Additionally, I implemented two new and prominent methods, each with their dedicated C functions: AMOVA (Analysis of Molecular Variance) and Hapstep (Analysis of Diversity and Differentiation Parameters). I am planning its inclusion in upcoming releases. This enhancement is poised to significantly accelerate analyses within the package, marking a major step forward in the efficiency and effectiveness of phylogeographic studies.

The following sections will delve into these enhancements, offering a comprehensive overview of each feature and function within the application and their impact on phylogeographic analysis.

Before delving into the details, it should be noted that in this thesis, the term “population” is employed in a broad sense to refer to geographical locations for the records. It is important to clarify that, due to the methodologies used for data collection and processing, particularly geocoding, the “populations” referenced may not always align with the traditional biological definition of a population. Instead, they may represent geographically approximated groups based on the data available. Consequently, in analyses such as AMOVA, the term “population differentiation” should be interpreted with this broader and more inclusive understanding. It implies differentiation among groups based on geographical coordinates, as indicated by the “combined_lat” and “combined_lon” columns in the Geocoder table.

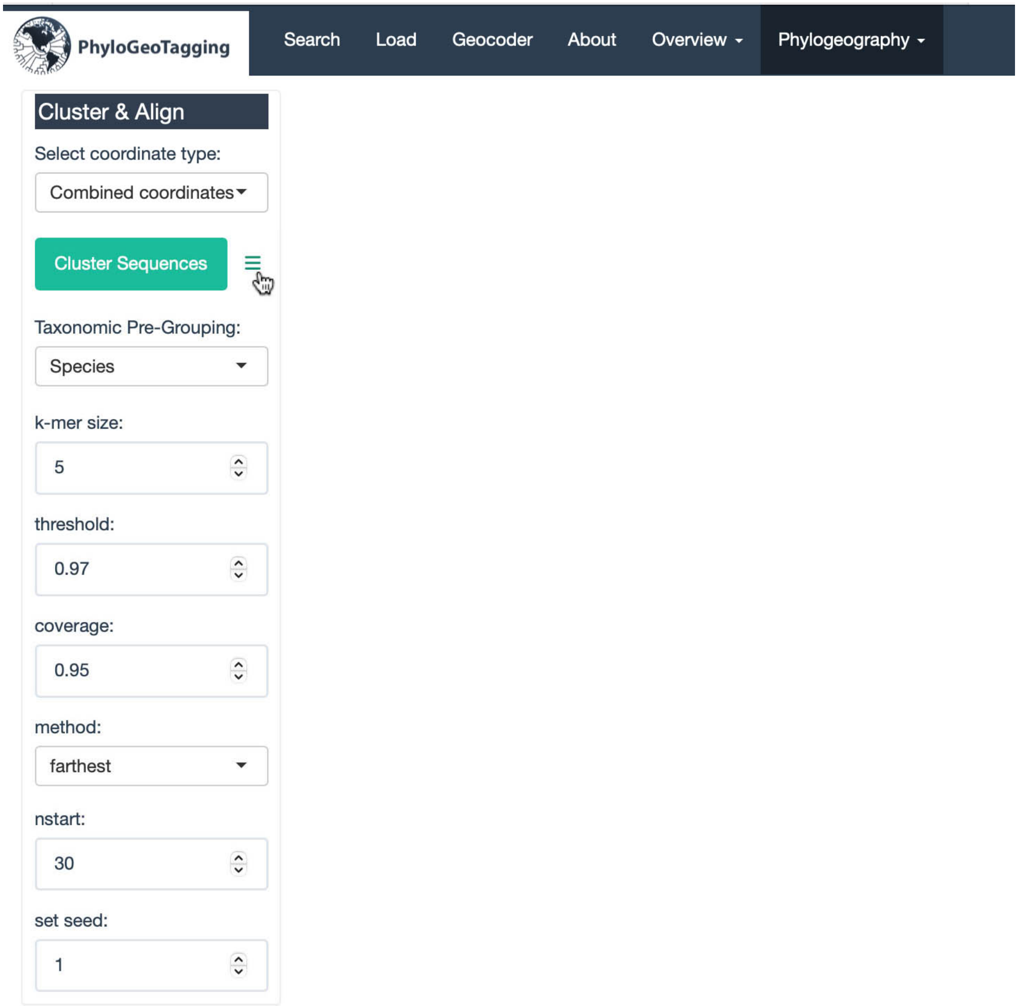
2.6.1 Cluster & Align

2.6.1.1. Cluster

Using vast amounts of sequence data in phylogeographic analysis represents a groundbreaking approach, made feasible by the package developed in this thesis. This package allows users to extract sequences from GenBank using a broad spectrum of search terms, including high-level taxonomic identifiers like a genus name, a family name, or even higher categories encompassing a wider array of biodiversity. Such searches often yield a highly diverse dataset, encompassing sequences from different lineages within the same family and sometimes even different genomic regions of the same species. To navigate this challenge, clustering becomes an indispensable step. It ensures that sequences originating from the same genomic regions or closely related taxa are methodically grouped together. This targeted grouping is critical for drawing meaningful comparisons, effectively aligning "apples with apples" rather than "apples with oranges." Additionally, clustering streamlines the subsequent sequence alignment by enhancing the similarity among sequences, thereby optimizing the alignment process. This strategic grouping achieves more than just simplifying the complexity of data; it narrows down the number of sequences to be aligned concurrently, which is a critical factor in improving both the quality and efficiency of multiple sequence alignment. Such meticulous organization not only makes the phylogeographic analysis more manageable but also elevates its precision and effectiveness, ensuring that the rich diversity of the data is harnessed without compromising the scientific integrity of the phylogeographic insights.

In this application, once the Geocoder table has been established, sequences, along with their associated metadata and geospatial attributes, are available for clustering. These sequences, representative of diverse genomic regions, are grouped based on sequence coverage and k-mer dissimilarity metrics, allowing for the identification of clusters that are phylogeographically comparable. The k-mer method involves divide DNA sequences into fragments of length-k. It calculates alignment-free distances between sequences by comparing the frequency and patterns of these k-mers. Similarities in k-mer composition provide a direct, computationally efficient measure of genetic distance without needing sequence alignment (Vinga & Almeida, 2003).

At the outset of the clustering process, users are presented with a control panel enabling the selection of coordinate types, such as *Combined coordinates*, *Geocoded coordinates*, or *Lat Lon*. These types are integrally connected to the spatial information stored in the Geocoder table (Figure 14).



Select coordinate type: Combined coordinates

Taxonomic Pre-Grouping: Species

k-mer size: 5

threshold: 0.97

coverage: 0.95

method: farthest

nstart: 30

set seed: 1

Figure 14. The Cluster & Align Tab, showcasing options for customizing clustering parameters.

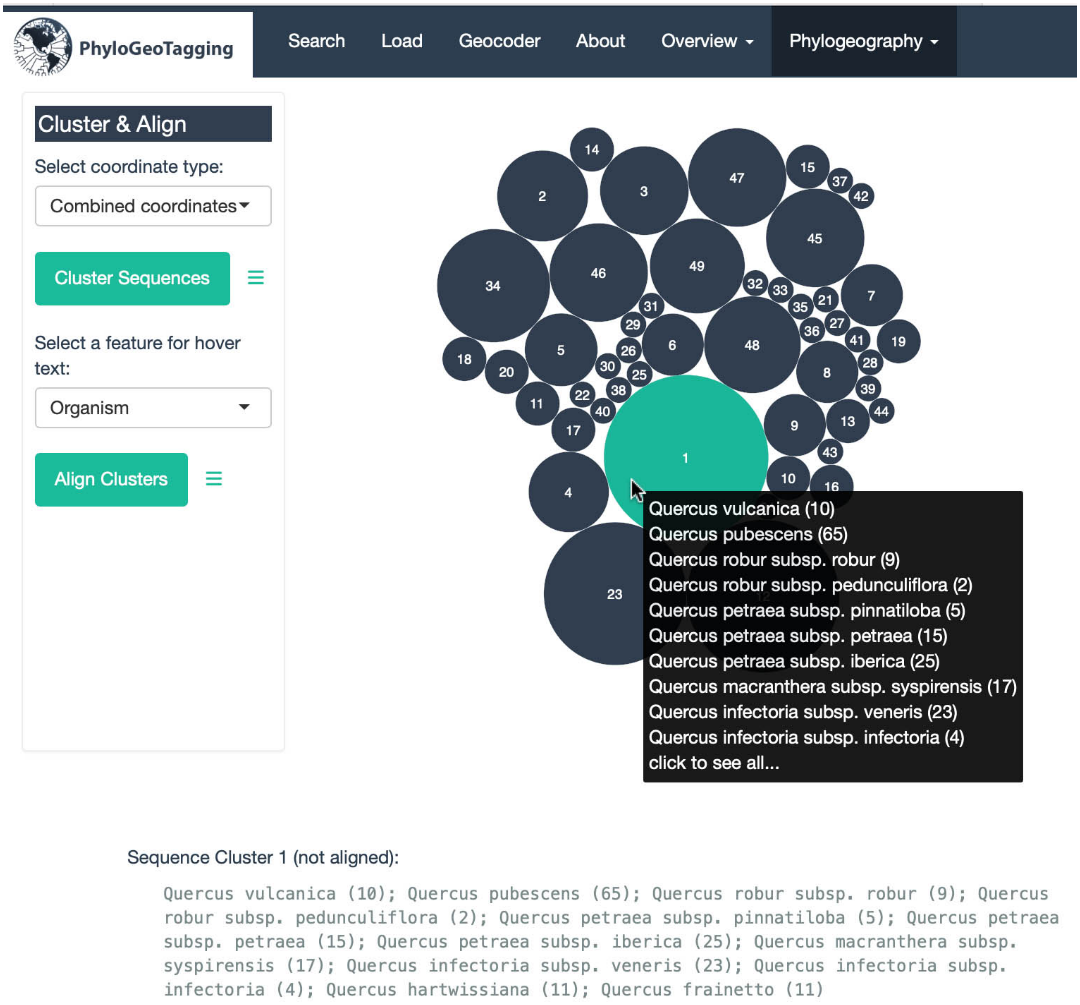
A sub-panel, accessed via the parameters button, allows for further customization of clustering parameters. The initial step in the clustering procedure in this package involves *Taxonomic Pre-Grouping*, where sequences are organized into preliminary groups based on taxonomic classification. When the *Unrestricted* option is chosen, sequences are grouped together without any predefined taxonomic constraints. This results in clusters that are diverse in their taxonomic origins, providing a panoramic view of the genetic landscape under investigation. Although such an approach is not recommended by default, it can be useful when exploring a wide range of genetic diversity between different taxa or when taxonomic classification is of secondary importance to the research objectives.

Selecting the *Genus* option refines the clustering process by only grouping together sequences that belong to the same genus. This restriction ensures that the resultant clusters are more genetically homogeneous, reflecting closer evolutionary relationships. This level of specificity is particularly advantageous when the focus of the study is to discern phylogenetic relations, especially phylogeographic patterns within a group of taxa that show putative interspecific hybridizations.

Opting for the *Species* level further narrows down the clustering, wherein only sequences that originate from the same species are grouped together. This stringent criterion is particularly valuable for investigations aimed at understanding intra-species genetic diversity and elucidating direct phylogeographic patterns. This approach essentially isolates the dimension of geographic differentiation as the primary variable, under the theoretical assumption that within a species, genetic isolation is minimized due to the potential for gene flow. Consequently, the phylogeographic patterns that emerge are more reflective of geographical influences rather than broader phylogenetic relationships. This methodology is crucial in filtering phylogeographic signals from mixed genetic data, accentuating the role of spatial distribution in shaping genetic variation. The most restrictive option, *Exact Match* demands that sequences be grouped only if they correspond exactly to the same name identified in the "Organism" feature of the Geocoder table. This level of precision results in highly specific clusters, each representing a unique name entity. These are mostly taxonomic names, but sometimes specimen names or codes may also be part of the Organism entries. This approach is invaluable in studies requiring detailed genetic comparison and contrast within very narrowly defined taxonomic boundaries, such as below species ranks including sub-species and varieties.

Users can customize the clustering process to meet their specific research needs by adjusting the *threshold* parameter for k-mer similarity and the *coverage* parameter for minimum coverage percentage, thereby controlling the tightness or looseness of clustering.

Initially, sequence lengths are utilized in a hierarchical clustering process to segment sequence records. The *coverage* parameter is instrumental in this phase, allowing users to define the acceptable range of sequence lengths that must match for sequences to be clustered together. When the number of sequences exceeds a specific limit, hierarchical clustering becomes impractical due to memory constraints. In such instances, initial clusters are created by sorting sequence lengths and applying a coverage-based similarity threshold. This initial segmentation can significantly accelerate the subsequent hierarchical clustering of selected sequences based on k-mer counts, which employs sequential k-means partitioning. Internally, the otu function from the R **kmer** package (Wilkinson, 2018) is used for this latter approach, accepting parameters such as k-mer size and a similarity cutoff threshold. The *set seed* parameter initializes the random number generator to produce the same results, thus ensuring reproducibility of the data clustering when the analysis is repeated with the same dataset. Upon completion of the clustering process, the results are visually depicted through interactive circular plots generated by the R **ggirafe** package (Gohe & Skintzos, 2023). This visual representation serves as an immediate, interpretable summary of the clustering outcomes and provides an interactive element. Users can explore cluster details on demand from the control panel using the *Select a feature for hover text* option. By hovering over or clicking on the circles, users can access a wealth of information about each cluster, including unique levels and the frequency of their occurrence (Figure 15).



Select coordinate type: Combined coordinates

Select a feature for hover text: Organism

Quercus vulcanica (10)
Quercus pubescens (65)
Quercus robur subsp. robur (9)
Quercus robur subsp. pedunculiflora (2)
Quercus petraea subsp. pinnatiloba (5)
Quercus petraea subsp. petraea (15)
Quercus petraea subsp. iberica (25)
Quercus macranthera subsp. syspirensis (17)
Quercus infectoria subsp. veneris (23)
Quercus infectoria subsp. infectoria (4)
click to see all...

Sequence Cluster 1 (not aligned):

```
Quercus vulcanica (10); Quercus pubescens (65); Quercus robur subsp. robur (9); Quercus
robur subsp. pedunculiflora (2); Quercus petraea subsp. pinnatiloba (5); Quercus petraea
subsp. petraea (15); Quercus petraea subsp. iberica (25); Quercus macranthera subsp.
syspirensis (17); Quercus infectoria subsp. veneris (23); Quercus infectoria subsp.
infectoria (4); Quercus hartwissiana (11); Quercus frainetto (11)
```

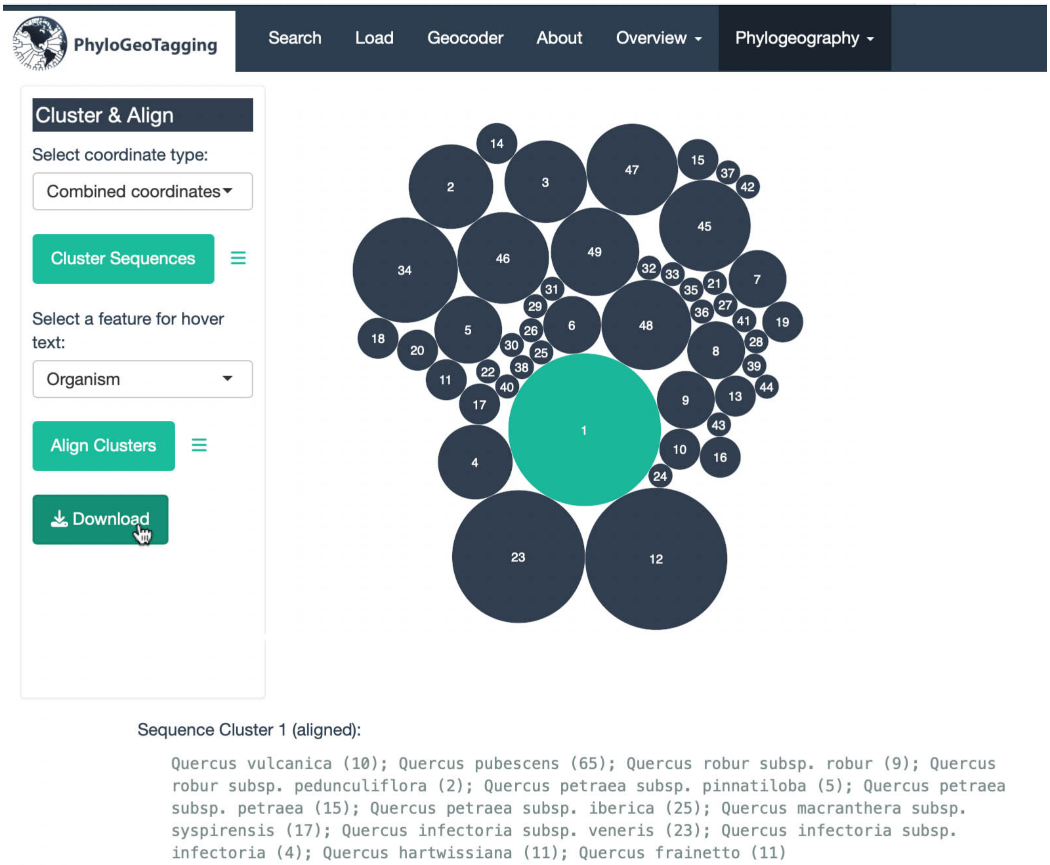
Figure 15. Interactive circular plot displaying clustered sequence data, with a highlighted example showing detailed taxonomic information upon user interaction.

2.6.1.2. Align

Upon clustering the sequences, the subsequent step in the analysis involves applying multiple sequence alignment (MSA) to each cluster. In this initial version of the package, due to stability issues with the R **msa** package (Bodenhofer et al., 2015), the alignment parameters are not user-configurable and are temporarily set to default values. This default setting employs the ClustalW algorithm, a widely used method in bioinformatics for aligning DNA or protein sequences.

Interactive features are embedded within the interface to enhance user engagement and data interpretation. By interacting with the circular plot, specifically clicking on the clusters, users trigger a footnote text area that provides comprehensive details for the selected cluster. The display is designed to show only the unique taxonomic levels, with the frequency of each level indicated in parentheses. Additionally, the alignment status of sequences is conveniently displayed in the header of this footnote area, providing a clear indication of the alignment progress or completion (Figure 16).

Once the alignment is complete, the “Download” button is enabled, allowing users to download the aligned sequences along with the Geocoder table. The downloaded .Rdata file carries an internal “CA” flag, signifying that it contains Cluster & Align data. This flag enables users to upload previously downloaded .Rdata files through the Load Tab and resume their analysis from the alignment step, including geocoder table and clusters, facilitating a seamless continuation of their work.



Select coordinate type: Combined coordinates

Select a feature for hover text: Organism

Sequence Cluster 1 (aligned):

```
Quercus vulcanica (10); Quercus pubescens (65); Quercus robur subsp. robur (9); Quercus
robur subsp. pedunculiflora (2); Quercus petraea subsp. pinnatiloba (5); Quercus petraea
subsp. petraea (15); Quercus petraea subsp. iberica (25); Quercus macranthera subsp.
syspirensis (17); Quercus infectoria subsp. veneris (23); Quercus infectoria subsp.
infectoria (4); Quercus hartwissiana (11); Quercus frainetto (11)
```

Figure 16. Interactive circular plot displaying clustered and aligned sequence data, with a highlighted example showing detailed taxonomic information upon user interaction, alignment status shown.

2.6.2. Haplotype Identification

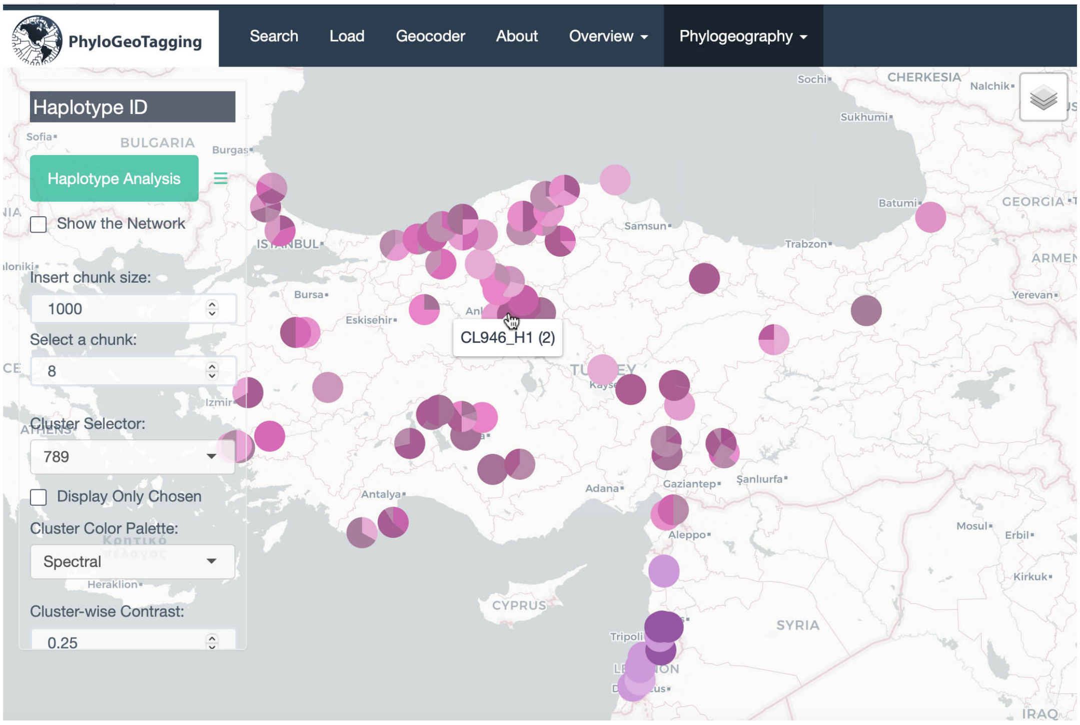
The journey through phylogeographic analysis within this package begins with Haplotype Identification, accessible via the Haplotype ID tab. This function is available after aligned sequences are prepared or loaded in the Cluster & Align tab, ensuring they are ready for in-depth analysis. At the core of this method is the collapsing of identical sequences into unique haplotypes within each cluster. This process is carried out by the haplotype function from the R package **haplotypes 2.0-beta**, which identifies haplotypes by evaluating the absolute pairwise genetic distance between sequences.

Further enhancing haplotype identification, the application also constructs statistical parsimony networks. These networks are created using the parsimnet function from the R package **haplotypes 2.0-beta** and are visually rendered with the **visNetwork** package (Almende et al., 2022). This approach represents genetic linkages and displays mutational steps between haplotypes within each cluster. Notably, the *prob* argument within parsimnet is disabled, allowing for unlimited connectivity of haplotypes and ensuring a single, comprehensive network.

Before running the analysis, parameters can be adjusted using control panel. One such parameter is the coding of indels (insertions or deletions), essential for haplotype identification and parsimony network construction. Users have the option to choose from *Simple Indel Coding,* where gaps are treated as a missing character and coded separately following the simple indel coding method (Simmons & Ochoterena, 2000); *5th State,* which treats gaps as a fifth state character; or *Missing,* where gaps are considered as missing data.

To ensure reproducibility, the *set seed* parameter guarantees that the random number generator produces consistent results for the same dataset across different runs, maintaining the integrity of the parsimony network topology for each cluster.

Upon executing the "Haplotype Analysis" under the Haplotype ID tab, users are presented with a graphical output that shares similarities with the Distribution Map tab. This interface can be thought of as a haplotype distribution map, where pie charts represent haplotype frequencies within each geographic location (Figure 17). Similar to the Distribution Map, markers are clustered above a certain zoom level to optimize performance and visual clarity. For large datasets, the clusters are segmented into chunks, allowing for a smoother rendering of the map elements, and ensuring a seamless user experience. Interactivity is a key feature of this interface. Clicking on the pie charts on the map or nodes on the parsimony network reveals detailed information about the selected item, enriching the user's understanding of the data. Additionally, the *Show the Network* checkbox controls the visibility of the "Parsimony Network" panel, located to the right of the window.



☐ Show the Network

Insert chunk size: 1000

Select a chunk: 8

Cluster Selector: 789

☐ Display Only Chosen

Cluster Color Palette: Spectral

Cluster-wise Contrast: 0.25

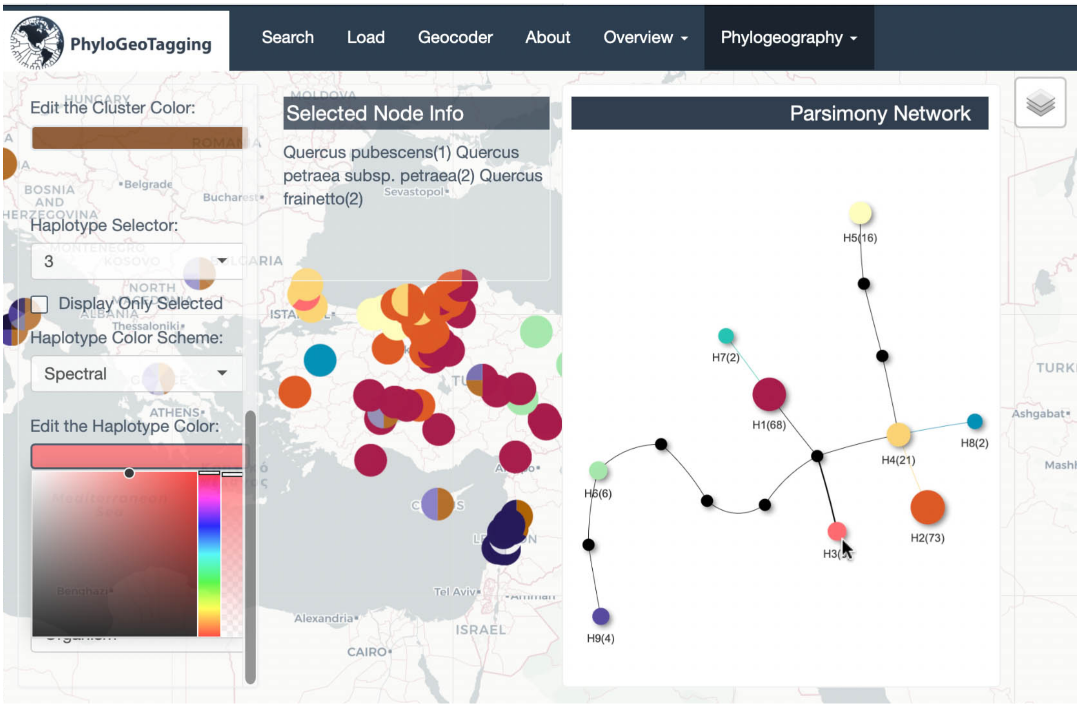
Figure 17. Interactive map with pie-chart markers indicating haplotype distribution across geographic locations, with options to adjust cluster visibility and color settings.

The control panel includes a *Cluster Selector* for choosing specific clusters, aided visually by color codes. Directly beneath this, the *Display Only Chosen* checkbox enables users to isolate and display only the selected cluster, hiding the others. The cluster-wise color control elements, *Cluster Color Palette*, *Cluster-wise Contrast*, and *Cluster-wise Saturation* provide control over the colors of all clusters simultaneously. These elements are essential for distinguishing between clusters. The color-picker, *Edit the Cluster Color*, enables the customization of pie chart color within a selected cluster. An internal algorithm initially assigns different contrast and saturation values to the main hue within a cluster, offer nuanced visual differentiation of haplotypes based on genetic distances. This can be adjusted using the *Cluster-wise Contrast* and *Cluster-wise Saturation* parameters.

Below the cluster sub-pane, control elements for haplotypes are available. The *Haplotype Selector* lets users choose a single haplotype from the selected cluster. The *Display Only Selected* feature isolates the pie charts of the selected haplotypes for focused analysis. The *Haplotype Color Scheme* changes the palette for haplotypes within the selected cluster, enhancing differentiation. Additionally, the *Edit the Haplotype Color* feature, with its dedicated color-picker, allows for the customization of the selected haplotype color (Figure 18).

At the bottom of the control panel, the *Piechart Opacity* feature enables adjustment of the pie charts' transparency, which is helpful when overlapping pie-charts are present on the map.

Finally, *Hover Labels* can be selected from the features available in the Geocoder table, corresponding to the coordinates types used in the Cluster & Align process. This feature, by default, displays the clusters and haplotypes codes, offering a quick reference to the underlying data.



Haplotype Selector: 3

☐ Display Only Selected

Haplotype Color Scheme: Spectral

Selected Node Info

Quercus pubescens(1) Quercus petraea subsp. petraea(2) Quercus frainetto(2)

Figure 18. Detailed view of the parsimony network and distribution map highlighting the genetic relationships among haplotypes, with user options for haplotype color customization and network display.

2.6.3. Haplotype Richness

The study of genetic diversity within and across populations is pivotal in evolutionary biology, ecology, and conservation genetics. A common measure employed in these studies is haplotype richness, which represents the number of unique genetic variants (haplotypes) within a population. Given the potential disparities in sampling efforts across different populations or clusters, a method that accounts for these differences is crucial for accurate representation of haplotype diversity. To accurately gauge this richness in studies where sampling effort varies, I introduce here the weighted haplotype richness calculation.

In the Weighted Haplotype Richness analysis, haplotype richness is calculated for each cluster by grouping the data by cluster and country features. This grouping process allows for the determination of the number of distinct haplotypes per group, thus quantifying the haplotype richness within that specific geographical and genetic context. The culmination of this analysis is the calculation of weighted haplotype richness, a process that adjusts the haplotype richness according to the size of each cluster. This adjustment averages the values for country i and is captured in the formula:

$$\mathrm{WHR}_i = \frac{\sum(n_k \times S_k)}{\sum S_k}$$

where WHR_i is the Weighted Haplotype Richness for country i. n_k is the unique number of haplotypes in cluster k. S_k is the size of cluster k.

This method places more significance on larger clusters to provide a more accurate representation of genetic diversity, thereby reduce the potential bias that could be caused by smaller, potentially under sampled clusters. This adjustment is crucial in studies where the availability of genetic data is uneven across different geographical locations or genetic clusters. However, it is essential to apply this method judiciously, considering its assumptions and potential limitations, especially if the larger clusters represent a more limited genetic scope due to other factors such as sampling bias or biological replicates from GenBank, which may represent a smaller number of haplotypes.

Focusing on a more detailed examination of genetic uniqueness, I introduce the “Cross-Country Weighted Haplotype Endemism” analysis, which centers on the endemism of haplotypes across various countries. This analysis specifically highlights haplotypes that are endemic, or unique, to each country within a particular cluster, considering the cluster's distribution across multiple countries.

The calculation method weights the endemic haplotypes by considering the frequency of the cluster's presence in each country relative to its global distribution.

The formula represents the calculation for the Cross-Country Weighted Haplotype Endemism for each country i within each cluster k, and total value for a country i is given as:

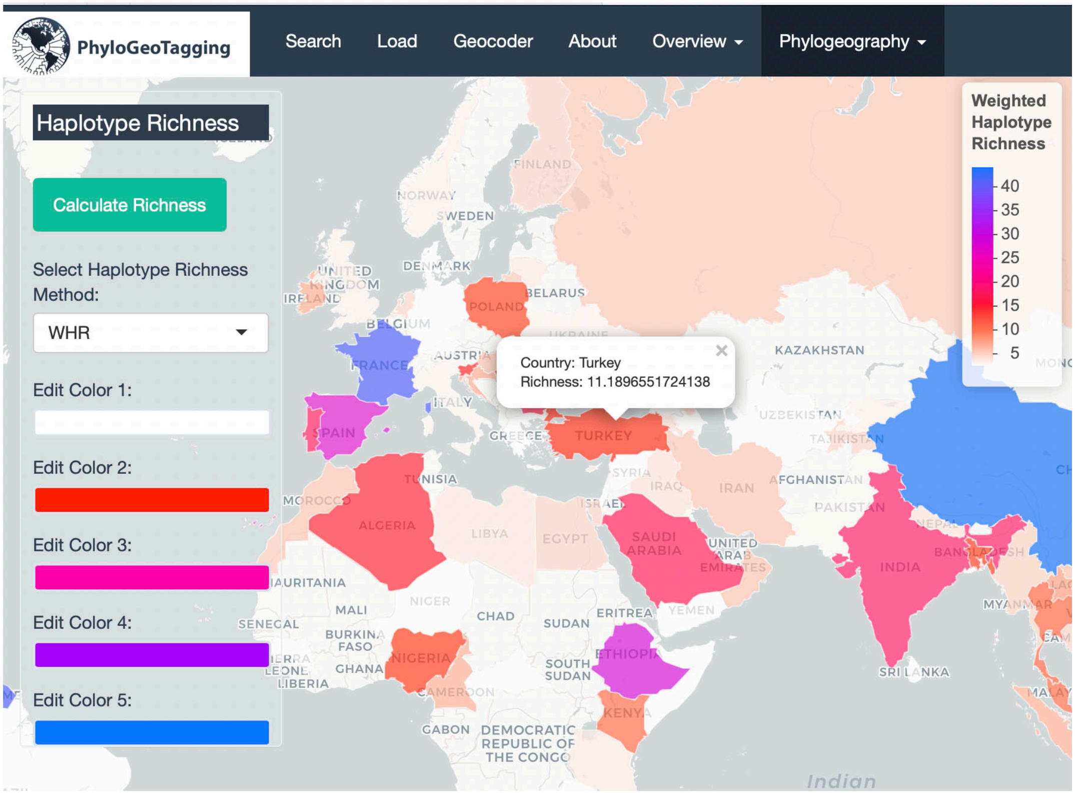
$$CCWHE_{i,k} = \sum_{j \in C} \left(\frac{E_{i,j,k}}{N_k} \right) \times \left(1 - \frac{S_{i,k}}{\sum_{j \in C} S_{j,k}} \right)$$

$$\text{Total } CCWHE_i = \sum_{k \in K} CCWHE_{i,k}$$

where $CCWHE_{i,k}$ is the Cross-Country Weighted Haplotype Endemism for country i and cluster k. C is the set of all countries, and j is indexes these countries. $E_{i,j,k}$ is the number of unique endemic haplotypes in country i for cluster k. N_k is the total number of sequences in cluster k. $S_{i,k}$ is the number of sequences of cluster k in country i, and $S_{j,k}$ represents the number of sequences of cluster k in any other country j. Total $CCWHE_i$ is the aggregate Cross-Country Weighted Haplotype Endemism for country i across all clusters K, which is the set of all clusters.

In the Cross-Country Weighted Haplotype Endemism analysis, haplotypes unique to specific countries within each cluster are identified and counted. This count is then adjusted based on the ratio of the number of sequences of that cluster found in a specific country to the total number of sequences in the cluster. These adjusted values are calculated separately for each cluster in every country. In this process, the endemism count for clusters sampled from only one country is weighted by a factor of zero. This is particularly necessity for clusters only sampled in one country, as it’s difficult to ascertain if these haplotypes are truly endemic. Consequently, the contribution of endemic haplotypes identified in studies spanning multiple countries is emphasized. Moreover, by weighting the relative frequencies of cluster sequences in each country, the rate of endemism is further incorporated into this initial weighting. This means that if a smaller number of sequences in a country compared to other countries yields a high number of endemic haplotypes, their impact is amplified by this initial weighting. Finally, similar to the Weighted Haplotype Richness calculation, these values are aggregated across clusters, weighted by the size of each cluster, to ensure a balanced representation of endemic haplotypes, unaffected by smaller or less-sampled clusters.

Under the "Haplotype Richness" tab, users can visualize the Weighted Haplotype Richness by selecting *WHR* and Cross-Country Weighted Haplotype Endemism by selecting *CCWHE* across countries on an interactive map. The map provides a color-coded representation of richness levels, with the ability to customize the color scheme according to five different user-specified colors. These colors correspond to different ranges of values. By clicking on a country, users can access the specific richness value for that location, engaging directly with the data to extract detailed insights (Figure 19).



Haplotype Richness

Select Haplotype Richness Method: WHR

Country: Turkey
Richness: 11.1896551724138

Figure 19. The color-coded map that illustrates the weighted haplotype richness across countries available.

2.6.4. ΦST Barriers

In the context of AMOVA (Analysis of Molecular Variance), the fixation index Φ_{ST} is specifically utilized to quantify genetic variation and differentiation among populations (Excoffier et al., 1992). In contrast to F_{ST} , a commonly used metric in population genetics based on haplotype (allele) frequencies alone, Φ_{ST} provides a more comprehensive view by incorporating genetic distances between haplotypes, thereby offering deeper insights into genetic differentiation.

Mapping pairwise Φ_{ST} values provides valuable insights into the genetic differentiation distributed across landscapes. The Φ_{ST} statistic is central to this analysis, measuring genetic variance between populations relative to the total genetic variance, thereby quantifies population differentiation. Pairwise Φ_{ST} values are calculated for each population pair within clusters and assigned to the midpoint coordinates between these populations, illustrating potential genetic barriers on the map.

Pairwise Φ_{ST} is calculated between every pair of populations using pairPhiST function from my package **haplotypes 2.0-beta**, which internally utilizes the AMOVA framework.

An intriguing aspect of the approach presented here is the use of both distance weighting (DW) and inverse distance weighting (IDW) for the Φ_{ST} values. The IDW method helps mitigate the effect of isolation by distance—a common phenomenon in population genetics where populations that are geographically closer tend to be more genetically similar than those further apart. By weighting the pairwise Φ_{ST} values based on geographical proximity, IDW ensures that interpretations of genetic differentiation extend beyond the mere effect of physical distance. This way, potential local barriers, where genetic gradients evident, can be detected. Scaling the geographic distances between populations relative to a maximum possible distance (derived from the Mercator projection limits), standardizes the normalization of geographic distance, allowing for reproducible application, regardless of the dataset's actual geographical extent.

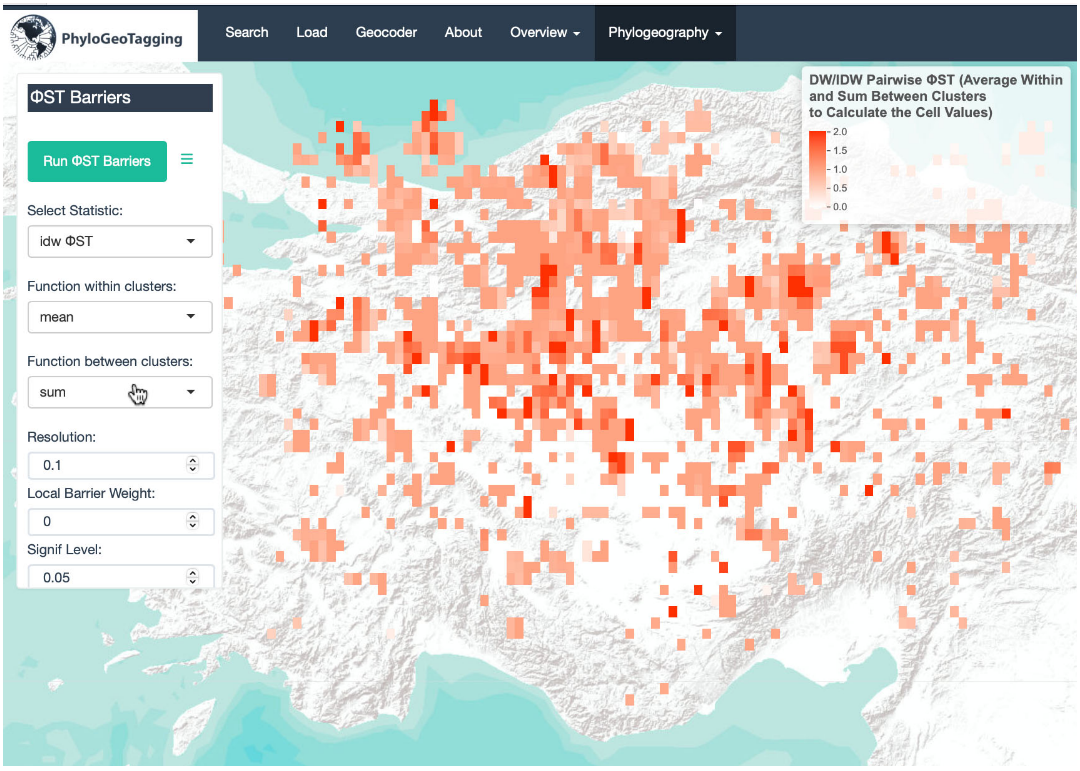
Innovatively, this tool allows the user to choose a negative value for the decay parameter, applying distance weighting (DW) to pairwise Φ_{ST} values and amplifying the influence of more distant populations in the calculation of the Φ_{ST} statistic. Unlike conventional IDW, where weight decreases with increasing distance, emphasizing closer populations, a negative decay value reverses this weighting mechanism. More distant populations are given greater weight, potentially uncovering genetic differentiation patterns that standard analyses might miss. This can be particularly useful, especially when data points congregate in certain local areas with high density, potentially overshadowing broader spatial patterns. However, results should be interpreted with caution, and the DW approach should be considered a supplementary tool, offering an alternate lens for spatial genetic data analysis, particularly where long-range genetic connections are of specific interest.

Under the ΦST Barriers tab, the tool provides an interactive platform for visualizing genetic differentiation as a raster map. Before commencing the analysis, users can adjust various parameters through the control panel. This includes the option to select indel coding methods, essential for the analysis, which are detailed in section 2.6.2 Haplotype Identification. Following the analysis, the main statistic to be displayed on the map can be chosen from the control panel, with options like *Geo-Dist*, indicating the geographic distance between pairwise populations, *Pairwise ΦST*, *Pairwise ΦST p-value*, and *IDW Pairwise ΦST* values (Figure 20).

The selected statistic undergoes a two-stage rasterization process on the map. In the first stage, within-cluster aggregation, the Φ_{ST} statistic is aggregated within each cluster for each raster cell using statistical methods like minimum, maximum, mean, or sum. This aggregation, determined by the *Function Within Clusters* parameter, is critical for understanding genetic structures within clusters and their spatial variation. The second stage, Between-Cluster Aggregation for Raster Cells, involves aggregating these within-cluster Φ_{ST} values across different clusters for each raster cell. This stage, guided by the *Function Between Clusters* parameter, is crucial for identifying shared phylogeographic patterns among lineages, potentially highlighting common phylogeographic barriers across taxa.

Users can adjust the resolution of the raster map using the *Resolution* parameter. The *Local Barrier Weight* settings, controlling the decay parameter in this modified IDW function, which is used in the package, allow for the input of negative values for DW or positive values for IDW, thereby influencing how the Φ_{ST} values are weighted.

Raster cells with non-significant Φ_{ST} values can be filtered out through the *Signif Level* parameter, enhancing the map's relevance and accuracy. Additionally, users can hide or filter cells by adjusting the final cell value cutoff using the *Cell Base Value* parameter.



Select Statistic: idw ΦST

Function within clusters: mean

Function between clusters: sum

Resolution: 0.1

Local Barrier Weight: 0

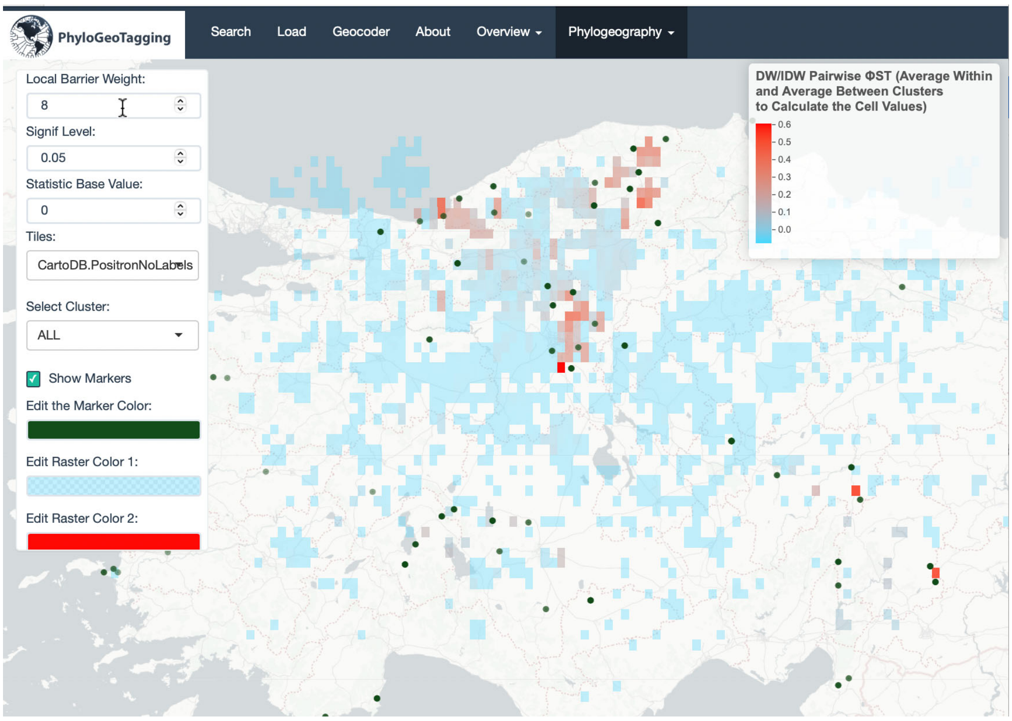
Signif Level: 0.05

Figure 20. An interactive raster map from the ΦST Barriers tab, showcasing the spatial distribution of the selected statistic. Color intensities vary to represent the differing values of the chosen statistic across the landscape.

The map's appearance can be further customized. The *Tiles* parameter in the tool allows for changes in the map's background tiles, offering various visual styles for the base map. This feature enables users to customize the map's appearance according to their preferences or the specific requirements of their analysis. The *Show Markers* checkbox enables the visualization of the original locations of selected clusters, which are pivotal in calculating the midpoints for the analysis (Figure 21).

Additionally, the *Select Cluster* parameter provides the functionality to isolate a specific cluster for focused analysis. By default, this parameter is set to display all clusters, offering an overview of the entire dataset. However, users can adjust it to concentrate on a cluster of interest, allowing for more detailed examination of specific areas or patterns. It's important to note that this analytical approach is applied only to clusters containing more than one population. Therefore, not all clusters may be available in the drop-down list.

Finally, two color picker controls, *Edit Raster Color 1* and *Edit Raster Color 2*, allow users to manipulate the color gradient, providing a personalized visual experience.


PhyloGeoTagging
Search
Load
Geocoder
About
Overview
Phylogeography
DW/IDW Pairwise ΦST (Average Within and Average Between Clusters to Calculate the Cell Values)
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Local Barrier Weight: 8

Signif Level: 0.05

Statistic Base Value: 0

Tiles: CartoDB.PositronNoLabels

Select Cluster: ALL

☑ Show Markers

Edit the Marker Color:

Edit Raster Color 1:

Edit Raster Color 2:

Figure 21. The ΦST Barriers visualization emphasizes local barriers using adjusted weight settings. The map is enhanced with a selected tile set for clarity, includes markers indicating the origins of barriers, and offers color and opacity customization for nuanced visual interpretation.

2.6.5. Isolation by Distance (IBD)

Isolation by Distance (IBD) is an essential principle in population genetics which posits that geographically proximate populations are more likely to exhibit genetic similarity compared to those that are more distantly spaced. This pattern emerges due to the restricted movement of individuals across space, leading to a decline in gene flow with increasing geographic distance. Consequently, gene exchange tends to be more frequent among populations that are closer together, tapering off as the distance grows.

A direct method to analyze Isolation by Distance (IBD) involves plotting genetic distances against geographic distances for pairs of populations. Two widely-used metrics for assessing genetic divergence are Φ_{ST} and Nei's Raw Distance D (Nei & Li, 1979). Nei's D is chosen for this analysis owing to its sensitivity to variations in haplotype (allele) frequencies and its capacity to provide an unbounded measure of genetic distance between populations, which is particularly beneficial when comparing highly divergent populations. It also avoids the potential confounding effects of within-population diversity that can affect other measures. Nei's Raw Distance, denoted as D, is calculated using the formula:

$$D = \sum_{i=1}^{k} \sum_{j=1}^{k'} x_{1i} x_{2j} \delta_{ij}$$

where k and k' represent the number of distinct haplotypes in the respective populations, x_{1i} and x_{2j} are the frequencies of the ith and jth haplotypes in populations 1 and 2, and δ_{ij} represents the mutational distance between haplotype i and haplotype j. This measure includes nucleotide substitutions as well as insertions and deletions, based on the indel coding method chosen. The pairwise Nei distances are calculated employing the pairnei function from the R package **haplotypes 2.0-beta**.

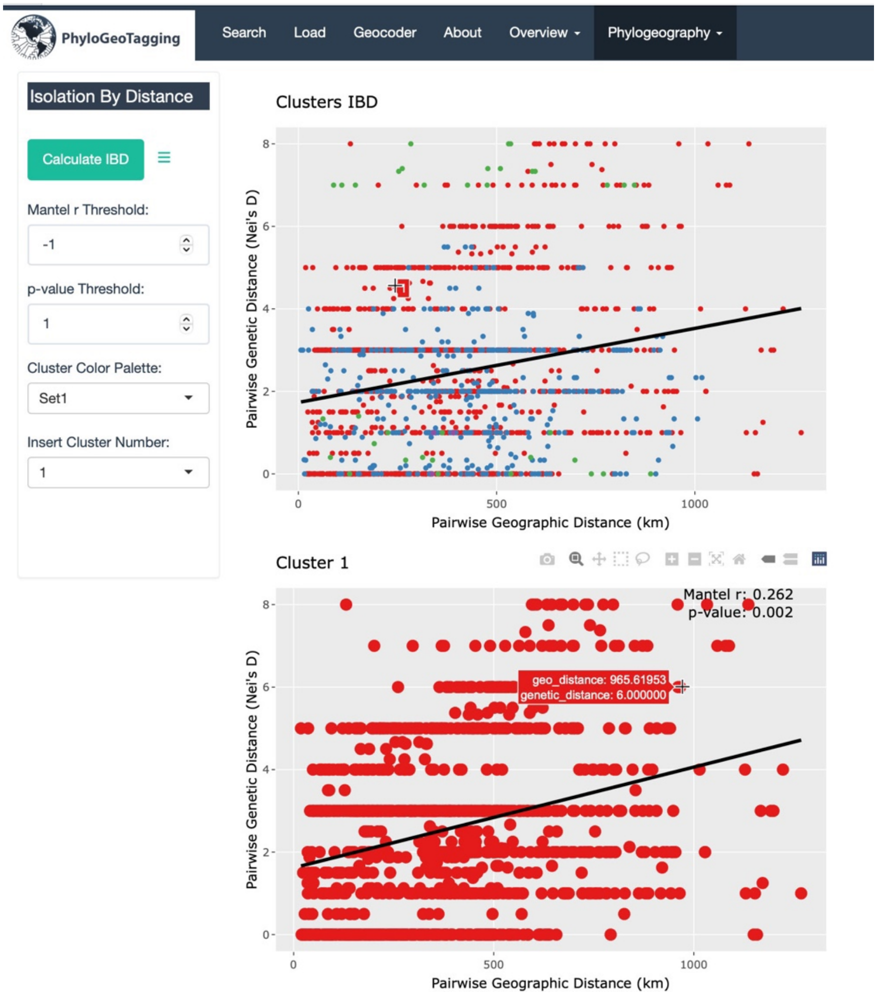
Geographic distances, measured in kilometers, are computed as geodetic distances using the st_distance function from the R package **sf**, ensuring precise and accurate spatial measurements between population locations.

To discern the correlation between genetic and geographic distances, the Mantel test (Mantel, 1967) is applied using Mantel function from R package **vegan** (Oksanen et al., 2022). This non-parametric statistical method evaluates the relationship between genetic and geographic distance matrices in IBD analyses. Its primary advantage lies in handling non-normal data distributions and managing spatial autocorrelation, which is often present in geographic data. This robustness makes the Mantel test particularly suitable for assessing the correlation between genetic similarity and geographic proximity, a crucial aspect of validating the IBD hypothesis. By quantifying this correlation, the Mantel test serves as an essential tool in elucidating the spatial patterns of genetic diversity.

Upon accessing the IBD section, users can initiate the analysis by clicking the "Calculate IBD" button (Figure 22). This action triggers the computation of pairwise genetic distances and their correlation with pairwise geographic distances. As in previous tabs, users have the option to select indel coding methods before running the analysis, as detailed in section 2.6.2. The control panel offers various post-analysis customization options for the output. The results of the IBD analysis are displayed in scatter plots, where each point represents the genetic distance versus the geographic distance for a pair of populations. The top plot provides an overall view with all clusters combined, while the bottom plot offers a detailed view for a specific cluster, chosen by *Insert Cluster Number* parameter.

Regression lines in these plots illustrate the data trends. Key Mantel test statistics, including the r correlation coefficient and its associated p-value, are displayed on the second plot. These statistics are crucial indicators of the strength and significance of the isolation by distance relationship. Users can interact with the plots by hovering over data points to reveal cluster information and graphic values. The *Mantel r*

Threshold parameter on the control panel allows users to set a cut-off value for the Mantel test, ranging from -1 to 1, to filter results based on the correlation coefficient. This feature is useful for identifying clusters with high levels of IBD. Additionally, the *p-value Threshold* parameter, ranging from 0 to 1, enables users to set a threshold for statistical significance, determining the validity of the correlations. Finally, the *Cluster Color Palette* settings allow users to select a color scheme for the data points, enhancing the visual distinction of clusters on the IBD plot.



Mantel r Threshold: -1

p-value Threshold: 1

Cluster Color Palette: Set1

Insert Cluster Number: 1

Mantel r: 0.262
p-value: 0.002

geo_distance: 965.61953
genetic_distance: 6.000000

Figure 22. Interactive IBD plot displaying the genetic versus geographic distances among clusters. Hovering over points reveals cluster information and distance metrics, while the control panel offers options for statistical thresholding and visual customization.

2.6.6. Analysis of Molecular Variance (AMOVA)

Analysis of Molecular Variance (AMOVA) is a statistical method employed in molecular systematics to quantify and partition genetic variation across different levels of spatial factors, such as populations, regions, or taxonomic factors including species, genera, and so on (Excoffier et al., 1992). AMOVA stands as one of the most robust tools in phylogeography for assessing genetic differentiation between geographically distinct populations. Following its primary application in population differentiation, AMOVA also plays an essential role in molecular systematics, particularly in defining taxonomic boundaries. The method elucidates the genetic structure of species, revealing the extent of genetic variation within species compared to that between them. This is crucial for understanding the genetic relationships and diversity patterns defining species and their subgroups. AMOVA's ability to dissect these relationships and patterns further underscores its power in both phylogeography and taxonomy.

In AMOVA, fixation indices denoted by Φ are used to measure the genetic variation linked to distinct levels of a given factor, offering a deeper understanding of genetic divergence within that context.

For a single factor, the sole index calculated Φ_{ST} , is formulated as:

$$\Phi_{ST} = \frac{\sigma^2_{among}}{\sum \sigma^2},$$

When multiple hierarchical factors are considered, such as species within genera, Φ_{SC} quantifies the variation between species within genera:

$$\Phi_{SC} = \frac{\sigma^2_{amongspecieswithingenera}}{\sigma^2_{amongspecieswithingenera} + \sigma^2_{withinspecies}}$$

For differentiation among genera, the index Φ_{CT} is applied:

$$\Phi_{CT} = \frac{\sigma^2_{amonggenera}}{\sum \sigma^2}$$

The Φ_{ST} statistic, when multiple hierarchical factors are present, provides a summary statistic for the overall genetic differentiation, is calculated as follows:

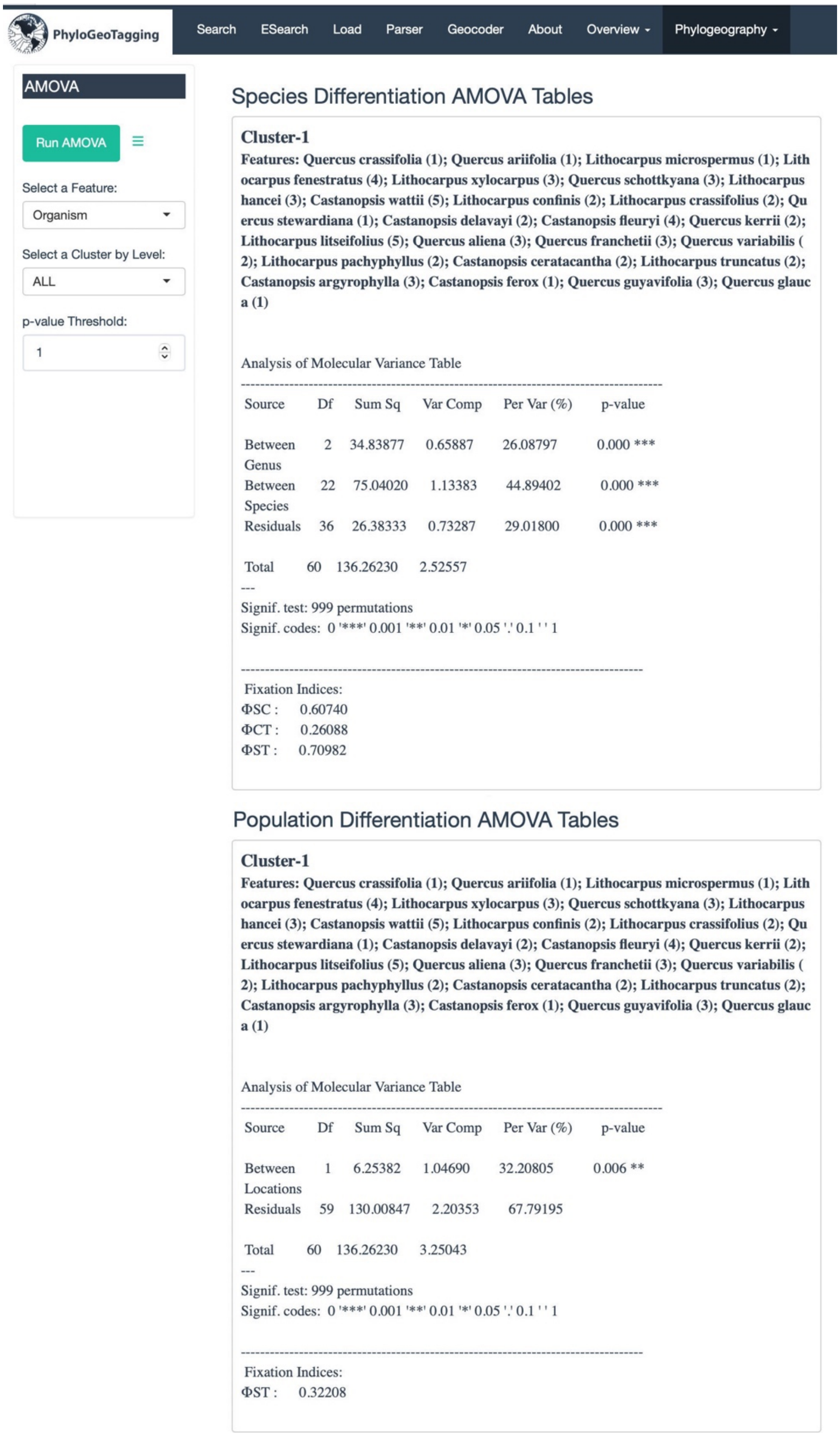
$$\Phi_{ST} = \frac{\sigma^2_{amonggenera} + \sigma^2_{amongspecieswithingenera}}{\sum \sigma^2}$$

Here, σ^2 represents the variance component associated with the level of differentiation being measured. A high Φ value suggests substantial genetic differentiation, while a low value indicates little differentiation. The validity of genetic patterns detected by AMOVA is tested through permutations. In cases where only a single factor is present, the permutation procedure involves freely shuffling all data points, disregarding any inherent group structure, to assess the significance of the observed variation. When multiple hierarchical factors are considered such as genus and species, genetic patterns are tested through systematic randomization process. Each species is reassigned to a random genus while keeping the number of species in each genus consistent. This step determines if the genetic differences observed between genera ($\sigma^2_{amonggenera}$) might have arisen by chance. Within each genus, species identities are interchanged to determine if the genetic variation within genera, between species ($\sigma^2_{amongspecieswithingenera}$), aligns with natural patterns. Finally, individuals are completely randomized across the dataset to assess the variation within species ($\sigma^2_{withinspecies}$) and to test the null hypothesis of no genetic structure at any level. If this comprehensive randomization significantly reduces the observed genetic differentiation, leading to a low probability of more extreme variance components occurring by chance (indicated by low p-values), it suggests that the observed genetic patterns are unlikely to be the result of random variation.

The AMOVA Tab in this application offers a comprehensive framework for assessing genetic variations. It generates both Species Differentiation AMOVA tables and Population Differentiation AMOVA tables when multiple genera or species are present within a cluster. Conversely, when a cluster contains only a single species, the tool exclusively calculates the Population Differentiation AMOVA tables. This specific analysis focuses on the genetic variations between populations, rather than on taxonomic distinctions. Users can initiate the AMOVA analysis by clicking the "Run AMOVA" button. Users have the option to select indel coding methods before running the analysis, as detailed in section 2.6.2. In the case of species differentiation, the AMOVA assesses genetic variance at different levels: between genera (if multiple genera are present), between species (within genera), and the residual variance, which represents variation within species. For population differentiation, the tool calculates variance between population and the residual variance, representing variations within each population. Associated p-values and fixation indices can also be seen in the tables (Figure 23).

The AMOVA is conducted using the amova function from the R package **haplotypes 2.0-beta**. Clusters resulting in NaN (Not a Number) for variance components are omitted from the results. This typically happens when the dataset is not suitable for the analysis, leading to an inability to calculate valid variance components. As a result, not all clusters may be represented in the final output.

Following the creation of AMOVA tables, users have several filtering options. By using the *p-value Threshold* settings, tables showing clusters with lower p-values can be displayed. This thresholding is particularly beneficial for large datasets. Additionally, specific features can be selected with the *Select a Feature* controller, to display levels of selected features as cluster table headers. The *Select a Cluster by Level* settings can be used to filter tables to include only the selected level. If *ALL* is selected, this filter can be disabled, showing all available clusters (Figure 24).



Select a Feature: Organism

Select a Cluster by Level: ALL

p-value Threshold: 1

Species Differentiation AMOVA Tables

Cluster-1

Features: Quercus crassifolia (1); Quercus ariifolia (1); Lithocarpus microspermus (1); Lithocarpus fenestratus (4); Lithocarpus xylocarpus (3); Quercus schottkyana (3); Lithocarpus hancei (3); Castanopsis wattii (5); Lithocarpus confinis (2); Lithocarpus crassifolius (2); Quercus stewardiana (1); Castanopsis delavayi (2); Castanopsis fleuryi (4); Quercus kerrii (2); Lithocarpus litseifolius (5); Quercus aliena (3); Quercus franchetii (3); Quercus variabilis (2); Lithocarpus pachyphyllus (2); Castanopsis ceratacantha (2); Lithocarpus truncatus (2); Castanopsis argyrophylla (3); Castanopsis ferox (1); Quercus guyavifolia (3); Quercus glauca (1)

```
Analysis of Molecular Variance Table
--------------------------------------------------------------------------------
 Source     Df   Sum Sq     Var Comp   Per Var (%)    p-value

 Between     2   34.83877    0.65887    26.08797      0.000 ***
 Genus
 Between    22   75.04020    1.13383     44.89402     0.000 ***
 Species
 Residuals  36   26.38333    0.73287     29.01800     0.000 ***

 Total      60   136.26230   2.52557
---
Signif. test: 999 permutations
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-----------------------------------------------------------------------------
 Fixation Indices:
ΦSC :     0.60740
ΦCT :     0.26088
ΦST :     0.70982
```

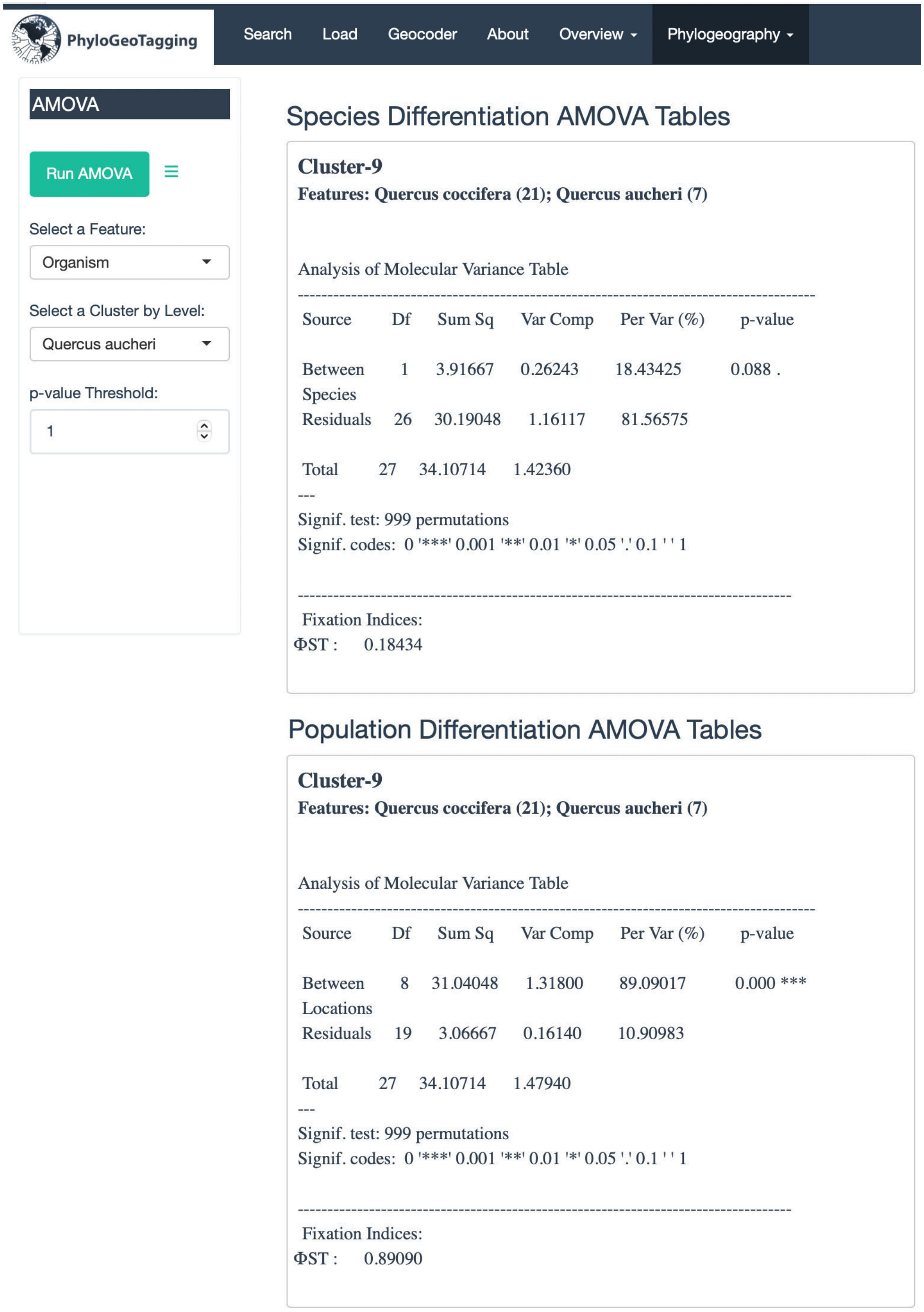
Population Differentiation AMOVA Tables

Cluster-1

Features: Quercus crassifolia (1); Quercus ariifolia (1); Lithocarpus microspermus (1); Lithocarpus fenestratus (4); Lithocarpus xylocarpus (3); Quercus schottkyana (3); Lithocarpus hancei (3); Castanopsis wattii (5); Lithocarpus confinis (2); Lithocarpus crassifolius (2); Quercus stewardiana (1); Castanopsis delavayi (2); Castanopsis fleuryi (4); Quercus kerrii (2); Lithocarpus litseifolius (5); Quercus aliena (3); Quercus franchetii (3); Quercus variabilis (2); Lithocarpus pachyphyllus (2); Castanopsis ceratacantha (2); Lithocarpus truncatus (2); Castanopsis argyrophylla (3); Castanopsis ferox (1); Quercus guyavifolia (3); Quercus glauca (1)

```
Analysis of Molecular Variance Table
--------------------------------------------------------------------------------
 Source     Df   Sum Sq     Var Comp   Per Var (%)    p-value

 Between     1    6.25382    1.04690   32.20805      0.006 **
 Locations
 Residuals  59   130.00847    2.20353    67.79195

 Total      60   136.26230   3.25043
---
Signif. test: 999 permutations
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-----------------------------------------------------------------------------
 Fixation Indices:
ΦST :     0.32208
```

Figure 23. The AMOVA Tab showcasing results with the presence of multiple genera and species. The displayed tables summarize the variance components, their percentage contributions, associated p-values, and fixation indices.



Select a Feature: Organism

Select a Cluster by Level: Quercus aucheri

p-value Threshold: 1

Species Differentiation AMOVA Tables

Cluster-9
Features: Quercus coccifera (21); Quercus aucheri (7)

```
Analysis of Molecular Variance Table
-----------------------------------------------------------------------------------
 Source      Df    Sum Sq     Var Comp    Per Var (%)     p-value

 Between      1    3.91667    0.26243     18.43425        0.088 .
 Species
 Residuals   26   30.19048    1.16117     81.56575

 Total       27   34.10714    1.42360
---
Signif. test: 999 permutations
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-------------------------------------------------------------------------------
 Fixation Indices:
ΦST :     0.18434
```

Population Differentiation AMOVA Tables

Cluster-9
Features: Quercus coccifera (21); Quercus aucheri (7)

```
Analysis of Molecular Variance Table
-----------------------------------------------------------------------------------
 Source      Df    Sum Sq     Var Comp    Per Var (%)     p-value

 Between      8   31.04048    1.31800     89.09017        0.000 ***
 Locations
 Residuals   19    3.06667    0.16140     10.90983

 Total       27   34.10714    1.47940
---
Signif. test: 999 permutations
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-------------------------------------------------------------------------------
 Fixation Indices:
ΦST :     0.89090
```

Figure 24. Overview of the AMOVA Tab showing the assessment of genetic variation within species and between populations, with a focus on a selected level, *Quercus aucheri*. The displayed tables summarize the variance components, their percentage contributions, associated p-values, and fixation indices.

2.6.7. Diversity and Differentiation Parameters

The comprehensive evaluation of genetic diversity and differentiation within and among populations is crucial for understanding the structure and evolutionary dynamics of species. These assessments are key to identifying the intricate patterns of genetic variation that have emerged over time through evolutionary processes and geographic distribution. Alongside the extensive analysis provided by AMOVA, partitions genetic variance across different hierarchical factors, the methodology introduced by Pons and Petit (1996) precisely targets the diversity and differentiation at the haplotype level. Haplotypes, which are specific sequences of DNA, carry a historical record that is crucial for understanding the genetic basis of population structures. This method provides a detailed view of genetic nuances and illuminates the phylogeographic structure, considering both the frequency of haplotypes (alleles) and their genetic distances.

In the field of population genetics, understanding the aspects of genetic diversity is vital for elucidating evolutionary history and phylogeographic patterns.

In the context of diversity parameter calculation, distinct measures are employed to capture the variations, h_S, v_S, h_T and v_T, respectively. The average within-population diversity v_S, which considers the genetic distances between haplotypes, is quantified as follows:

$$v_s = \sum_{i,j} \pi_{ij}\left(p_i p_j + c_{ij}\right)$$

The total diversity parameter is calculated as:

$$v_T = \sum_{i,j} \pi_{ij}\, p_i p_j$$

where p_i and p_j are the frequencies of the i^{th} and j^{th} haplotypes within the population, and the term π_{ij} denotes the genetic distance between these two haplotypes. Meanwhile, c_{ij} is the covariance between the frequencies of the two haplotypes. In these calculations, both frequency and sequence distances are considered. Conversely, the classical measure of diversity h_S and h_T does not consider the genetic distances between haplotypes. When all genetic distances π_{ij} are set to 1, the measure v simplifies to h, aligning with the classic definition of the probability that two randomly chosen haplotypes from the population will be different based solely on their frequency. Limitations inherent in working with real-world populations of limited size necessitated the development of unbiased multinomial estimators for v_S and v_T. The estimators $\widehat{v_S}$ and $\widehat{v_T}$ provide adjusted measures to account for the finite sizes of real-world populations:

$$\widehat{v_S} = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{n_k}{n_k - 1}\right)\sum_{i,j} \pi_{ij} x_{ki} x_{kj}$$

$$\widehat{v_T} = \sum_{i,j} \pi_{ij} x_{.i} x_{.j} - \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{i,j} \pi_{ij}(x_{ki} - x_{.i})(x_{kj} - x_{.j})$$

The estimator $\widehat{v_S}$, reflecting the average within-population diversity, takes into account the genetic distances π_{ij}, between haplotypes i and j, as well as the observed frequencies x_{ki} and x_{kj} of these haplotypes within each sampled population k. In the formula, n is the number of populations, n_k is the sample size of the k^{th} population. This estimator is adjusted for sample size to provide an unbiased measure of within-population diversity.

The estimator $\widehat{v_T}$, representing the total diversity, is concerned with the genetic variance within and between populations. In the formula, $x_{.i}$ and $x_{.j}$ are the average frequencies of haplotypes across all populations. The first term represents the expected genetic variance under random mating, while the second term adjusts for deviations due to population structure. These estimators, $\widehat{v_S}$ and $\widehat{v_T}$, are derived from the principles of unbiased multinomial estimators and are specifically tailored to address the challenges posed by finite sample sizes in empirical population genetics studies.

This framework proposes two key measures of differentiation, G_{ST} and $\widehat{N_{ST}}$. G_{ST} is derived solely from haplotype frequencies, offering a view of differentiation that does not consider the similarities between haplotypes. Conversely, $\widehat{N_{ST}}$ incorporates these similarities, thereby providing a measure that reflects phylogeographic structures. They are derived using diversity parameters. The formulas for G_{ST} and $\widehat{N_{ST}}$ are expressed as:

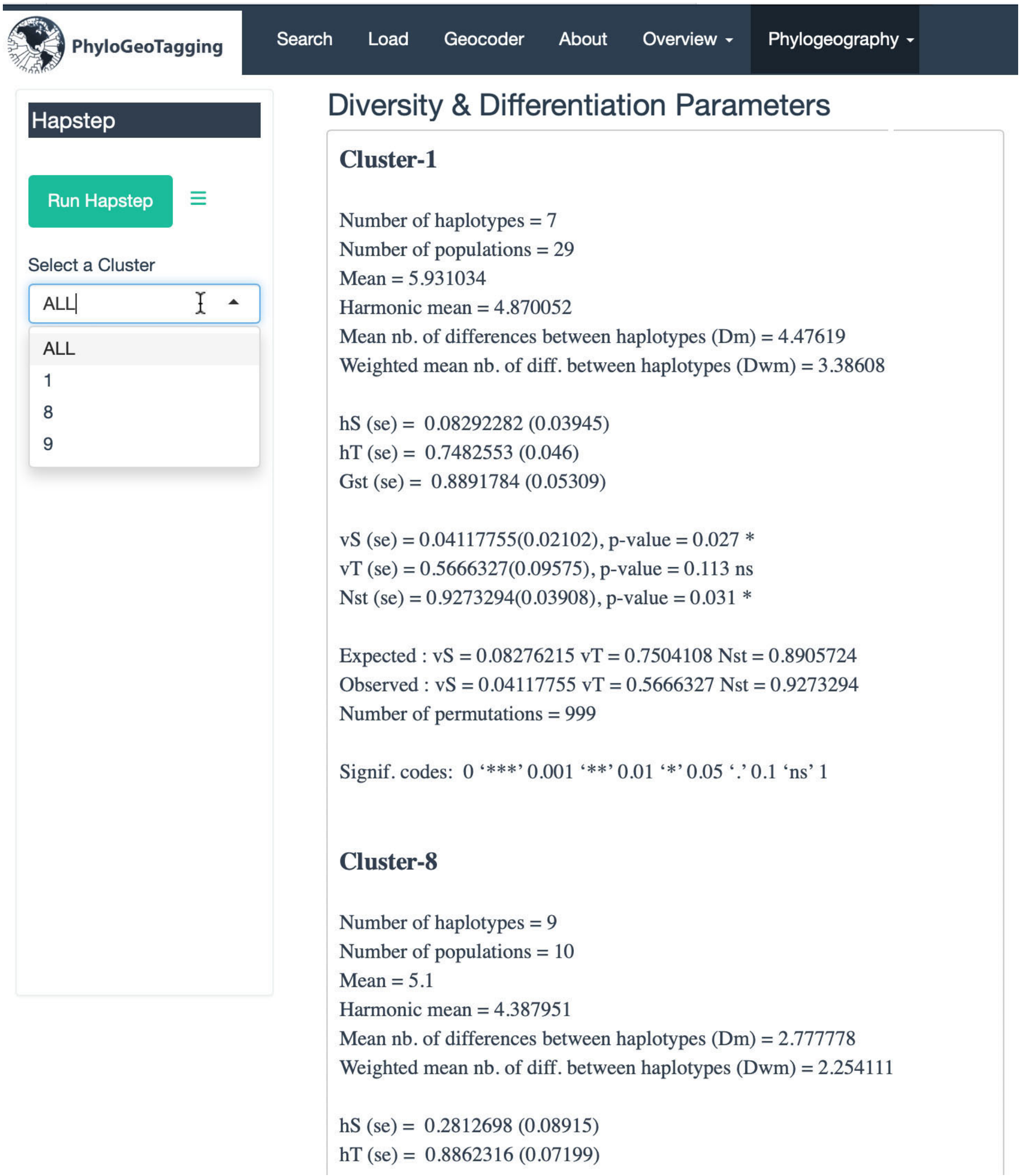
$$G_{ST} = \frac{h_T - h_S}{h_T}$$

and

$$\widehat{N_{ST}} = \frac{\widehat{v_T} - \widehat{v_S}}{\widehat{v_T}}$$

To ensure robust statistical inference, permutation tests are utilized to assess the significance of the observed differentiation. By randomly shuffling haplotypes among populations and recalculating $\widehat{v_S}$, $\widehat{v_T}$, and $\widehat{N_{ST}}$, the distribution of these statistics under the null hypothesis of no differentiation can be created. Observed values are then compared against this distribution to determine their statistical significance. In summary, these parameters offer a comprehensive framework for deciphering the genetic structure of populations, guiding conservation strategies, and illuminating the evolutionary history of taxa.

These parameters are calculated using the hapstep function from **haplotypes 2.0-beta** package. This function implements the TURBO PASCAL program **Hapstep** (version 2001, Pons and Petit, 1996). The Hapstep Tab within this application displays tables generated by the hapstep function. Users can initiate the calculation by clicking the “Run Hapstep” button (see Figure 25). Subsequent to table generation, users have the option to utilize the *Select a Cluster* settings to filter the clusters. Selecting *ALL* displays all available clusters.



```
Cluster-1

Number of haplotypes = 7
Number of populations = 29
Mean = 5.931034
Harmonic mean = 4.870052
Mean nb. of differences between haplotypes (Dm) = 4.47619
Weighted mean nb. of diff. between haplotypes (Dwm) = 3.38608

hS (se) =  0.08292282 (0.03945)
hT (se) =  0.7482553 (0.046)
Gst (se) =  0.8891784 (0.05309)

vS (se) = 0.04117755(0.02102), p-value = 0.027 *
vT (se) = 0.5666327(0.09575), p-value = 0.113 ns
Nst (se) = 0.9273294(0.03908), p-value = 0.031 *

Expected : vS = 0.08276215 vT = 0.7504108 Nst = 0.8905724
Observed : vS = 0.04117755 vT = 0.5666327 Nst = 0.9273294
Number of permutations = 999

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 'ns' 1


Cluster-8

Number of haplotypes = 9
Number of populations = 10
Mean = 5.1
Harmonic mean = 4.387951
Mean nb. of differences between haplotypes (Dm) = 2.777778
Weighted mean nb. of diff. between haplotypes (Dwm) = 2.254111

hS (se) =  0.2812698 (0.08915)
hT (se) =  0.8862316 (0.07199)
```

Figure 25. Interface of the Hapstep Tab in the application.

It should be noted that the analysis calculates the estimates $\widehat{v_S}$, $\widehat{v_T}$, and $\widehat{N_{ST}}$. However, in the output of the program and in this thesis work, these notations are presented as v_S, v_T, and N_{ST} to maintain consistency with the original implementation and follows the conventions established by the initial version of the software.

Beside parameters h_S, h_T, G_{ST}, v_S, v_T, and N_{ST}, with their standard errors (se) given in parentheses, tables include a variety of other parameters. The tables provide the total number of haplotypes and populations, along with the mean number of haplotypes in each population. A key statistic included is the harmonic mean of populations, which is sensitive to low-frequency data, thus reducing the impact of high outliers. Dm, the mean number of differences between haplotypes measures the average genetic distance across pairs of haplotypes, calculated from the haplotype distance matrix. Another important measure is Dwm, the weighted mean number of differences between haplotypes. This differs from Dm by factoring in the frequency of each haplotype within the populations, thereby weighting the differences based on the relative abundance of each haplotype. The tables also contrast expected and observed values for v_S, v_T, and N_{ST}, allowing a comparison between expected measures (under a null model with no differentiation) and the observed values. Additionally, the number of permutations used in the analysis and significance codes are provided.

This method accepts only populations comprising a minimum of three individuals. Populations smaller than this threshold are excluded. Additionally, clusters that yield “NA” for parameters are also omitted from the final presentation, meaning not all clusters may be represented in the results.

CHAPTER 3

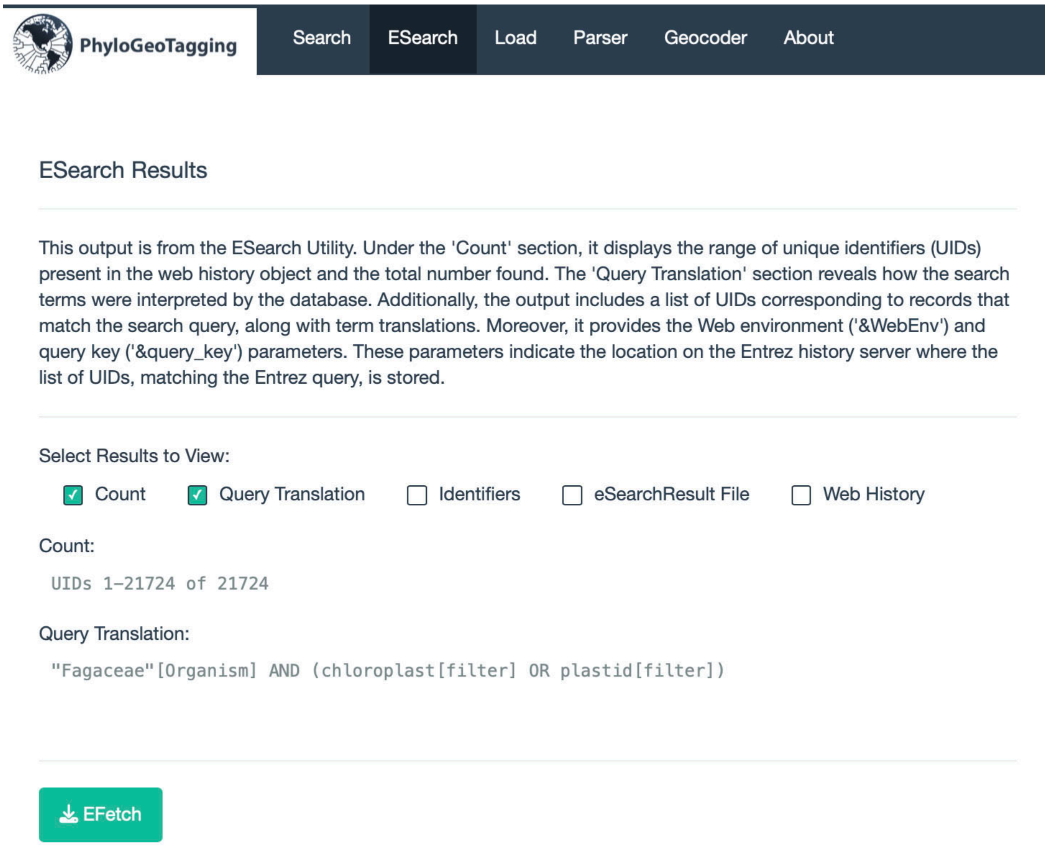
RESULTS and DISCUSSON

In the Results Section, sample analyses are demonstrated using GenBank Release 257.0, available during the period of study at the end of 2023. Results will be given under different sub-sections: 3.1.1 for the plant Family Fagaceae (Beech/Oak Family), 3.1.2 for the Genus *Salvia* (Sage), and 3.1.3 for the insect Genus *Apis* (Honey bee). In the first example, more information will be provided about the methods. In the second and third examples, the methodology will be less mentioned. Additionally, to highlight different aspects of toolkit use, different methods will be emphasized in the analysis of these examples. This section of the thesis is deliberately crafted to include trial and error results, guiding researchers in the practical use of the package, rather than presenting polished, publication-ready results for the given examples. Not all possible combinations are demonstrated due to their vast number. The analysis also explores the positive and negative impacts of certain selected parameters on the outcomes.

3.1. Examples

3.1.1. Comprehensive Example: Family Fagaceae

The Family Fagaceae, one of the most important woody families globally, encompasses nearly 1000 species distributed across a wide range of ecosystems. Chloroplast DNA sequences (cpDNA) of the Fagaceae Family were thoroughly searched using the terms “Fagaceae [ORGANISM] AND (chloroplast[filter] OR plastid[filter])”. This search yielded a total of 21,724 sequences, as detailed in Figure 26. During the parsing process, 10,532 sequences were successfully obtained. Notably, addresses are unavailable for 11,187 of these records, and five of them lack sequence data, as detailed in the log file. Of the parsed 10,532 records, a significant majority, 10,420, have been geocoded. Five sequences were not included in the geocoded data set, as ISO codes could not be obtained from their addresses. The remaining sequences, amounting to 10,420, were successfully geocoded. Additionally, 107 sequences identified as complete genome sequences were excluded from the analysis.



ESearch Results

This output is from the ESearch Utility. Under the 'Count' section, it displays the range of unique identifiers (UIDs) present in the web history object and the total number found. The 'Query Translation' section reveals how the search terms were interpreted by the database. Additionally, the output includes a list of UIDs corresponding to records that match the search query, along with term translations. Moreover, it provides the Web environment ('&WebEnv') and query key ('&query_key') parameters. These parameters indicate the location on the Entrez history server where the list of UIDs, matching the Entrez query, is stored.

Select Results to View:

☑ Count ☑ Query Translation ☐ Identifiers ☐ eSearchResult File ☐ Web History

Count:

```
UIDs 1-21724 of 21724
```

Query Translation:

```
"Fagaceae"[Organism] AND (chloroplast[filter] OR plastid[filter])
```

Figure 26. Overview of the ESeach Tab showing the output of search for Chloroplast DNA sequences (cpDNA) of Fagaceae.

A total of 380 organism names were found under the “Organism” feature. The largest one, including 1,587 species, is *Quercus serrata*, and the predominant records are the *Quercus* taxa in the dataset (Figure 27). However, within these names, specimens of unknown species (e.g., “Quercus sp.”), or ambiguous names such as “Lithocarpus environmental sample” or “fossil Fagus pollen” that lack specific taxon information, are present. Hybrids names such as “Quercus laevis x Quercus incana” are also found. Samples that do not have a valid species name are kept separate during some analyses, like AMOVA.

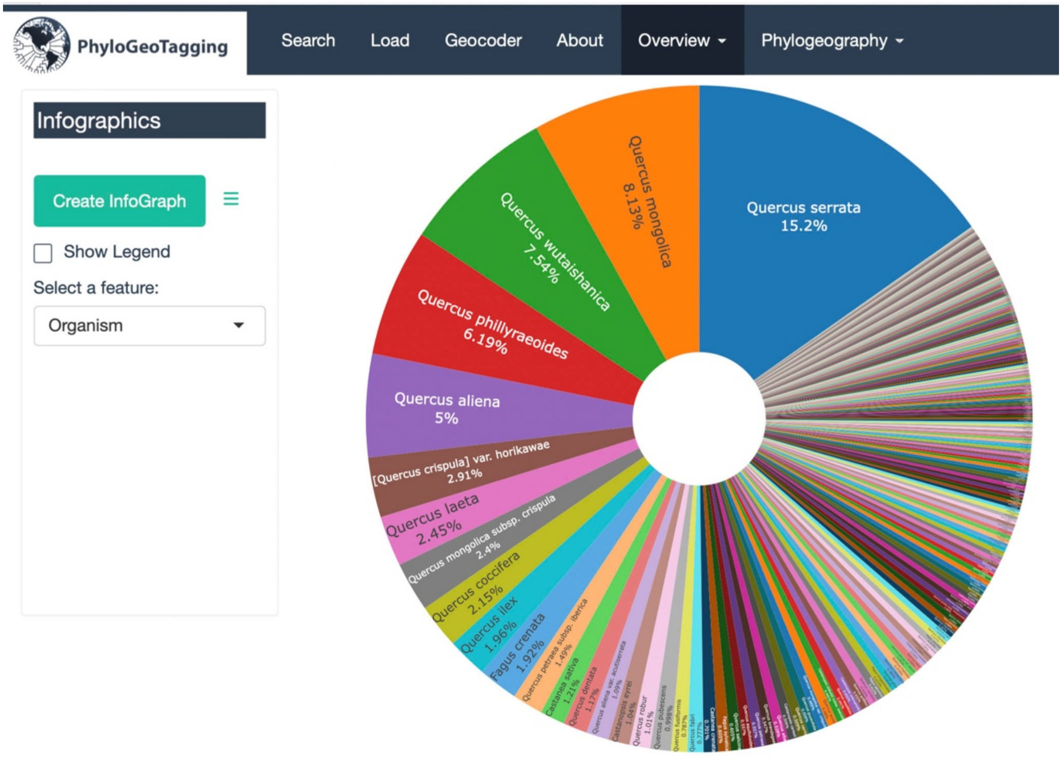
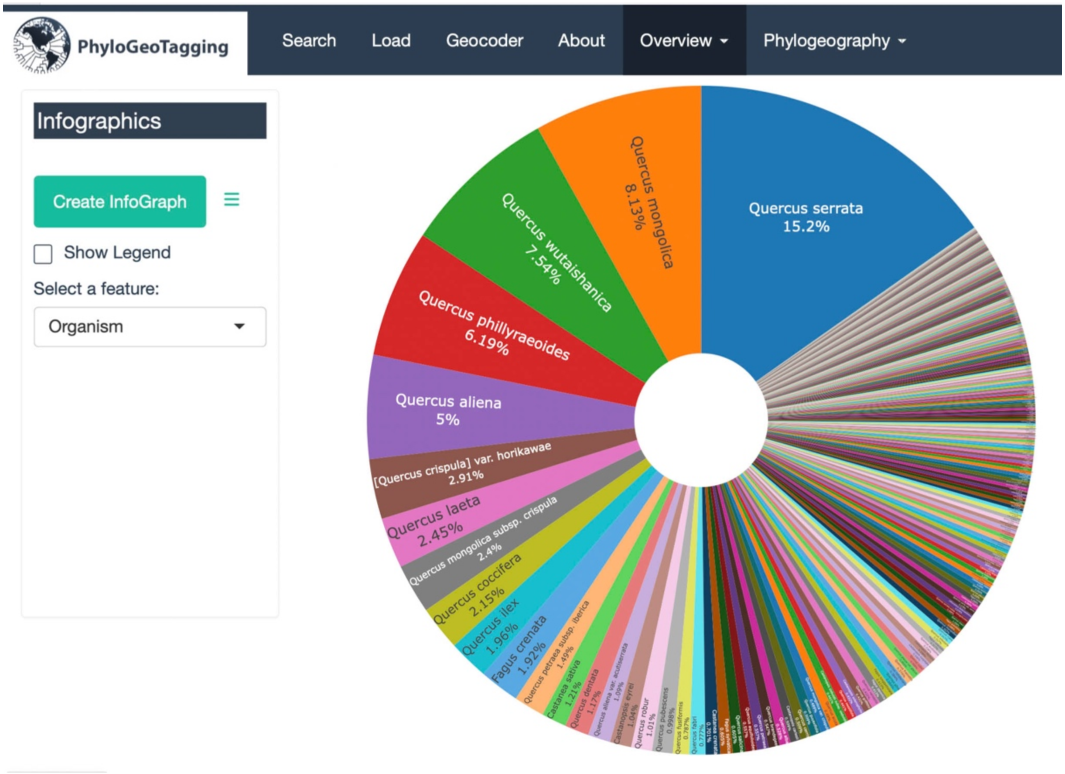


Figure 27. Graphical representation of Organism feature of the GenBank cpDNA records for Fagaceae.

Samples represent 56 countries, of which Japan, China, Mexico, Italy, and Turkey are the top five countries contributing the most to the Fagaceae dataset (Figure 28).

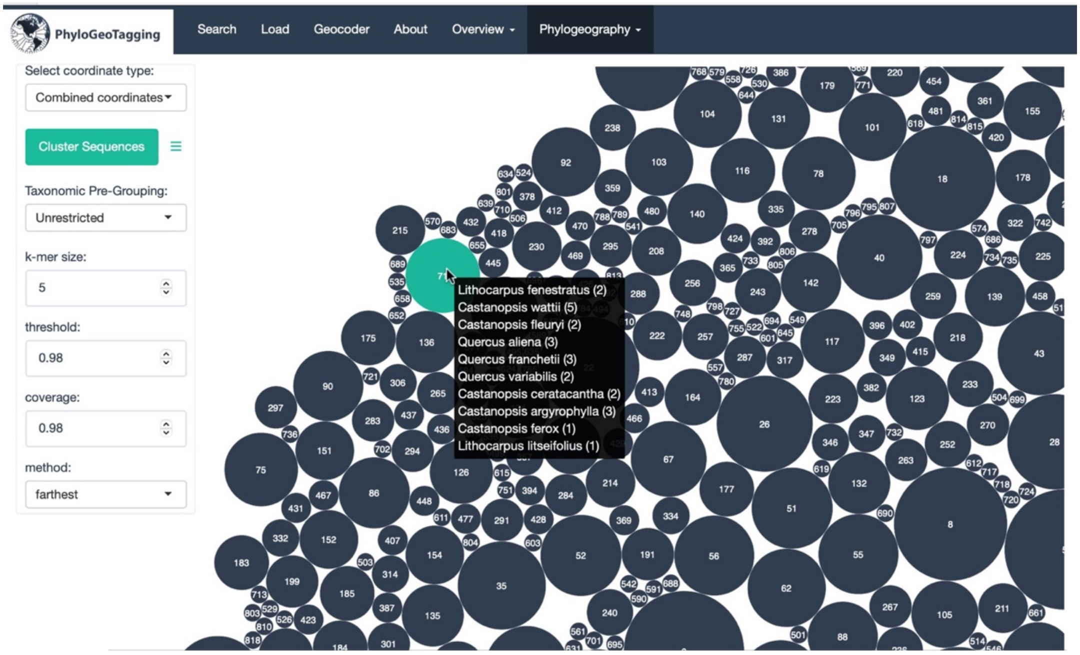


☐ Show Legend

Select a feature: Organism

Figure 28. Graphical representation of Organism feature of the GenBank cpDNA records for Fagaceae.

In the clustering process following geocoding, the *Taxonomic Pre-Grouping* setting can be assigned as *Unrestricted*. This approach allows for the comparison of a broad spectrum of species and even genera within the family, thereby enriching the analysis. To ensure accuracy and to prevent the inclusion of unrelated sequences from different taxa in the same cluster, the k-mer similarity threshold and the coverage threshold were initially set at a high proportion of 98%, as shown in Figure 29. However, it is important to emphasize that, depending on the specific requirements of the analysis and the data characteristics, an increase in this threshold value may be necessary. As will be discussed in subsequent part of the section, the threshold was indeed adjusted to 99% in later analyses. This modification was implemented to more accurately and effectively capture phylogeographic information. Additionally, this subsequent parameter adjustment provides an opportunity to delineate the effects of both threshold values on the representation of phylogeographic relationships in this first example.



Select coordinate type: Combined coordinates

Taxonomic Pre-Grouping: Unrestricted

k-mer size: 5

threshold: 0.98

coverage: 0.98

method: farthest

Lithocarpus fenestratus (2)
Castanopsis wattii (5)
Castanopsis fleuryi (2)
Quercus aliena (3)
Quercus franchetii (3)
Quercus variabilis (2)
Castanopsis ceratacantha (2)
Castanopsis argyrophylla (3)
Castanopsis ferox (1)
Lithocarpus litseifolius (1)

Figure 29. Overview of the ESeach Tab showing the output of search for Chloroplast DNA sequences (cpDNA) of Fagaceae.

Setting the k-mer similarity threshold to 98% led to the formation of 818 clusters of varying sizes. In analyzing these clusters, the user has the flexibility to adopt a broader analytical perspective, focusing on aspects like haplotype richness and ΦST barriers. Alternatively, a more targeted approach can be taken, concentrating on specific clusters that exhibit phylogeographic differentiation, as revealed through methods like AMOVA, IBD, and Hapstep analyses. These multiple approaches allow for both a comprehensive overview and detailed examination of distinct phylogeographic patterns within the clusters.

In the context of Haplotype Richness, specifically in the Weighted Haplotype Richness tab, calculations for each cluster were performed by aggregating data based on cluster and country features, considering the sizes of the clusters. This method emphasizes the importance of larger clusters, providing a more accurate representation of genetic diversity and reducing the potential for bias from smaller clusters. This metric serves as an indicator of the diversity of haplotypes in a country, essential for genetic studies on reproductive taxa. It is crucial to ensure accurate grouping of taxa; including unrelated taxa to increase cluster sizes may inflate the haplotype richness value, but this would not reflect true biological significance. The weighted haplotype richness value, therefore, represents the average diversity of genetic variants that can be produced by a fertile lineage within the Fagaceae family. For instance, the observed value of 9.42 for Turkey represents the average number of distinct genetic variants produced by naturally reproducing Fagaceae taxa (Figure 30). This includes only taxa from the same species, or different species which can form hybrids, such as the Roburoid oaks, known for their natural hybridization (Tekpinar et al., 2021). However, it is important to note that taxa with highly conserved sequences might be grouped together, despite their inability to hybridize. This is an inherent limitation when adopting an Unrestricted pre-grouping approach, and if necessary, a different *Taxonomic Pre-Grouping* setting or a threshold might be more appropriate. This issue is addressed in later discussions.

The geographic distribution of the calculated metric, as depicted on the map, shows the highest haplotype richness in Mexico, China, and Turkey. This observation corresponds with regions where the Fagaceae family, particularly the Genus *Quercus*, exhibits significant diversity (Axelrod, 1983). However, it is crucial to interpret these findings in the correct context; for instance, the lower value in the USA should not be misconstrued as a lack of diversity but rather attributed to the relatively smaller number of samples in the GenBank. This highlights the importance of sample size in influencing the results of such analyses.

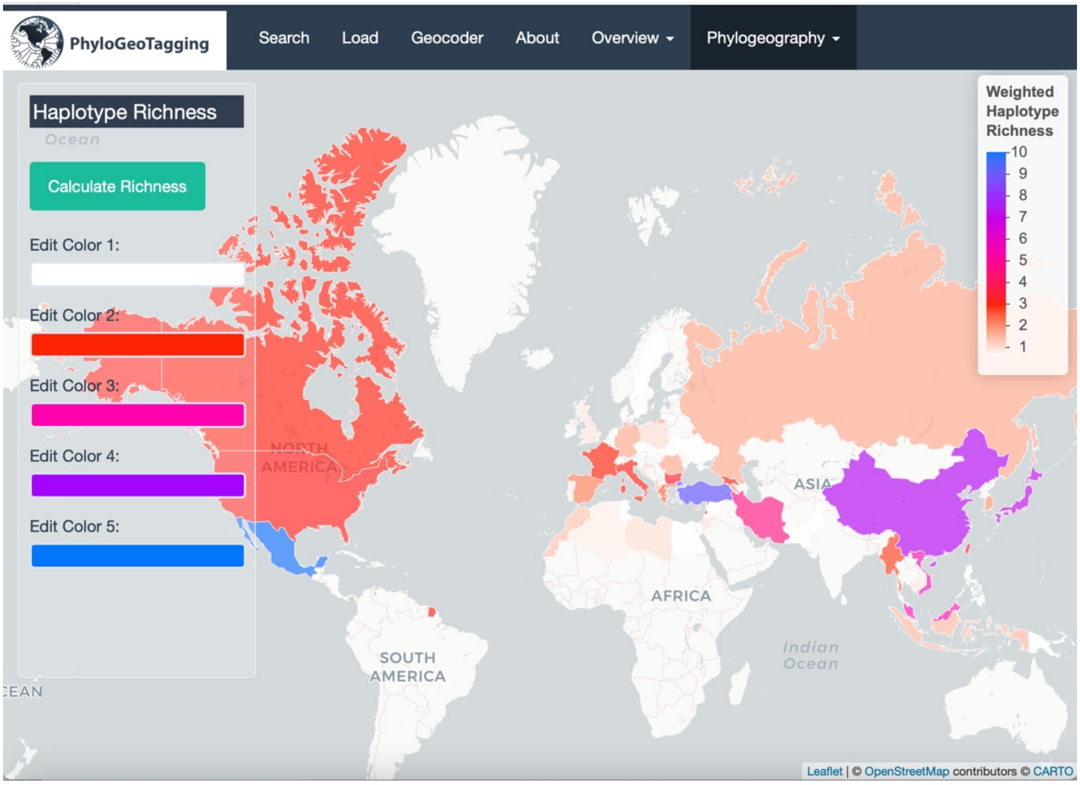
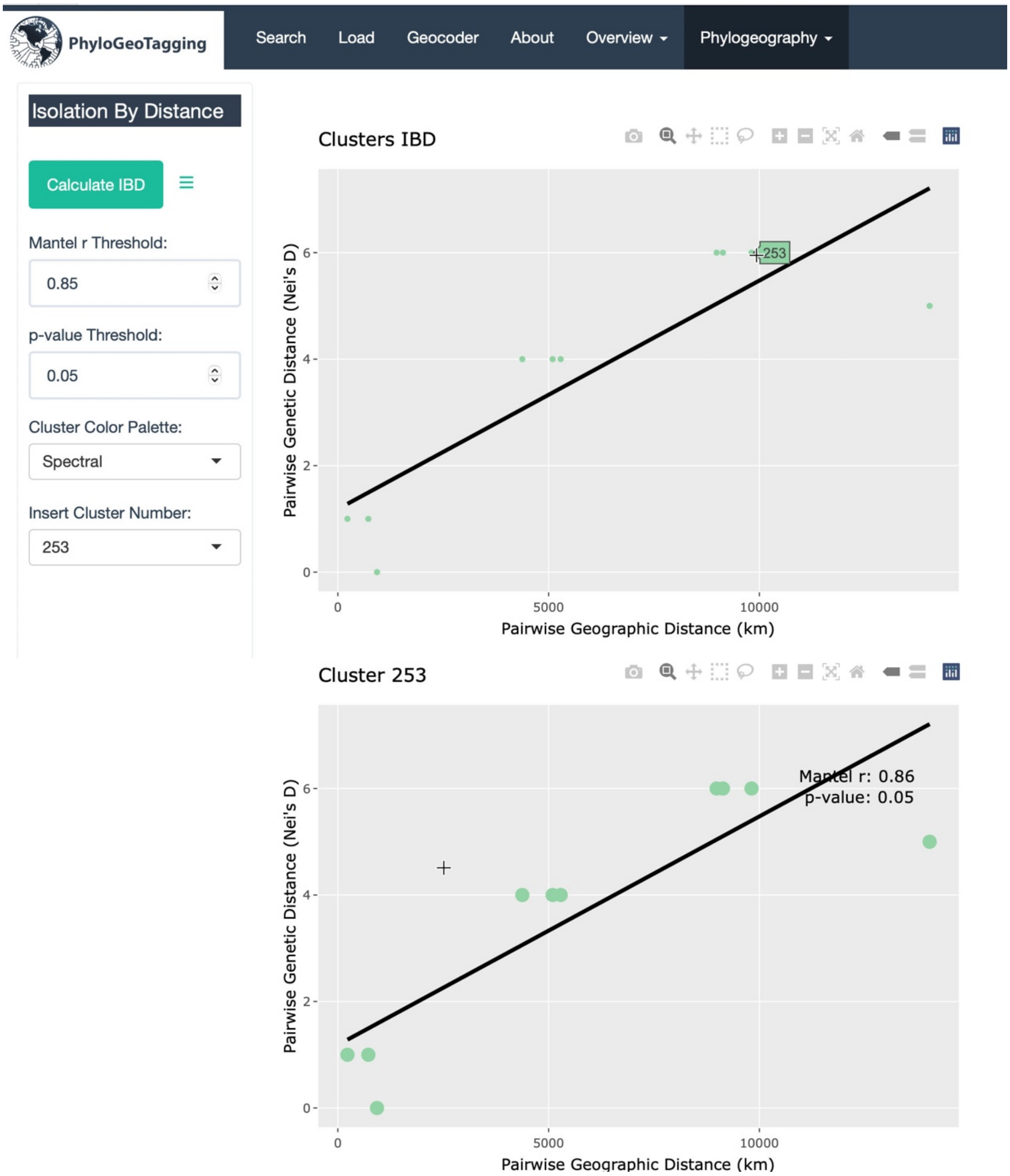


Figure 30. Weighted Haplotype Richness analysis results for Fagaceae. The Taxonomic Pre-Grouping setting was assigned as Unrestricted, and the k-mer similarity threshold and the coverage threshold were set to 98.

Before delving into the ΦST Barrier analysis, it is beneficial to examine which clusters exhibit geographic differentiation using haplotypes maps, IBD, AMOVA, and Hapstep analysis. Subsequently, the barriers created by these clusters can be explored in detail. Therefore, the latter analyses will be carried out first and the results will be examined in detail.

As a result of the IBD analysis, it is observed that the Mantel correlation coefficient reaches 1 in some clusters with low sample sizes. However, for a more robust data analysis, the p-value threshold is initially set. A p-value of < 0.05 is widely accepted as the threshold for statistical significance, indicating that the observed correlation would occur by chance less than 5% of the time under the null hypothesis. As such, clusters exhibiting a p-value above 0.05 are not further addressed in this analysis, focusing instead on those with statistically significant results. By progressively decreasing the Mantel r Threshold filter step-by-step, the cluster with the highest yet significant correlation can be identified. For instance, for Cluster-253, the r value found is 0.86, and the rounded p-value is 0.05, as shown in Figure 31. Upon examining the taxa in this cluster, a notable divergence between populations is evident, serving as an exemplary case of both intra-breed and inter-geographical regional divergence. The H1 and H2 haplotypes, differentiated by a single mutation in Japan, are represented by species from the genus *Castanopsis*, as depicted in Figure 32. The H4 haplotype, found in Singapore and 4 mutations away from the H2 haplotype, is comprised of individuals from the genus *Lithocarpus*. Meanwhile, the H3 haplotype in the USA, which is 6 mutations distant from both H2 and H4, is formed by individuals belonging to the genus *Quercus*. The sequences forming these haplotypes, which are relatively long (around 700 bases) and appear to belong to a part of the matK cpDNA region as per the “gene” feature in the records, are highly conserved across genera, leading to grouping of species from different genera in the same cluster. Nevertheless, the accumulated mutations between these genera, alongside geographical factors, play a role in the phylogeographic differentiation observed in this example.



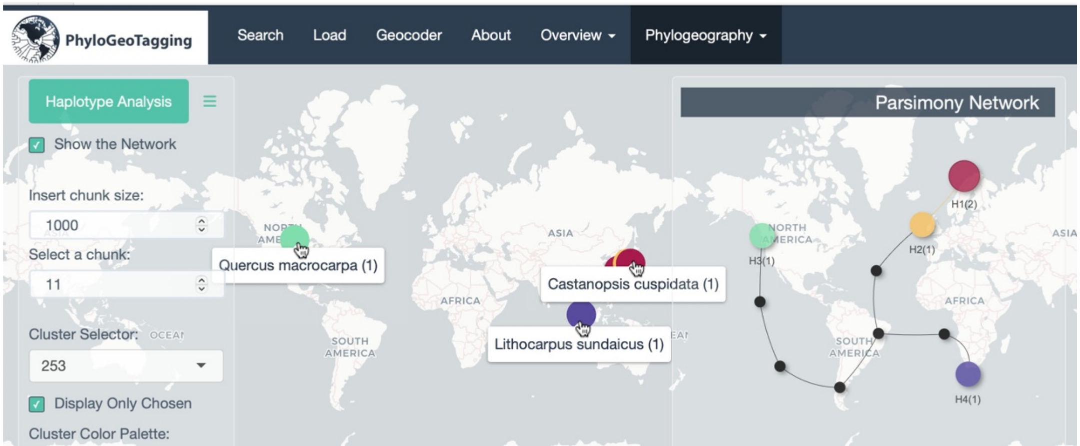
Mantel r Threshold: 0.85

p-value Threshold: 0.05

Cluster Color Palette: Spectral

Insert Cluster Number: 253

Figure 31. Mantel test results for Cluster-253. The Mantel r value found is 0.86, which is the highest correlation found in the dataset and is significant.



☑ Show the Network

Insert chunk size: 1000

Select a chunk: 11

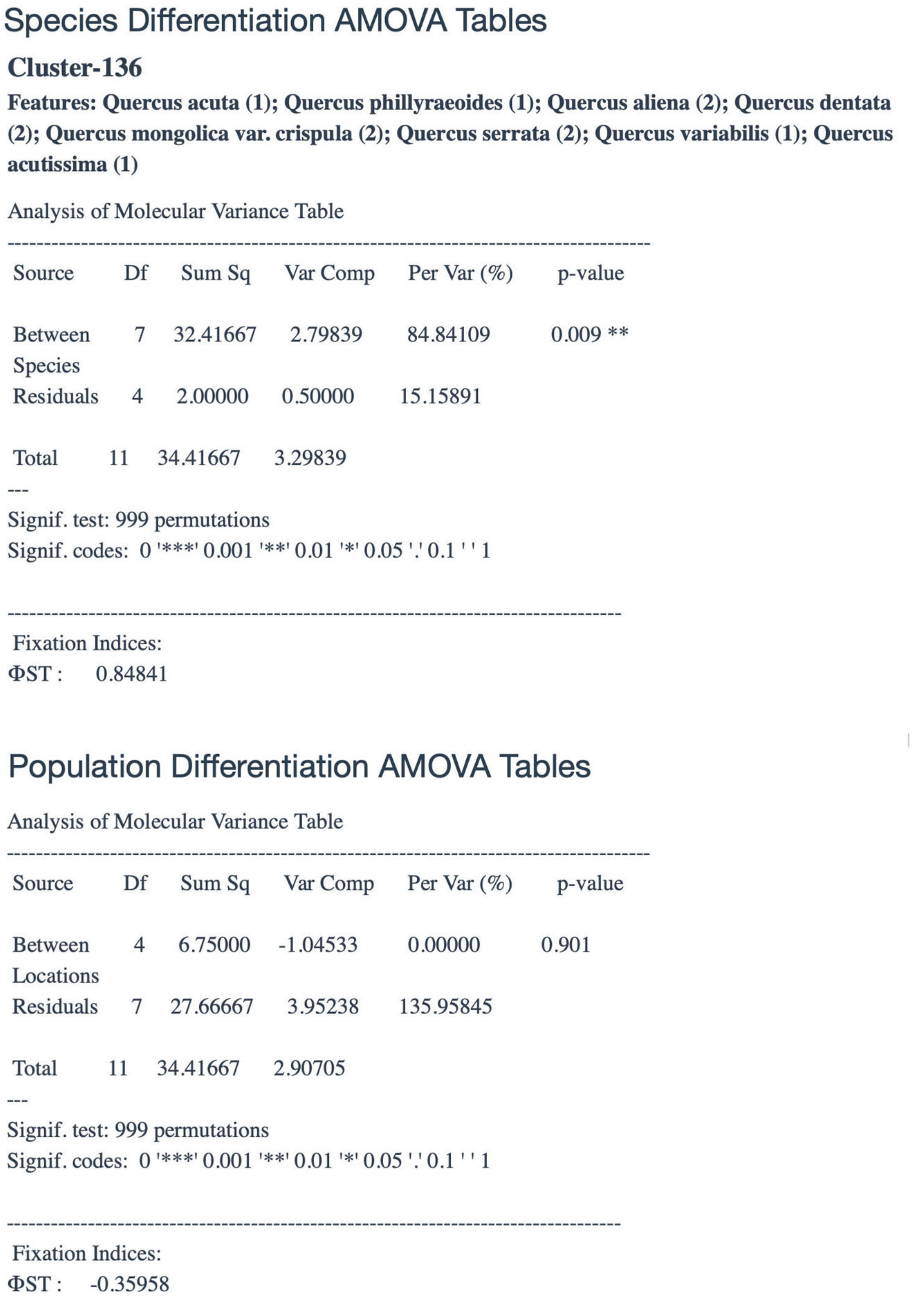
Cluster Selector: 253

☑ Display Only Chosen

Figure 32. Haplotype map and parsimony network for the Cluster-253.

When the Mantel r value is slightly lowered, a notable observation emerges in Cluster-136, which exhibits an r value of 0.826. The high correlation observed in this case is primarily due to the taxonomic heterogeneity within the cluster. This cluster encompasses specimens from the subgenus Quercus (e.g., *Quercus aliena*) and the subgenus Cerris (e.g., *Quercus phillyraeoides*), both belonging to the genus *Quercus* yet incapable of natural hybridization. The phylogenetic distance discerned between these specimens contributes to an amplified perception of population differentiation. In this context, examining the AMOVA results for this cluster can be beneficial. As indicated in the AMOVA tables presented in Figure 33, a substantial portion of the variation within the cluster is observed to occur between species. Furthermore, the occurrence of notably high negative Φ_{ST} values, as indicated in the lower AMOVA table for population separation, suggests that the differentiation within populations in this cluster significantly exceeds that observed between populations. Typically, Φ_{ST} values are expected to fall between 0 and 1. However, negative values do arise when there is an exceptional amount of genetic variation within populations compared to the variation between them. In this case, the negative values can be attributed to the confluence of different lineages within the population, resulting in an exceptionally high variance. This amalgamation leads to increased genetic diversity within the population, which, in contrast, diminishes when compared across different populations, thus resulting in the observed negative Φ_{ST} values. Such results are mirrored in the negative N_{ST} values derived from Hapstep analyses for this cluster, further emphasizing the disparity in genetic diversity. These negative N_{ST} values highlight that the within-population haplotypic diversity, which accounts for genetic distances, is greater than the total genetic diversity across populations. This unusual situation indicates a mix of species from different subgeneric clades within the cluster, akin to combining apples and oranges, which inflates the within-population variance. The implications of these negative Φ_{ST} and N_{ST} values underscore the importance of critically evaluating the cluster's composition and may call for a reassessment of the methodological framework. Modifying parameters, such as increasing the k-mer threshold or selecting a more conservative *Taxonomic Pre-Grouping* value, could yield a more accurate reflection of the genetic structure. Nonetheless, to showcase the capabilities of this package and the impact of specific selected parameters, it is beneficial to use the existing settings.

The subsequent cluster exhibiting the highest Mantel correlation is Cluster-129, with an Mantel r value of 0.813 AMOVA results for this cluster, which consists of the species *Fagus grandifolia* (North American beech), *Fagus crenata* (Japanese beech), and *Fagus sylvatica* (European beech), indicate significant phylogeographic differentiation, with interspecific differentiation (Φ_{ST}=0.83253) and population differentiation components at an intercontinental scale (Φ_{ST}=1).



Species Differentiation AMOVA Tables

Cluster-136

Features: Quercus acuta (1); Quercus phillyraeoides (1); Quercus aliena (2); Quercus dentata (2); Quercus mongolica var. crispula (2); Quercus serrata (2); Quercus variabilis (1); Quercus acutissima (1)

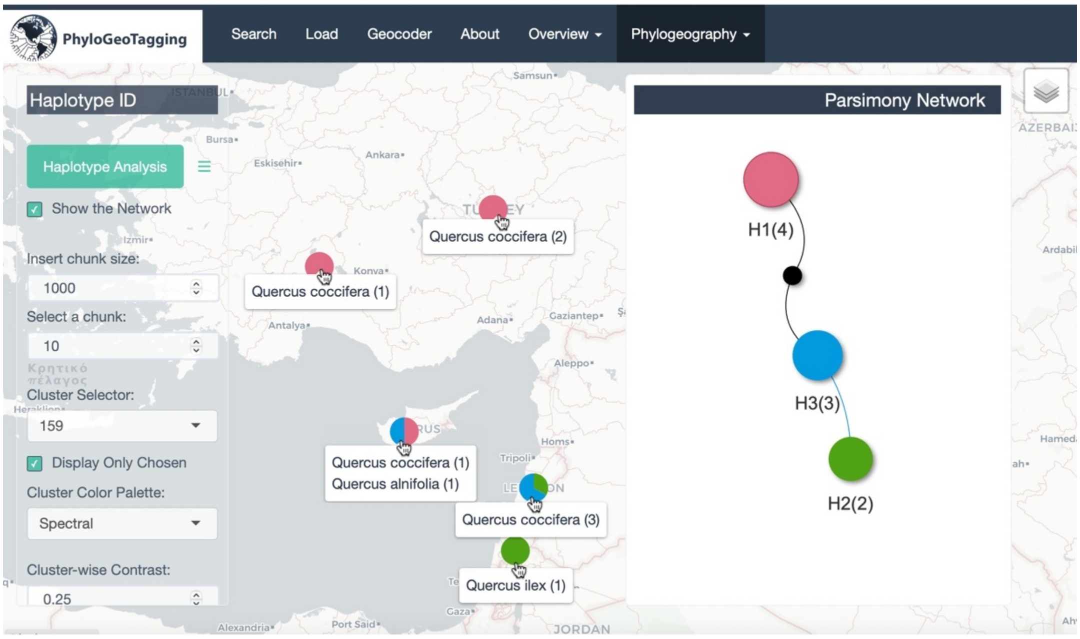
```
Analysis of Molecular Variance Table
-------------------------------------------------------------------------------------
 Source      Df    Sum Sq     Var Comp    Per Var (%)      p-value

 Between      7    32.41667    2.79839    84.84109        0.009 **
 Species
 Residuals    4    2.00000    0.50000     15.15891

 Total       11    34.41667    3.29839
---
Signif. test: 999 permutations
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

---------------------------------------------------------------------------------
 Fixation Indices:
ΦST :     0.84841
```

Population Differentiation AMOVA Tables

```
Analysis of Molecular Variance Table
-------------------------------------------------------------------------------------
 Source      Df    Sum Sq     Var Comp    Per Var (%)      p-value

 Between      4    6.75000    -1.04533     0.00000       0.901
 Locations
 Residuals    7    27.66667    3.95238    135.95845

 Total       11    34.41667    2.90705
---
Signif. test: 999 permutations
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

---------------------------------------------------------------------------------
 Fixation Indices:
ΦST :    -0.35958
```

Figure 33. Detailed AMOVA tables for Cluster-136, in which the taxonomic heterogeneity contributed to the variance.

Cluster-159 also exhibits a high Mantel value (Mantel r = 0.771). Within this cluster of closely related species, AMOVA results indicate low but significant species differentiation (Φ_{ST}= 0.10714, p = 0.030). Population differentiation is moderate, with a lower level of significance (Φ_{ST}= 0.48161, p = 0.092). However, an examination of the haplotype map reveals sharing of some haplotypes among *Quercus coccifera*, *Quercus alnifolia*, and *Quercus ilex*, attributable to phylogeographic stratification (Figure 34).



☑ Show the Network

Insert chunk size: 1000

Select a chunk: 10

Cluster Selector: 159

☑ Display Only Chosen

Cluster Color Palette: Spectral

Cluster-wise Contrast: 0.25

Figure 34. Haplotype map and parsimony network for the Cluster-159, illustrating moderate level of haplotype sharing between *Quercus coccifera*, *Quercus alnifolia*, and *Quercus ilex*.

Analysis can also be directed toward clusters incorporating a specific type of interest. Such clusters, featuring the selected type, can be readily identified by filtering through the AMOVA tables for targeted analysis (see Also Figure 24). For instance, upon examining Cluster-9, which includes *Quercus vulcanica*—an oak species endemic to Western and Central Anatolia in Turkey—we encounter a scenario similar to the one observed in Cluster-136. This cluster encompasses species from diverse groups: *Quercus vulcanica* from the Section Quercus, *Quercus ilex* and *Quercus coccifera* from the Section Ilex, and *Quercus brantii* from the Section Cerris. The admixture of these distinct lineages has led to notable interspecific divergence, as evidenced by the AMOVA results (Φ_{ST} = 0.52487, p < 0.001). In contrast to Cluster-136, the inter-population differentiation in Cluster-9 is markedly high and statistically significant, primarily due to the geographical separation of these distant species (Φ_{ST} = 0.87298, p < 0.001). To further elucidate the genetic dynamics within this cluster, Hapstep analysis was also conducted. Inspection of the Hapstep tables revealed that $N_{ST} < G_{ST}$. These differentiation parameters, highlighting the admixture event, suggest a disruption in the phylogeographic structure. The admixture within some populations, where rare haplotypes from different sub-generic clades come nearby, increases the variation synthetically, resulting in very low v_T compared to h_T.

Adjusting the initial parameters of clustering can enhance the capture of phylogeographic structure. Specifically, setting the k-mer threshold to 99 results in more phylogenetically homogeneous groups. This higher threshold narrows the clusters, increasing their number from 818 to 1184. Focusing on Cluster-7, previously Cluster-9, under this new configuration, where *Quercus vulcanica* is present, it is observed that only species from Section *Quercus* are included. This taxonomic narrowing significantly reduces interspecies divergence, though it remains statistically significant (Φ_{ST} = 0.12199, p < 0.001). Additionally, as the admixture decreases, thereby reducing variation within populations, the differentiation between populations shows a slight increase (Φ_{ST} = 0.92583, p < 0.001). N_{ST} increased from 0.784 to 0.927, become significant and exceeding G_{ST}, which reflects the phylogeographic structure and evident interspecific hybridizations within the cluster. v_T is also increased and represents more realistic metrics for genetic diversity.

Finally, to ensure - at least theoretically - the accuracy of the phylogeographic analysis, the *Taxonomic Pre-Grouping* is set to *Species,* k-mer threshold and coverage also lowered to 90%. This adjustment leads to an increase in the number of clusters to 1,718. However, due to the smaller sizes of these clusters, a limited number are included in the subsequent analysis. Upon merging the results of the Mantel test from IBD, AMOVA, and Hapstep, only 41 clusters are found to be common across all three outputs, as detailed in Table 1. At this stage, conducting a ΦST Barrier analysis facilitates both a broader inspection and a detailed examination of these clusters, which exhibit significant phylogeographic stratification.

The first line in Table 1 highlights Cluster-1336, which exhibits a high number of haplotypes. This cluster, comprising *Quercus serrata*, is one of the largest clusters in Japan due to extensive sampling, as can be seen in Figure 35. The ΦST Barriers specific to this cluster are displayed at the top of the figure, with the barriers for all clusters presented below. Such visualization allows for a direct comparison of the genetic variation within Cluster-1336 against the broader genetic landscape encompassing all clusters.

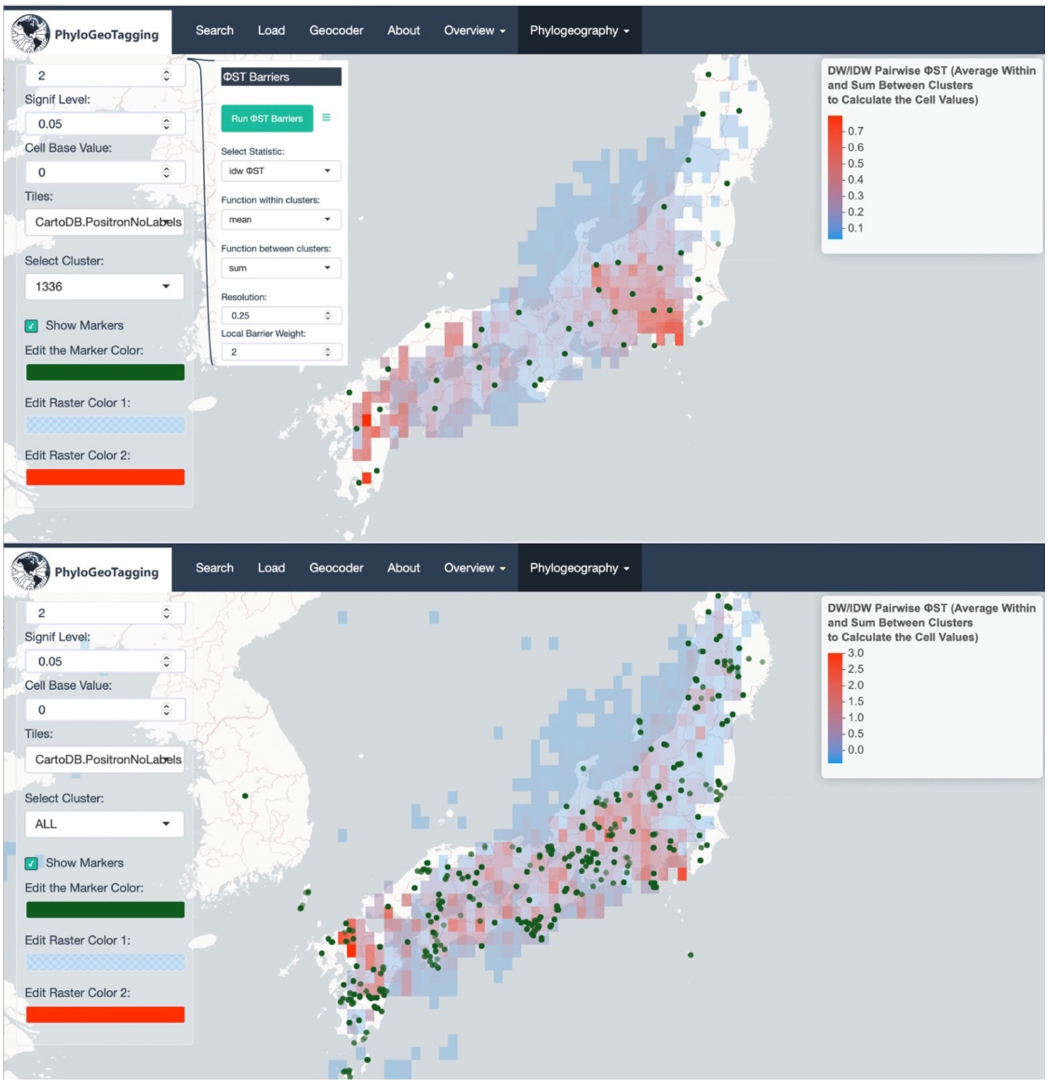
When examining the collective data from all clusters, the *Function Between Clusters* parameter is set to *sum*. This setting is crucial for identifying barriers that align at the same midpoint between different clusters. It effectively captures a superposition effect, revealing wider and overlapping phylogeographic signals. This approach can significantly elevate the idw Φ_{ST} value, as demonstrated in this example, where the highest value observed is around three. Such a high idw Φ_{ST} value indicates strong genetic differentiation at certain points, underscoring the tool's ability to detect nuanced phylogeographic patterns across the studied area.

Table 1. Merging the results of the Mantel test from IBD, AMOVA, and Hapstep for 41 clusters. Significant values are highlighted with colors. Expansions of abbreviations are given in the footer of the table.

Cluster	Species	hap	npop	harm	M r	M r p	Φ_{ST}	Φ_{ST} p	h_T	v_T	v_T exp	v_T p	G_{ST}	N_{ST}	N_{ST} exp	N_{ST} p
1336	*Q. serrata*	22	41	7.038	0.241	0	0.82	0	0.793	0.649	0.795	0.217	0.752	0.828	0.744	0.012
1114	*Q. mongolica*	16	4	7.416	0.439	0.04	0.37	0	0.877	0.97	0.878	0.754	0.258	0.479	0.25	0.003
32	*Q. coccifera*	12	4	4.565	0.088	0.13	0.49	0	0.979	1.003	0.981	0.683	0.151	0.434	0.146	0.04
1333	*Q. aliena*	7	13	5.747	0.125	0.12	0.94	0	0.806	0.749	0.8	0.283	0.837	0.884	0.839	0.021
932	*Q. serrata*	8	44	7.892	0.225	0	0.86	0	0.683	0.566	0.684	0.231	0.851	0.86	0.848	0.278
1465	*Q. petraea*	3	12	3.618	0.513	0	0.85	0	0.577	0.801	0.569	0.841	0.798	0.834	0.767	0.479
169	*Q. coccifera*	13	5	4.724	0.046	0.25	0.54	0	0.987	1.021	0.986	0.699	0.325	0.618	0.32	0.001
929	*Q. aliena*	4	13	5.747	0.172	0.04	0.91	0	0.706	0.473	0.708	0.094	0.9	0.855	0.895	0.924
1466	*Q. phillyraeoides*	5	2	5.833	0.479	0.02	0.37	0.011	0.914	0.914	0.914	0.942	0.229	0.229	0.229	0.138
18	*Q. serrata*	7	44	7.943	-0.04	0.64	0.84	0	0.415	0.38	0.414	0.433	0.819	0.84	0.814	0.156
38	*Q. ilex*	10	4	5.653	0.054	0.34	0.52	0	0.932	1.11	0.926	0.988	0.316	0.49	0.306	0.044
190	*Q. petraea*	2	12	3.429	0.083	0.02	0.72	0	0.527	0.527	0.527	1	0.71	0.71	0.71	1
1115	*Q. phillyraeoides*	14	15	7.612	-0.1	0.73	0.64	0	0.566	0.619	0.567	0.714	0.401	0.646	0.395	0.021
178	*Q. ilex*	12	4	6.15	-0.09	0.83	0.58	0	0.964	1.025	0.969	0.726	0.268	0.53	0.269	0.013
1240	*Q. mongolica*	10	3	10.39	0.771	0.21	0.27	0	0.691	0.416	0.687	0.037	0.262	0.192	0.246	0.741
802	*C. sativa*	6	2	11.82	-0.35	0.7	0.37	0.012	0.703	0.499	0.705	0.133	-0.06	0	-0.057	0.044
131	*Q. pubescens*	4	13	3.842	0.107	0.21	0.95	0	0.674	0.562	0.675	0.168	0.954	0.961	0.955	0.332
621	*Q. aliena*	3	13	5.747	0.097	0.15	0.9	0	0.452	0.367	0.453	0.328	0.801	0.817	0.803	0.328
129	*Q. petraea*	5	5	5	0.026	0.4	0.94	0	0.9	0.967	0.902	0.654	0.852	0.908	0.844	0.337
120	*Q. cerris*	4	4	3.429	-0.03	0.57	0.94	0	0.889	0.965	0.886	0.686	0.812	0.945	0.799	0.174
936	*Q. serrata*	3	7	5.144	0.155	0.15	0.85	0	0.32	0.245	0.323	0.323	0.872	0.875	0.86	0.672
1334	*Q. crispula*	2	7	7.396	0.332	0.12	0.83	0	0.245	0.245	0.245	1	0.833	0.833	0.833	1
624	*Q. serrata*	3	44	7.923	-0.01	0.53	0.94	0	0.424	0.337	0.426	0.161	0.931	0.935	0.93	0.692
931	*Q. mongolica*	2	7	7.84	0	0.39	1	0	0.571	0.571	0.571	1	1	1	1	1
46	*Q. phillyraeoides*	10	13	7.095	-0.06	0.68	0.67	0	0.651	0.572	0.65	0.442	0.604	0.638	0.6	0.167
127	*Q. libani*	3	3	3.83	-0.46	0.8	0.86	0	0.933	1	0.934	1	0.857	0.9	0.856	0.343
2	*Q. aliena*	3	13	5.747	-0.05	0.7	0.94	0	0.487	0.567	0.491	0.828	0.921	0.898	0.922	0.843
930	*Q. crispula*	3	10	7.467	-0.07	0.57	0.91	0	0.569	0.569	0.569	0.674	0.906	0.906	0.906	1
1673	*Q. mongolica*	11	3	10.36	0.184	0.19	0.28	0	0.752	0.761	0.754	0.524	0.288	0.292	0.285	0.476
128	*Q. macranthera*	2	2	4.444	-0.22	0.75	1	0	1	1	1	1	1	1	1	1
542	*Q. coccifera*	5	4	4.66	-0.05	0.75	0.61	0	0.867	0.734	0.861	0.152	0.606	0.62	0.598	0.419
1464	*Q. mongolica*	5	3	8.759	0.101	0.29	0.49	0	0.557	0.821	0.552	0.909	0.38	0.299	0.359	0.644
125	*Q. infectoria*	2	5	3.158	-0.06	0.67	0.85	0	0.6	0.6	0.6	1	0.778	0.778	0.778	1
122	*Q. frainetto*	2	2	3.429	0.073	0.47	0.79	0	0.667	0.667	0.667	1	0.5	0.5	0.5	1
660	*Q. phillyraeoides*	6	12	8.112	0.03	0.43	0.4	0	0.23	0.117	0.23	0.093	0.276	0.203	0.249	0.745
1373	*Q. phillyraeoides*	2	4	4.898	-0.19	0.73	0.58	0.002	0.583	0.583	0.583	1	0.714	0.714	0.714	1
626	*Q. laeta*	16	15	6.866	-0.21	0.91	0.19	0	0.474	0.303	0.474	0.045	0.188	0.204	0.186	0.325
644	*Q. laeta*	13	15	6.866	-0.15	0.81	0.25	0	0.683	0.377	0.678	0.002	0.314	0.275	0.313	0.78
1296	*Q. phillyraeoides*	4	7	4.565	-0.32	0.85	-0	0.617	0.385	0.297	0.385	0.048	-0	-0.01	-0.001	0.791
981	*C. sativa*	4	2	11.82	0.866	0.33	-0.1	0.616	0.237	0.197	0.242	0.379	0.147	0.097	0.136	0.761
552	*Q. ilex*	3	4	5.933	-0.08	0.71	0.04	0.452	0.569	0.458	0.568	0.343	0.385	0.264	0.376	1

hap: Number of haplotpes.
npop: Number of populations.
harm: Harmonic mean number of haplotypes in populations.
M r: Mantel r - value.
M r p: p -value of Mantel r - value.
Φ_{ST}: AMOVA fixation index for population differentiation.
Φ_{ST} p: p-value of Φ_{ST}.

h_T: Total diversity parameter.
v_T : Total diversity parameter (considers genetic distances).
v_T exp: Expected value (permutation average) of v_T.
v_T p: p-value of v_T.
G_{ST}: Population differentiation parameter.
N_{ST}: Population differentiation parameter (considers genetic distances).
N_{ST} exp: Expected (permutation average) of N_{ST}.
N_{ST} p: p-value of N_{ST}.



2
Signif Level: 0.05
Cell Base Value: 0
Tiles: CartoDB.PositronNoLabels
Select Cluster: 1336
☑ Show Markers
Select Statistic: idw ΦST
Function within clusters: mean
Function between clusters: sum
Resolution: 0.25
Local Barrier Weight: 2

2
Signif Level: 0.05
Cell Base Value: 0
Tiles: CartoDB.PositronNoLabels
Select Cluster: ALL
☑ Show Markers

Figure 35. ΦST Barrier analysis for *Quercus serrata* in Cluster-1336 and comparative view across all clusters. The top section shows the specific ΦST Barriers for Cluster-1336, one of the largest in Japan, while the lower section presents barriers for all clusters. This comparison highlights significant genetic differentiation, with the *Function Between Clusters* parameter set to sum, revealing superposition effects and high idw ΦST values.

The IDW in the ΦST Barrier analysis plays a significant role in highlighting local barriers, particularly accentuating genetic changes over shorter distances. However, it's essential to note that in scenarios of intensive sampling, such as that of Cluster-

1336, the ΦST Barrier analysis sometimes tends to reflect the genetic variation within the region more than the actual genetic barriers.

Users should consider that while ΦST Barrier analysis is a powerful tool for visualizing and understanding genetic differentiation, the context of sampling and parameter settings significantly influences the interpretation of results. Therefore, the outcomes of the ΦST Barrier analysis should be integrated with other phylogeographic data for a more comprehensive understanding of the genetic structures at play.

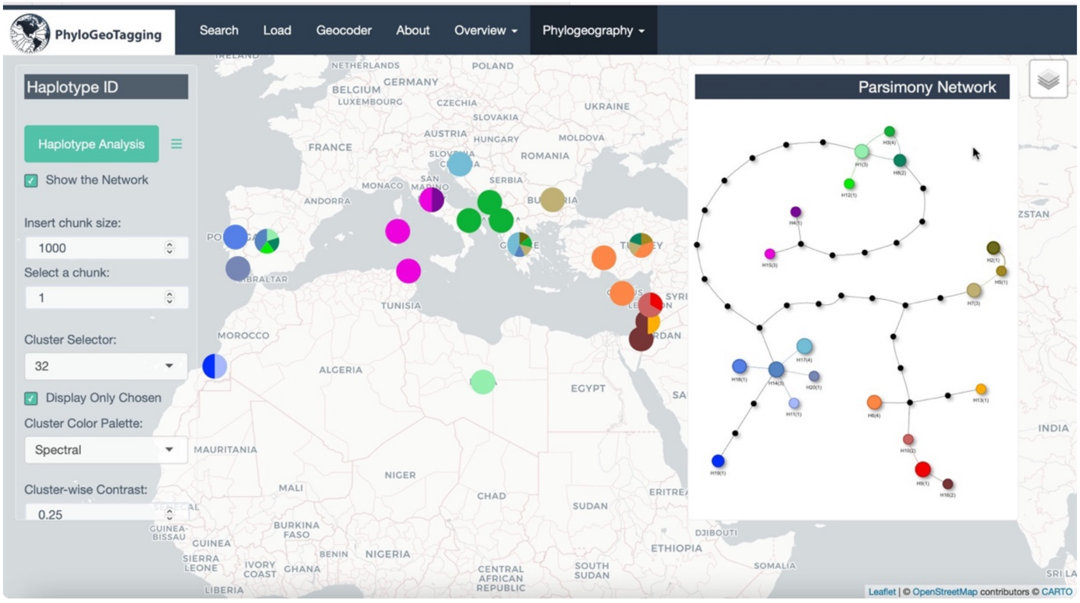
The next example investigated from the table is *Quercus coccifera*, represented in Cluster-32 at line 4. In this cluster, a notable number of mutational distances between haplotypes have been observed, as illustrated in the haplotype distribution and parsimony network in Figure 36**.**

This observation prompts the inquiry into how such genetically distant haplotypes could be grouped within the same cluster. It is important to note that at the beginning of this analysis, the k-mer threshold and coverage were lowered, which can lead to the grouping of more distant sequences within the same cluster.

This situation calls to mind a remarkable phenomenon: chloroplast capture. To further investigate this, it is beneficial to examine GenBank records for Cluster-32. This can be done using the *Hover Labels* from the control panel and selecting either the “Gene”, “Product” or “Definition” feature. A detailed inspection of these records reveals that the sequences belong to the trn-H-psbA region. Additionally, further examination of the “Title” feature facilitates the identification of a related study (Vitelli et al., 2017).

Chloroplast capture, a unique process in plant evolution, involves the acquisition of the chloroplasts of one species by another through hybridization and subsequent backcrossing. This typically occurs in plants where chloroplasts are maternally inherited, resulting in a plant that maintains its phenetic form and nuclear DNA traits of the main species, but possesses the chloroplast DNA of another species.

In the context of Cluster-32, evidence from additional studies including *Quercus coccifera*, which examined different regions of the chloroplast genome, supports the hypothesis of an ancient chloroplast capture events and other range expansions that form distinct lineages within the Section *Ilex* in the Mediterranean basin. These reticulate hybridizations occurred between different oak sections through the evolution of the genus, specifically, took place between a member of Section Cerris and Section Ilex (Simone et al., 2016; Tekpınar et al., 2021). This scenario could explain the high number of mutations observed among different haplotypes of same species in Cluster-32, despite the sequence's relatively short length of about 560 base pairs. Such findings underscore the complexity and evolutionary significance of chloroplast capture in plant genetics.



☑ Show the Network

Insert chunk size: 1000

Select a chunk: 1

Cluster Selector: 32

☑ Display Only Chosen

Cluster Color Palette: Spectral

Cluster-wise Contrast: 0.25

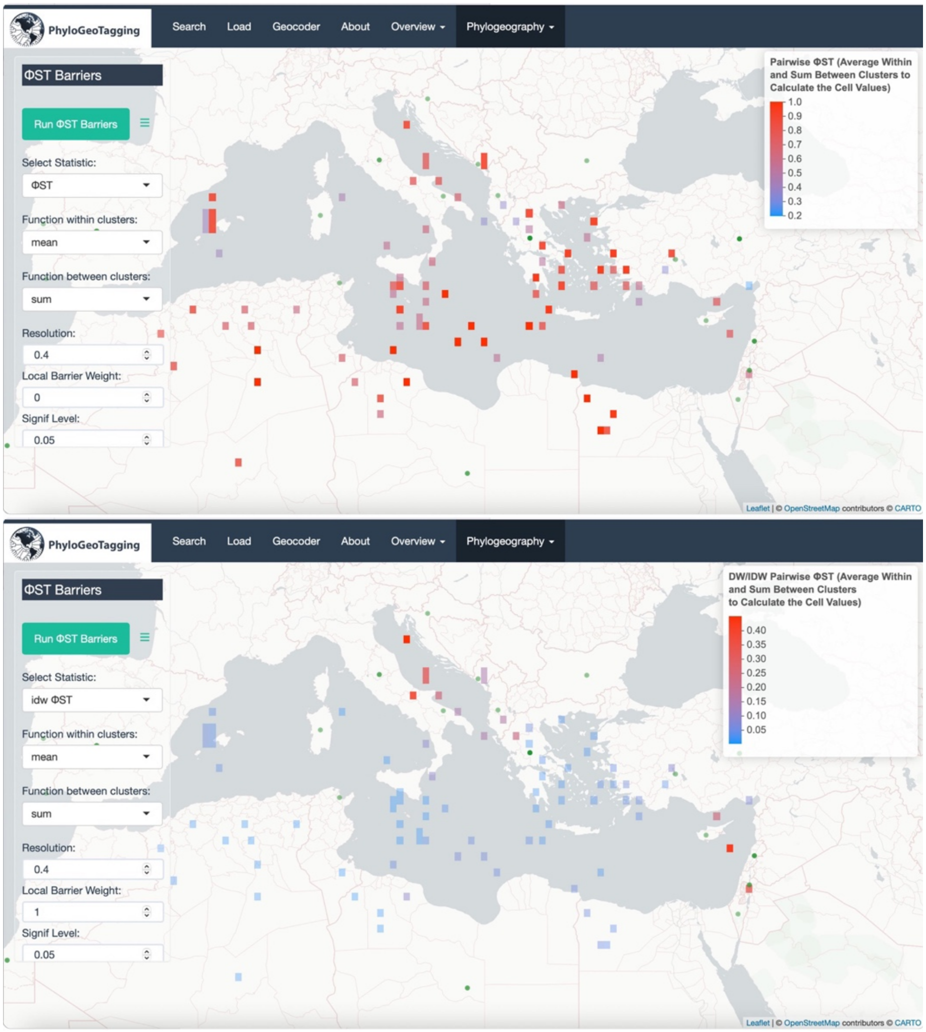
Figure 36. Haplotype distribution and parsimony network for *Quercus coccifera* in Cluster-32. This illustrates the high number of mutations between haplotypes, indicative of chloroplast capture events.

In Figure 37, the ΦST Barriers are examined, focusing on their characteristics under different visualization conditions. When Inverse Distance Weighting (IDW) is not applied, and raw ΦST barriers are visualized, it is observed that the strongest barriers, typically occurring between populations of different lineages, predominantly cluster in the Mediterranean Sea area. This pattern highlights the significant role of geographic separation, suggesting that lineages can exist without intermixing for extended periods. Such spatial isolation contributes to the distinct genetic identities and evolutionary paths of these lineages within the Mediterranean basin.

By contrast, when IDW is used to highlight local barriers, barriers arise from differentiation both between and within lineages. A detailed inspection of the haplotype maps and ΦST barriers maps in the East Mediterranean reveals that closely related haplotypes, such as H6 from Cyprus and H9 and H10 from Lebanon, create a local barrier within the lineage. Furthermore, more southern haplotypes, specifically H13 and H16, which are on the same branch of the parsimony network, form another barrier with the population where H9 and H10 are found.

Significant barriers are also observed around the Adriatic Sea, a region where different lineages meet in proximity. This observation suggests a complex interplay of genetic differentiation influenced by geographical and lineage factors, highlighting the intricate nature of phylogeographic patterns.

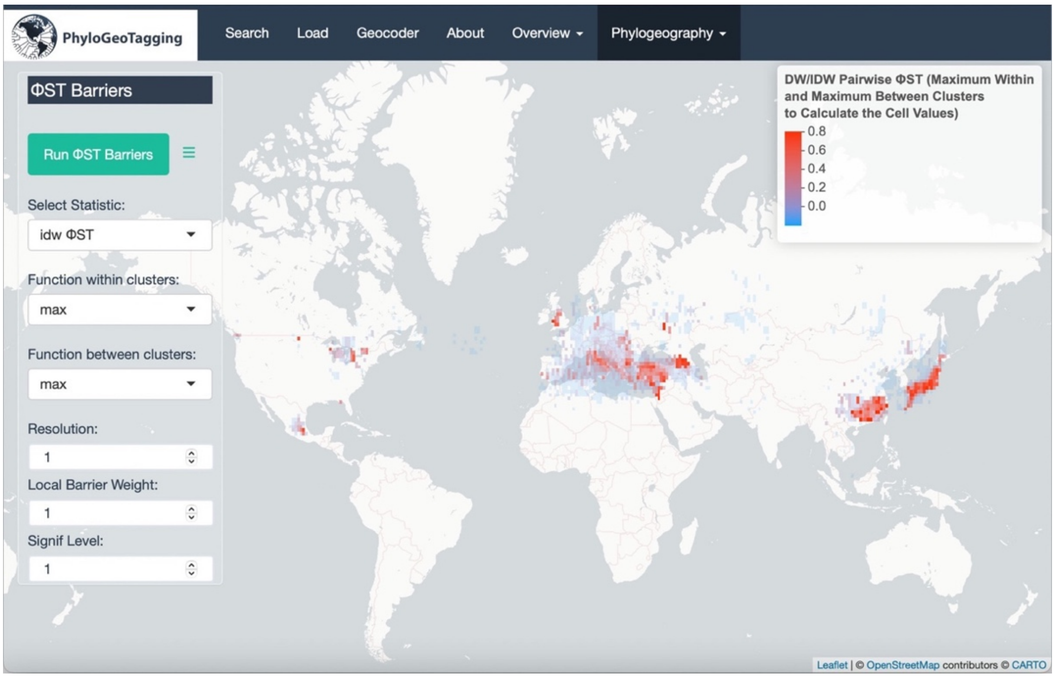
Finally, a global representation of ΦST Barriers is provided. This representation includes all clusters that are represented by two or more populations and serves to visualize the data available in GenBank regarding population differentiation within the Fagaceae family (Figure 38). Significance parameters have been disabled to highlight all potential barriers. Both the *Function within clusters* and *Function between clusters* parameters are set to *max* to capture the maximum values coinciding in each raster cell. This approach is adopted to balance the effects of sampling differences as much as possible. While IDW effectively highlights local barriers, it can also be utilized from another perspective to depict genetic hotspots on the world map, offering a comprehensive view of genetic diversity and differentiation across the globe.


PhyloGeoTagging
Search
Load
Geocoder
About
Overview
Phylogeography
ΦST Barriers
Run ΦST Barriers
Pairwise ΦST (Average Within and Sum Between Clusters to Calculate the Cell Values)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Leaflet | © OpenStreetMap contributors © CARTO
PhyloGeoTagging
Search
Load
Geocoder
About
Overview
Phylogeography
ΦST Barriers
Run ΦST Barriers
DW/IDW Pairwise ΦST (Average Within and Sum Between Clusters to Calculate the Cell Values)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
Leaflet | © OpenStreetMap contributors © CARTO

Select Statistic: ΦST
Function within clusters: mean
Function between clusters: sum
Resolution: 0.4
Local Barrier Weight: 0
Signif Level: 0.05

Select Statistic: idw ΦST
Function within clusters: mean
Function between clusters: sum
Resolution: 0.4
Local Barrier Weight: 1
Signif Level: 0.05

Figure 37. ΦST Barrier analysis in the Mediterranean showing the effects of chloroplast capture. The top section displays raw ΦST Barriers, primarily highlighting cross-lineage barriers. In the bottom section, which uses Inverse Distance Weighting (IDW), local barriers within lineages become apparent. Both sections collectively underscore the complex interplay of geography and lineage in genetic differentiation.


PhyloGeoTagging
Search
Load
Geocoder
About
Overview
Phylogeography
ΦST Barriers
Run ΦST Barriers
DW/IDW Pairwise ΦST (Maximum Within and Maximum Between Clusters to Calculate the Cell Values)
0.8
0.6
0.4
0.2
0.0
Leaflet | © OpenStreetMap contributors © CARTO

Select Statistic: idw ΦST

Function within clusters: max

Function between clusters: max

Resolution: 1

Local Barrier Weight: 1

Signif Level: 1

Figure 38. Global ΦST Barriers for the Fagaceae Family. This figure demonstrates how Inverse Distance Weighting (IDW) not only highlights local barriers but also reveals genetic hotspots worldwide. It provides a comprehensive view of genetic diversity and differentiation across various global regions, showcasing the intricate patterns within the Fagaceae family.

3.1.2. Example: Genus *Salvia*

The next example explored the genus *Salvia*, which represents the largest genus in the Lamiaceae family. Encompassing over 900 species, *Salvia* species are distributed worldwide, presenting a rich diversity.

An extensive search for chloroplast DNA sequences (cpDNA) of the Genus *Salvia* was conducted using the search terms “Salvia [ORGANISM] AND (chloroplast[filter] OR plastid[filter]).” This search resulted in a total of 7,131 sequence records. After thorough parsing and subsequent geocoding, a dataset of 4,263 sequences and their metadata, was obtained. Totally, 362 unique organism names were found in the dataset, representing the *Salvia* genus (Figure 39). The majority of these, amounting to 310 entries, were identified as distinct species. In addition, there were 5 subspecies. The dataset also included 33 varieties and 7 form names. Among these, there were 6 ambiguous species, indicating cases where specific species identification remains uncertain, denoted by using the "sp." abbreviation in organism names. The dataset contained a single record marked as a hybrid.

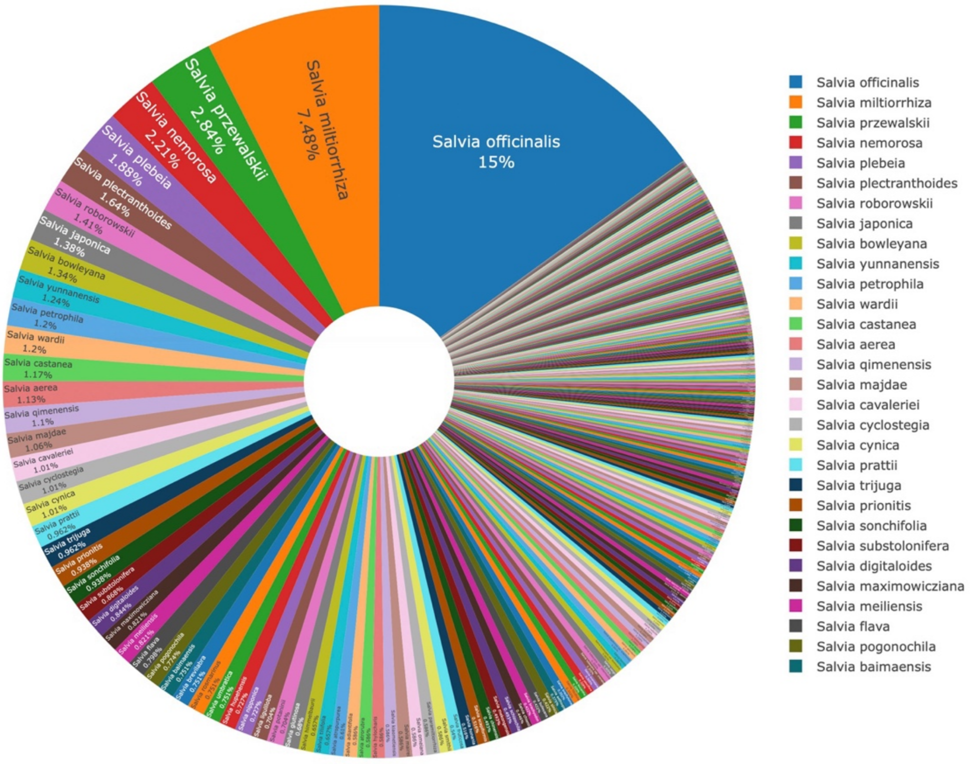


Figure 39. Donut chart representation of unique organism names in the *Salvia* dataset. The legend shows the first 30 of 362 organism names.

First five largest sampling countries are China, Mexico, Croatia, Japan, and Italy. First 10 country compromise more than 90% of unique species available for cpDNA sequences in GenBank (Figure 40).

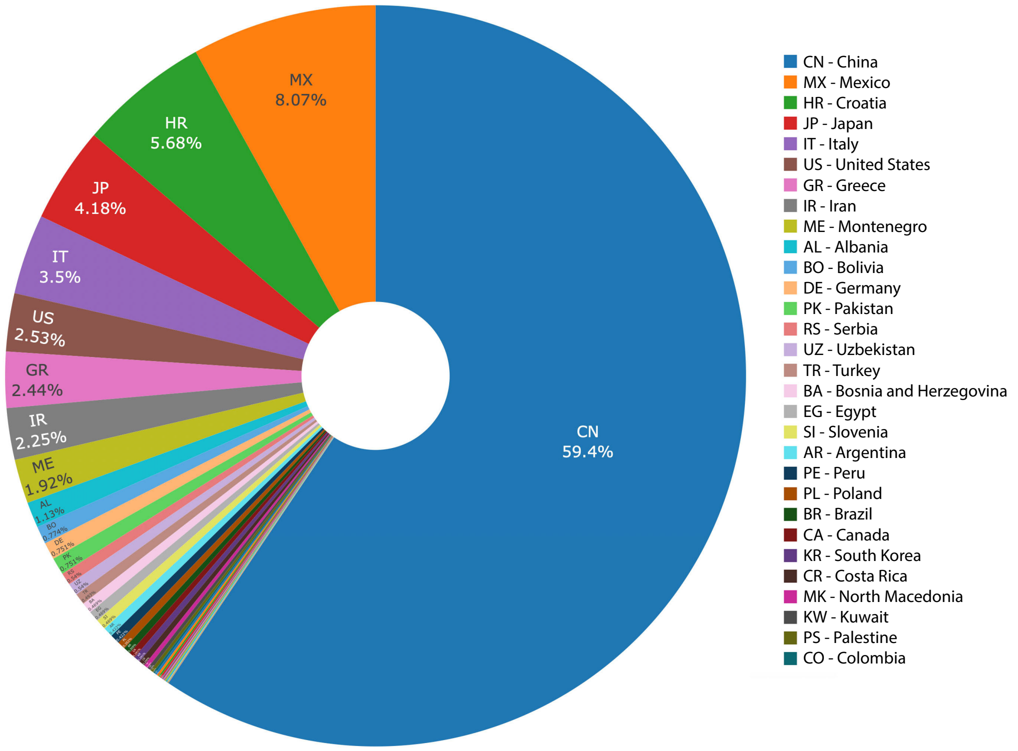
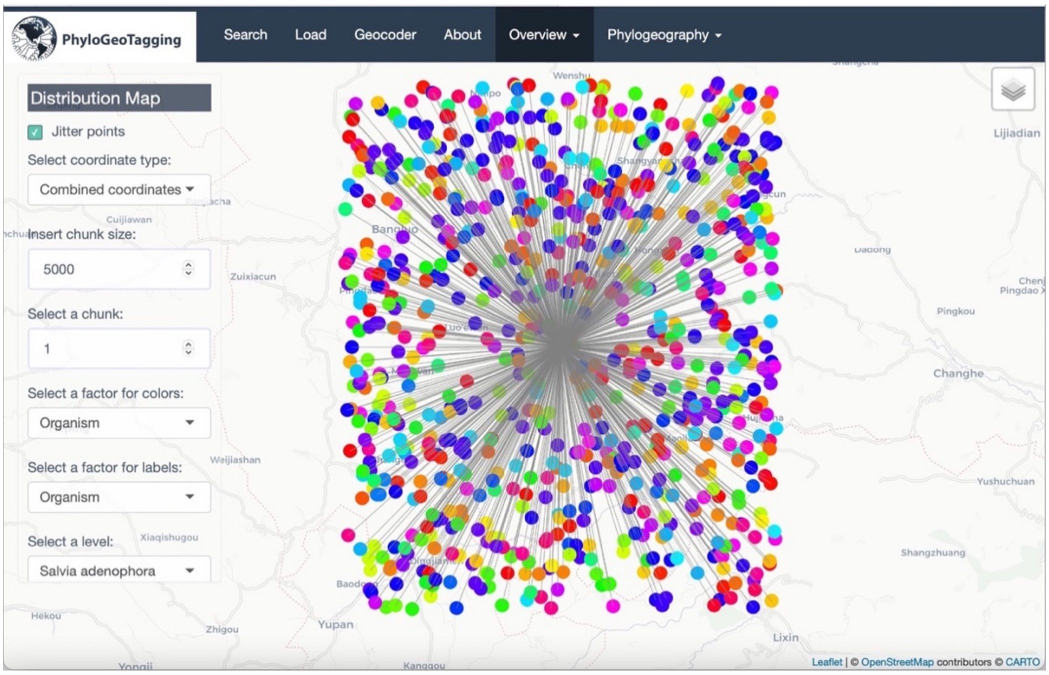


Figure 40. Donut chart of cpDNA sequence sampling by country. The legend lists the first 30 countries.

Similar to the approach taken with the Fagaceae family, both unrestricted and species-based clustering methods were applied to the *Salvia* dataset. Given that the search was confined to sequences from the genus *Salvia*, selecting *Unrestricted* is equivalent to setting the *Taxonomic Pre-Grouping* parameter to the *Genus*. To maintain a high level of accuracy and to prevent the inclusion of unrelated sequences from different taxa in the same cluster, both the k-mer similarity threshold and the coverage threshold were initially set at a stringent 99%, and 98%, respectively. In other run, *Taxonomic Pre-Grouping* parameter is set to *Species,* and k-mer threshold and coverage lowered to 90%.

In the first run of the clustering, 1,179 clusters formed in various sizes. Initial insights show that the *Salvia* dataset, when unrestrictedly aligned with strict similarity parameters, does not produce appropriate data for phylogeographical analysis. A quick look at the Hapstep tables reveals that none of these clusters show significant phylogeographical signals. N_{ST} is larger than G_{ST} in only two clusters, Cluster-11 and Cluster-116, yet not statistically significant. One reason for the non-availability of the data for phylogeographic analysis is the absence of fine-level geographic coordinates for many clusters. These clusters are assigned to a single location, the country's center coordinates. These single-location clusters cannot be used for phylogeographic analysis. This situation is clearly seen in Figure 41, where jittered points show the center coordinate of China. To inspect the issue, China, which is the largest country contributing to the *Salvia* dataset, is filtered using the Geocoder table. All available 2,534 records are represented in 68 locations, but already 70.8% of them don't have a detailed address in GenBank, just country names. Incorporating "lat-lon" data available in GenBank records contributes to increasing the precision by 3.2%. However, some available detailed addresses don't return a result in geocoding, so they are tagged at the county scale. In sum, the available locations in China decreased slightly, and 71.6% of them are assigned to the center coordinate for the country. A similar problem emerges from other clusters as well. As a result, Hapstep analysis can only be applied to 15 clusters out of 1,179, due to the small number of available populations.

Since the existing data is not suitable for phylogeographic analysis, it would be better to focus on species separation using AMOVA. The Analysis of Molecular Variance (AMOVA), presented in Table 2, offers a comprehensive review of the *Salvia* dataset on a global scale. However, it underscores a previously mentioned difficulty in applying phylogeographical analysis to this dataset. The representation of clusters by single locations inherently limits the effectiveness of population differentiation in AMOVA. This situation is vividly illustrated in the contrast between the columns SΦ_{ST} and Sp versus PΦ_{ST} and Pp in the table. The SΦ_{ST} values, which represent species differentiation in AMOVA Φ_{ST}, provide more abundant data compared to the PΦ_{ST} values that signify population differentiation. For better visualization and focus on species differentiation, of the 144 results, 88, which are statistically significant for



Distribution Map

☑ Jitter points

Select coordinate type: Combined coordinates

Insert chunk size: 5000

Select a chunk: 1

Select a factor for colors: Organism

Select a factor for labels: Organism

Select a level: Salvia adenophora

Figure 41. Visualization of cluster locations with jittered points indicating the central coordinates of China, highlighting the challenge of missing fine level geographic data for phylogeographic analysis.

species differentiation, are shown. While most significant results for population differentiation are also significant for species differentiation and included in the table, there are only two clusters, clusters 90 and 108, which are significant solely for population differentiation not represented in the table. From a broader perspective, the *Salvia* species exhibit a fair level of differentiation when multiple species are compared on a worldwide basis using AMOVA. Although species differentiation apparent in the dataset suggests that the *Salvia* species have diverged to some extent, these findings should be approached with caution (see the Discussion Section for more information).

Table 2. AMOVA results for the *Salvia* dataset. The $S\Phi_{ST}$ represents species differentiation, and $P\Phi_{ST}$ denotes population differentiation. Sp and Pp are the respective p-values.

Cluster	Countries	Taxa	S*ΦST*	S*p*	P*ΦST*	P*p*
20	China	*S. przewalskii, S. prattii, S. nipponica, S. kiaometiensis, S. hylocharis, S. glutinosa, S. flava, S. digitaloides, S. cyclostegia, S. castanea, S. bulleyana, S. atrorubra, S. aerea*	1	0		
30	China	*S. smithii, S. roborowskii, S. potaninii, S. pogonochila, S. omeiana, S. maximowicziana, S. cynica, S. brevilabra*	1	0		
35	Mexico, China	*S. semiatrata, S. scapiformis, S. qimenensis, S. prionitis, S. plectranthoides, S. miltiorrhiza, S. japonica, S. chienii, S. cavaleriei, S. bowleyana*	1	0	0,647	0
43	China	*S. stibali, S. tricuspis, S. smithii, S. roborowskii, S. przewalskii* var. *przewalskii, S. pogonochila, S. paohsingensis, S. omeiana, S. maximowicziana* var. *maximowicziana, S. maximowicziana* var. *floribunda, S. kiaometiensis, S. heterochroa, S. evansiana* var. *scaposa, S. castanea* f. *castanea*	1	0	-0,278	0,853
48	China	*S. scapiformis* var. *scapiformis, S. prionitis, S. plectranthoides, S. nanchuanensis* var. *pteridifolia, S. miltiorrhiza, S. kiangsiensis, S. honania, S. dabieshanensis, S. chunganensis, S. chinensis, S. cavaleriei* var. *simplicifolia, S. subbipinnata, S. bowleyana* var. *bowleyana*	1	0	0,184	0,101
52	Mexico, China, Bolivia, Brazil	*S. tubifera, S. tiliifolia, S. oxyphora, S. miniata, S. leucantha, S. karwinskii, S. iodantha, S. guaranitica, S. fallax, S. curviflora, S. adenophora*	1	0	0,244	0,097
62	China	*S. chinensis, S. sinica, S. qimenensis, S. prattii, S. miltiorrhiza, S. japonica, S. chienii*	1	0		
63	China	*S. paramiltiorrhiza, S. miltiorrhiza, S. meiliensis, S. honania, S. bowleyana, S. baimaensis*	1	0		
74	China	*S. wardii, S. przewalskii, S. hupehensis, S. evansiana, S. cyclostegia, S. castanea*	1	0		
82	China	*S. wardii, S. przewalskii, S. evansiana, S. cyclostegia, S. castanea*	1	0		
90	China	*S. wardii, S. przewalskii, S. evansiana, S. cyclostegia, S. castanea*	1	0		
97	China	*S. kiaometiensis, S. flava, S. castanea, S. bulleyana, S. aerea*	1	0		
98	China	*S. kiaometiensis, S. flava, S. castanea, S. bulleyana, S. aerea*	1	0		
101	China, Mexico	*S. tiliifolia, S. semiatrata, S. miniata, S. fallax, S. curviflora, S. adenophora*	1	0	0,196	0,078
104	China, Mexico	*S. tiliifolia, S. miniata, S. leucantha, S. iodantha, S. fallax, S. curviflora*	1	0	0,265	0,053
107	China	*S. wardii, S. sikkimensis, S. prattii, S. flava, S. digitaloides, S. chienii, S. campanulata* var. *codonantha*	1	0	0,61	0,163
110	Mexico	*S. filipes, S. tiliifolia, S.* sp. *Ledesma 20366, S. xalapensis, S. roscida*	1	0		
112	China, Mexico	*S. tiliifolia, S. leucantha, S. iodantha, S. fallax, S. curviflora*	1	0	0,368	0,028
114	Japan, United States	*S. nipponica* var. *kisoensis, S. isensis, S. japonica, S. nipponica* var. *trisecta, S. miltiorrhiza*	1	0		
115	Turkey	*S. pseudeuphratica, S. sericeotomentosa* var. *hatayica, S. sericeotomentosa* var. *sericeotomentosa, S. kronenburgii, S. cerinopruinosa, S. euphratica* var. *leiocalycina, S. euphratica* var. *euphratica*	1	0		
116	China, Mexico	*S. tiliifolia, S. leucantha, S. iodantha, S. fallax, S. curviflora*	1	0	-0,091	0,898
118	Mexico, China	*S. iodantha, S. tiliifolia, S. leucantha, S. fallax, S. curviflora*	1	0	-0,091	0,707
132	Turkey	*S. pseudeuphratica, S. sericeotomentosa* var. *hatayica, S. sericeotomentosa* var. *sericeotomentosa, S. kronenburgii, S. cerinopruinosa, S. euphratica* var. *euphratica*	1	0		
145	China	*S. trijuga, S. plebeia*	1	0		
151	Greece, Uzbekistan, Canada	*S. officinalis, S. spinosa, S. verticillata*	1	0	1	0
154	China	*S. mekongensis, S. mairei*	1	0		
169	China	*S. schizochila, S. digitaloides, S. castanea*	1	0		
170	China	*S. pogonochila, S. hupehensis*	1	0		
171	China	*S. officinalis, S. nemorosa, S. grandifolia*	1	0		
174	China, Japan	*S. plectranthoides, S. lutescens* var. *lutescens, S. lutescens* var. *stolonifera, S. pygmaea* var. *simplicior, S. omerocalyx* var. *prostrata*	1	0	0,679	0,1

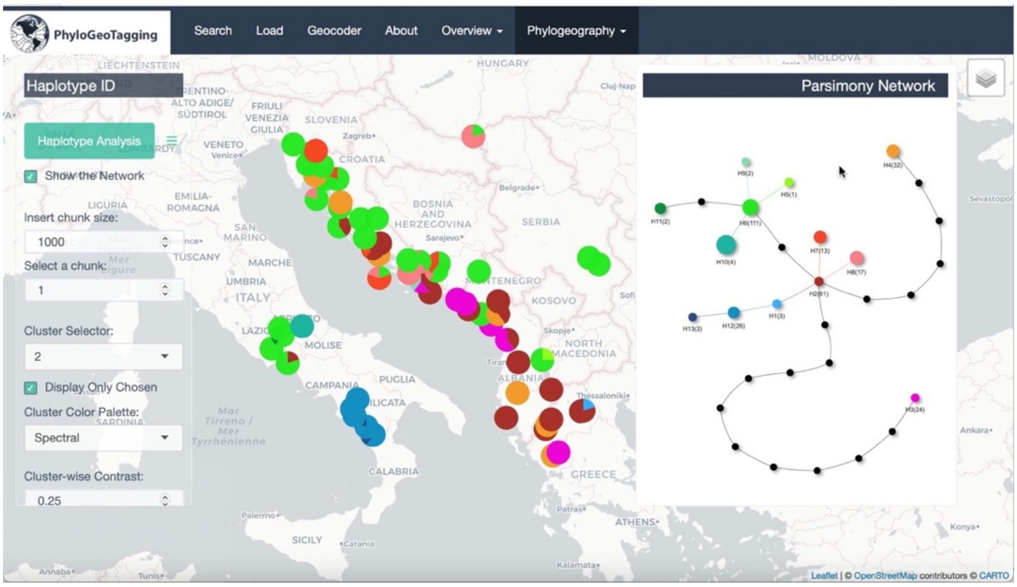
Table 2. Continued

Cluster	Countries	Taxa	S*ΦST*	S*p*	P*ΦST*	P*p*
179	Bolivia, Argentina	*S. retinervia, S. cuspidata* subsp. *rosea, S. cuspidata* subsp. *gilliesii, S. calolophos*	1	0		
183	China	*S. umbratica, S. maximowicziana*	1	0		
186	China	*S. prionitis, S. bowleyana, S. baimaensis*	1	0		
203	China	*S. przewalskii, S. mairei*	1	0		
207	China	*S. maximowicziana, S. hupehensis*	1	0		
213	China	*S. przewalskii, S. mairei*	1	0		
222	Japan	*S. japonica, S. isensis*	1	0	-0,667	0,332
223	Japan	*S. x sakuensis, S. glabrescens* var. *repens, S. glabrescens* var. *glabrescens*	1	0		
226	Greece	*S. pomifera, S. fruticosa*	1	0	1	0
236	China	*S. evansiana, S. cyclostegia*	1	0		
240	China	*S. paramiltiorrhiza, S. bowleyana*	1	0		
252	Mexico	*S. microphylla, S. roscida*	1	0		
263	China	*S. cyclostegia, S. castanea*	1	0		
281	Mexico	*S. roscida, S. longispicata*	1	0		
282	Egypt, Costa Rica	*S. splendens, S. lasiocephala*	1	0	1	0
290	China	*S. trijuga, S. plebeia*	1	0		
11	Croatia, Montenegro, Greece, Albania	*S. brachyodon, S. officinalis*	0,953	0	0,88	0
41	China	*S. wardii, S. przewalskii, S. potaninii, S. omeiana, S. maximowicziana, S. cynica, S. castanea*	0,929	0		
51	China	*S. przewalskii, S. prattii, S. kiaometiensis, S. flava, S. bulleyana, S. brachyloma, S. aerea*	0,902	0		
53	China	*S. trijuga, S. substolonifera, S. plebeia*	0,88	0		
289	China	*S. trijuga, S. substolonifera*	0,846	0	0,846	0
40	United States, Pakistan, Canada, Egypt, South Africa, United Kingdom	*S. hierosolymitana, S. moorcroftiana, S. rosmarinus, S. sclarea, S. pratensis, S. verticillata, S. officinalis, S. viridis, S. lanceolata, S. verbenaca*	0,778	0,001	0,484	0,05
160	United States, Italy, Pakistan	*S. yangii, S. rosmarinus*	0,714	0		
259	China	*S. wardii, S. campanulata* var. *campanulata*	0,714	0		
10	China	*S. umbratica, S. smithii, S. roborowskii, S. przewalskii, S. potaninii, S. pogonochila, S. omeiana, S. maximowicziana, S. himmelbaurii, S. cynica, S. brevilabra, S. aerea*	0,71	0		
6	China	*S. wardii, S. umbratica, S. smithii, S. roborowskii, S. przewalskii, S. potaninii, S. pogonochila, S. pauciflora, S. omeiana, S. maximowicziana, S. hupehensis, S. himmelbaurii, S. evansiana, S. cynica, S. cyclostegia, S. castanea, S. brevilabra, S. aerea*	0,696	0		
7	China, Mexico	*S. wardii, S. schizochila, S. przewalskii, S. prattii, S. plectranthoides, S. mekongensis, S. mairei, S. madrensis, S. lankongensis, S. kiaometiensis, S. hylocharis, S. flava, S. digitaloides, S. cyclostegia, S. cavaleriei, S. castanea, S. bulleyana, S. brachyloma, S. atrorubra, S. atropurpurea, S. aerea*	0,693	0	-0,986	0,366
49	Japan, China	*S. lutescens* var. *intermedia, S. hayatana, S. pygmaea, S. lutescens* var. *lutescens, S. japonica, S. ranzaniana, S. lutescens* var. *crenata, S. isensis, S. japonica* var. *japonica*	0,676	0	-0,573	0,758
100	China	*S. miltiorrhiza, S. meiliensis, S. honania, S. bowleyana, S. baimaensis*	0,669	0,027		
165	China, Italy, Kuwait	*S. deserta, S. aethiopis, S. spinosa*	0,667	0	0,667	0
164	China	*S. qimenensis, S. chienii, S. cavaleriei* var. *cavaleriei, S. baimaensis*	0,571	0	-0,429	0,394
68	China	*S. tricuspis, S. przewalskii* var. *przewalskii, S. pogonochila, S. maximowicziana* var. *maximowicziana, S. maximowicziana* var. *floribunda, S. kiaometiensis, S. heterochroa, S. evansiana* var. *scaposa, S. cynica, S. castanea* f. *castanea*	0,546	0,042	-0,298	0,918
99	China	*S. sonchifolia, S. petrophila*	0,538	0		
121	China, Mexico	*S. nipponica, S. glutinosa*	0,538	0	0,695	0

Table 2. Continued

Cluster	Countries	Taxa	SΦ*ST*	S*p*	PΦ*ST*	P*p*
166	China	*S. wardii, S. mairei, S. castanea* f. *tomentosa, S. castanea* f. *glabrescens, S. atropurpurea*	0,526	0	0,25	0,118
26	China	*S. miltiorrhiza, S. umbratica, S. cynica, S. brevilabra*	0,519	0,024	0,484	0,008
12	China	*S. wardii, S. schizochila, S. przewalskii, S. prattii, S. lankongensis, S. hylocharis, S. flava, S. digitaloides, S. cyclostegia, S. castanea, S. brachyloma, S. atrorubra, S. atropurpurea, S. aerea*	0,517	0		
3	China, Mexico	*S. wardii, S. umbratica, S. sonchifolia, S. smithii, S. schizochila, S. roborowskii, S. qimenensis, S. przewalskii, S. prionitis, S. prattii, S. potaninii, S. pogonochila, S. plectranthoides, S. pauciflora, S. omeiana, S. nipponica, S. nanchuanensis, S. mekongensis, S. meiliensis, S. maximowicziana, S. lankongensis, S. japonica, S. hylocharis, S. hupehensis, S. himmelbaurii, S. glutinosa, S. flava, S. digitaloides, S. daiguii, S. cynica, S. cyclostegia, S. cavaleriei, S. castanea, S. bulleyana, S. brevilabra, S. brachyloma, S. bowleyana, S. baimaensis, S. atrorubra, S. atropurpurea, S. aerea*	0,516	0	-0,322	0,825
13	China	*S. wardii, S. schizochila, S. przewalskii, S. prattii, S. mekongensis, S. lankongensis, S. hylocharis, S. digitaloides, S. cyclostegia, S. castanea, S. brachyloma, S. atrorubra, S. atropurpurea, S. aerea*	0,512	0		
17	China	*S. umbratica, S. smithii, S. roborowskii, S. przewalskii, S. pogonochila, S. omeiana, S. maximowicziana, S. hupehensis, S. himmelbaurii, S. cynica, S. brevilabra*	0,511	0		
32	China	*S. sonchifolia, S. prionitis, S. plectranthoides, S. paramiltiorrhiza, S. nanchuanensis, S. miltiorrhiza, S. japonica, S. filicifolia, S. daiguii, S. cavaleriei, S. aerea, S. adiantifolia*	0,506	0,006		
28	China	*S. smithii, S. roborowskii, S. przewalskii, S. pogonochila, S. himmelbaurii, S. cynica, S. brevilabra*	0,505	0		
208	China	*S. przewalskii, S. mairei*	0,5	0		
15	China	*S. sinica, S. scapiformis, S. qimenensis, S. prionitis, S. plectranthoides, S. miltiorrhiza, S. liguliloba, S. japonica, S. chinensis, S. chienii, S. cavaleriei, S. bowleyana, S. baimaensis*	0,5	0,004		
33	China, Mexico	*S. schizochila, S. przewalskii, S. prattii, S. mekongensis, S. madrensis, S. flava, S. digitaloides, S. castanea, S. bulleyana, S. aerea*	0,493	0,025	0,626	0
123	China, Mexico	*S. nipponica, S. glutinosa*	0,489	0	0,378	0,041
96	China	*S. sonchifolia, S. petrophila*	0,455	0		
23	China	*S. yunnanensis, S. scapiformis, S. qimenensis, S. prionitis, S. plectranthoides, S. paramiltiorrhiza, S. miltiorrhiza, S. liguliloba, S. japonica, S. chienii, S. cavaleriei, S. bowleyana*	0,447	0,007		
78	China	*S. przewalskii, S. miltiorrhiza, S. tricuspis*	0,426	0,047	-0,435	0,689
111	Pakistan, Poland, Italy, Kuwait	*S. lanata, S. yangii, S. aethiopis, S. spinosa, S. aegyptiaca*	0,406	0,013	0,048	0,282
14	China	*S. sinica, S. scapiformis, S. qimenensis, S. prionitis, S. plectranthoides, S. paramiltiorrhiza, S. miltiorrhiza, S. liguliloba, S. japonica, S. honania, S. chinensis, S. chienii, S. cavaleriei, S. bowleyana, S. baimaensis, S. adiantifolia*	0,393	0,002		
67	Peru	*S. tubiflora, S. paposana, S. rhombifolia, S. formosa, S.* sp. *Peru18582, S.* sp. *Peru170164*	0,382	0,05		
9	China	*S. umbratica, S. smithii, S. roborowskii, S. przewalskii, S. potaninii, S. pauciflora, S. omeiana, S. maximowicziana, S. hupehensis, S. himmelbaurii, S. cynica, S. brevilabra, S. aerea*	0,34	0,018		
5	China, Mexico	*S. sonchifolia, S. sinica, S. semiatrata, S. scapiformis, S. qimenensis, S. prionitis, S. plectranthoides, S. paramiltiorrhiza, S. nanchuanensis, S. miltiorrhiza, S. meiliensis, S. liguliloba, S. japonica, S. honania, S. filicifolia, S. daiguii, S. chinensis, S. chienii, S. cavaleriei, S. bowleyana, S. baimaensis, S. adiantifolia*	0,334	0,001	-0,85	0,242
262	China	*S. qimenensis, S. bowleyana*	0,333	0		
163	Egypt	*S. officinalis, S. multicaulis, S. viridis*	0,3	0		
24	China	*S. wardii, S. przewalskii, S. prattii, S. lankongensis, S. hylocharis, S. flava, S. digitaloides, S. cyclostegia, S. castanea, S. brachyloma, S. atrorubra, S. atropurpurea, S. aerea*	0,284	0,026		
292	Japan	*S. lutescens* var. *crenata, S. omerocalyx*	0,2	0		

This high level of variance partitioning between species may indicate restricted gene flow between these species. In this case, selecting a high k-mer and coverage threshold is a good strategy to avoid including genetically isolated entities within clusters. However, this strictness also means losing some potential haplotypes within species, as well as the categorization of identical gene regions into different clusters due to polymorphisms within species. To handle this, Taxonomic Pre-Grouping is set to "Species", and the k-mer threshold and coverage is lowered to 90%. This alteration increases the total number of clusters to 1,714. By decreasing similarity restrictions and constraining groups to species as a fully reproductive unit, at least one significant phylogeographic group is identified. Hapstep analysis, which can still only be applied to very few clusters, returns significant N_{ST} in Cluster-2. The original study of Cluster-2 sequences, which examined chloroplast DNA of *Salvia officinalis* (common sage), focused on resolving its phylogeographic structure and evolutionary history (Radosavljević et al., 2022). It is reported that genetic differentiation was assessed using G_{ST} and N_{ST}, with N_{ST} values being significantly higher, indicating a phylogeographic structure. The results suggest the presence of ancient populations, colonization events, and the role of microrefugia in the evolutionary history of *Salvia officinalis*. Similar outcomes were automatically replicated with this package using sequences submitted to GenBank. Haplotypes are geographically stratified, N_{ST} is significantly higher than G_{ST}, reflecting phylogeographic patterns (Figure 42).



☑ Show the Network

Insert chunk size: 1000

Select a chunk: 1

Cluster Selector: 2

☑ Display Only Chosen

Cluster Color Palette: Spectral

Cluster-wise Contrast: 0.25

Figure 42. Haplotype distributions and parsimony network for Cluster-2, reflecting the phylogeographic patterns of *Salvia officinalis*.

Clustering is a crucial step for automated phylogeographic analysis using vast amounts of sequence data obtained through GenBank. However, focusing on large datasets can lead to some data being overlooked if detailed analysis is the main target. Neither similarity thresholds nor taxonomic constraints may fail to group the correct set of sequences from a given study, particularly apparent within-lineage variation represented by different species. Radosavljević et al., 2022 reported 16 haplotypes detected, however, here only 13 haplotypes are achieved in Cluster-2. Although this dataset comprises a single species, the risk of overlooking is very low with the “Species” grouping, further inspection can be beneficial to understand this discrepancy. To quickly investigate this issue and assess the efficacy of clustering, one can filter the Geocoder table to specifically retrieve records relevant to the study of interest. The total number of entries for this study is 598. When these subsets of sequences are re-clustered with highly relaxed parameters (Unrestricted taxonomic grouping, k-mer, and coverage thresholds 70%), one large cluster occurs with a single species, *Salvia officinalis*, and in this cluster, a parsimony network shows two groups of haplotypes separated from each other by 349 mutations. In a way that leaves no room for doubt, they are representing two different sequence regions. Increasing k-mer and coverage thresholds to 80% results in correct clustering, where two different groups are partitioned into their associated clusters. Results are the same as the initial analysis, with 13 and nine haplotypes attained for Cluster-1 and Cluster-2, respectively. Further reading of the methods applied in the inspected study reveals that the indel-coding method they used created this discrepancy.

Focusing on a specific country can also be a popular choice among users. For instance, filtering the country_code column of the Geocoder table using the isocode “TR” yields only the results from Turkey (Table 3). The returned records point out that there is only a single study that contributes to the current *Salvia* dataset (Dizkirici et al., 2015). Turkey is characterized by a high level of *Salvia* diversity, representing almost 12% of all species worldwide and exhibiting a high endemism rate of 55% (Celep & Doğan, 2023). However, the number of cpDNA sequences from Turkey submitted to GenBank is disproportionately low.

Table 3. Filtered geocoding results showcasing data from Turkey. The country_code column is refined using the ISO code “TR” to display *Salvia* dataset specific to the Turkey.

Exported data										
Row	Accession Numbers	Adresses	lat	lon	combined lat	combined lon	country code	Organism	Product	Journal
5967	KM519783	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia pseudeuphratica*	tRNA-Val	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5968	KM519782	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia sericeotomentosa var. hatayica*	tRNA-Val	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5969	KM519781	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia sericeotomentosa var. sericeotomentosa*	tRNA-Val	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5970	KM519780	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia kronenburgii*	tRNA-Val	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5971	KM519779	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia cerinopruinosa*	tRNA-Val	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5972	KM519778	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia euphratica var. leiocalycina*	tRNA-Val	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5973	KM519777	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia euphratica var. euphratica*	tRNA-Val	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5974	KM519776	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia pseudeuphratica*	tRNA-Phe	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5975	KM519775	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia sericeotomentosa var. hatayica*	tRNA-Phe	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5976	KM519774	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia sericeotomentosa var. sericeotomentosa*	tRNA-Phe	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5977	KM519773	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia kronenburgii*	tRNA-Phe	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5978	KM519772	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia cerinopruinosa*	tRNA-Phe	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5979	KM519771	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia euphratica var. leiocalycina*	tRNA-Phe	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5980	KM519770	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia euphratica var. euphratica*	tRNA-Phe	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5981	KM519769	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia pseudeuphratica*	trnT-trnL intergenic spacer; IGS	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5982	KM519768	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia sericeotomentosa var. hatayica*	trnT-trnL intergenic spacer; IGS	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5983	KM519767	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia sericeotomentosa var. sericeotomentosa*	trnT-trnL intergenic spacer; IGS	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5984	KM519766	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia kronenburgii*	trnT-trnL intergenic spacer; IGS	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5985	KM519765	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia cerinopruinosa*	trnT-trnL intergenic spacer; IGS	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5986	KM519764	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia euphratica var. leiocalycina*	trnT-trnL intergenic spacer; IGS	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)
5987	KM519763	Turkey	n/a	n/a	38,960	34,925	TR	*Salvia euphratica var. euphratica*	n/a	Plant Syst. Evol. 301 (10), 2313-2323 (2015) Submitted (12-SEP-2014)

As in the previous example, clustering with relaxed parameters created the exact same three clusters as the analysis of the initial overall dataset. These clusters are attributed to three different cpDNA regions examined in the selected study. This subset of sequences doesn't have associated detailed address information in GenBank. Upon inspecting the study, sampling regions, while not coordinate information, can be achieved from the related publication. Combining locations and haplotype distributions manually demonstrates some levels of phylogeographic differentiation, especially for Cluster-122 and Cluster-132 (Figure 43). However, since species distributions do not overlap much for each cluster, and species are represented by a single sample in each population for each cluster, care must be taken when interpreting the results to make reliable phylogeographic inferences. The main purpose of this study is to investigate the phylogenetic relationships based on nuclear and chloroplast DNA regions of *Salvia euphratica* sensu lato and its closely related species, rather than conducting a phylogeographic analysis.

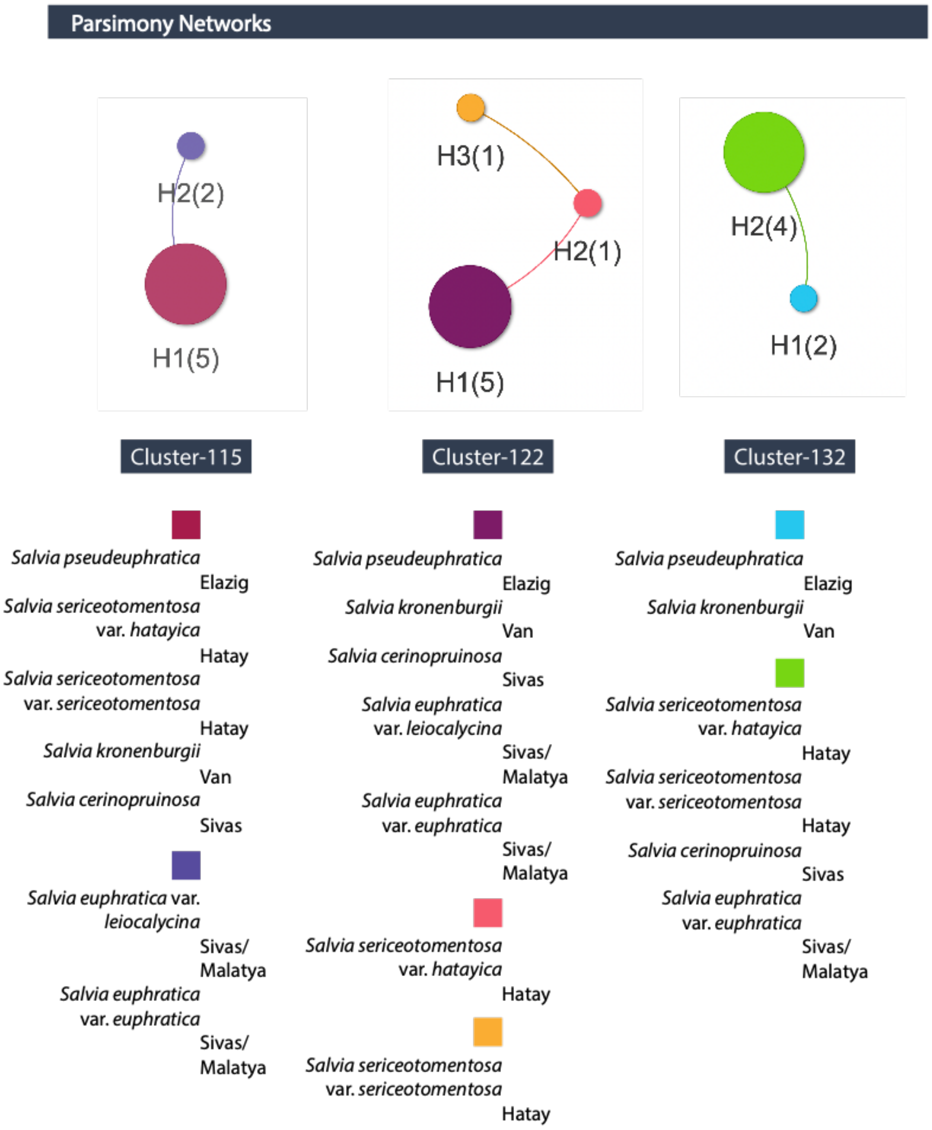

Parsimony Networks
H2(2)
H1(5)
H3(1)
H2(1)
H1(5)
H2(4)
H1(2)
Cluster-115
Cluster-122
Cluster-132
Salvia pseudeuphratica Elazig
Salvia sericeotomentosa var. hatayica Hatay
Salvia sericeotomentosa var. sericeotomentosa Hatay
Salvia kronenburgii Van
Salvia cerinopruinosa Sivas
Salvia euphratica var. leiocalycina Sivas/ Malatya
Salvia euphratica var. euphratica Sivas/ Malatya
Salvia pseudeuphratica Elazig
Salvia kronenburgii Van
Salvia cerinopruinosa Sivas
Salvia euphratica var. leiocalycina Sivas/ Malatya
Salvia euphratica var. euphratica Sivas/ Malatya
Salvia sericeotomentosa var. hatayica Hatay
Salvia sericeotomentosa var. sericeotomentosa Hatay
Salvia pseudeuphratica Elazig
Salvia kronenburgii Van
Salvia sericeotomentosa var. hatayica Hatay
Salvia sericeotomentosa var. sericeotomentosa Hatay
Salvia cerinopruinosa Sivas
Salvia euphratica var. euphratica Sivas/ Malatya

Figure 43. Manual integration of locations and haplotype distributions. Location data is sourced from reported sampling regions due to the lack of detailed address information in GenBank.

3.1.3. Mitochondrial DNA Example: Genus *Apis*

In the last example, available mithocondrial DNA records for genus *Apis* (Honey Bee) were retrieved using search terms "Apis [ORGANISM] AND mitochondrion[filter]". This search yielded a total of 10,735 records. The final dataset obtained after geocoding comprised 7,820 sequences. These 7,820 records encompass 9 *Apis* species represented by binomial names, 27 subspecies from three species which have trinomial names, and one hybrid (Figure 44). The sequence region information, available in the metadata of associated records, is derived from the "Product" feature is illustrated in the Figure 45.

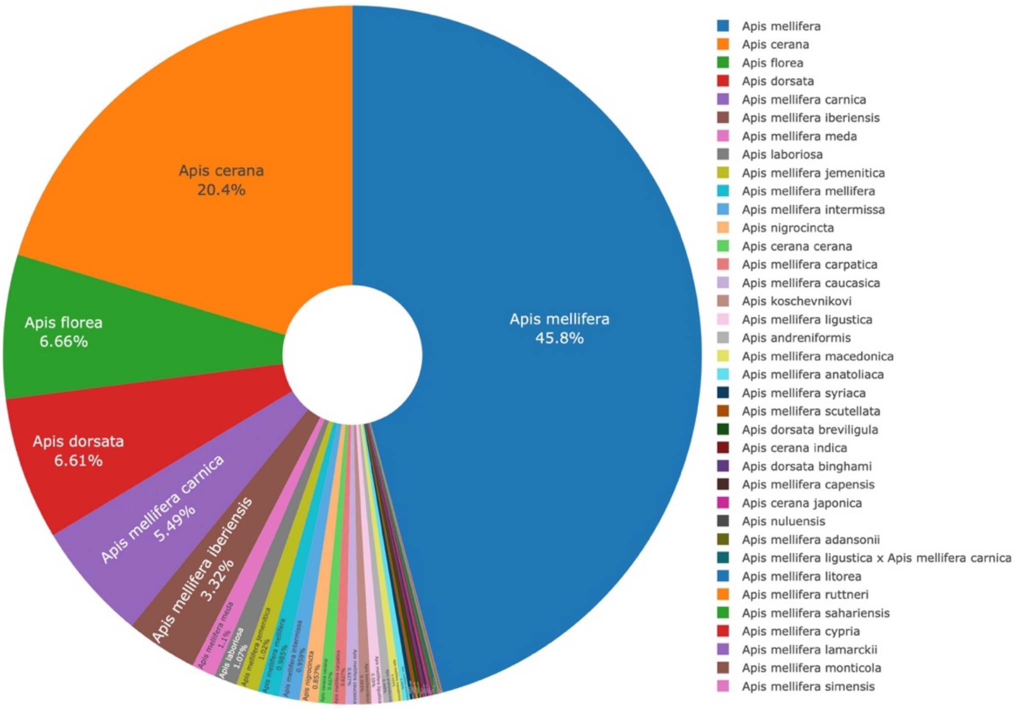


Figure 44. Donut chart representation of unique organism names in the *Apis* dataset. The legend displays all available organism names within the dataset.

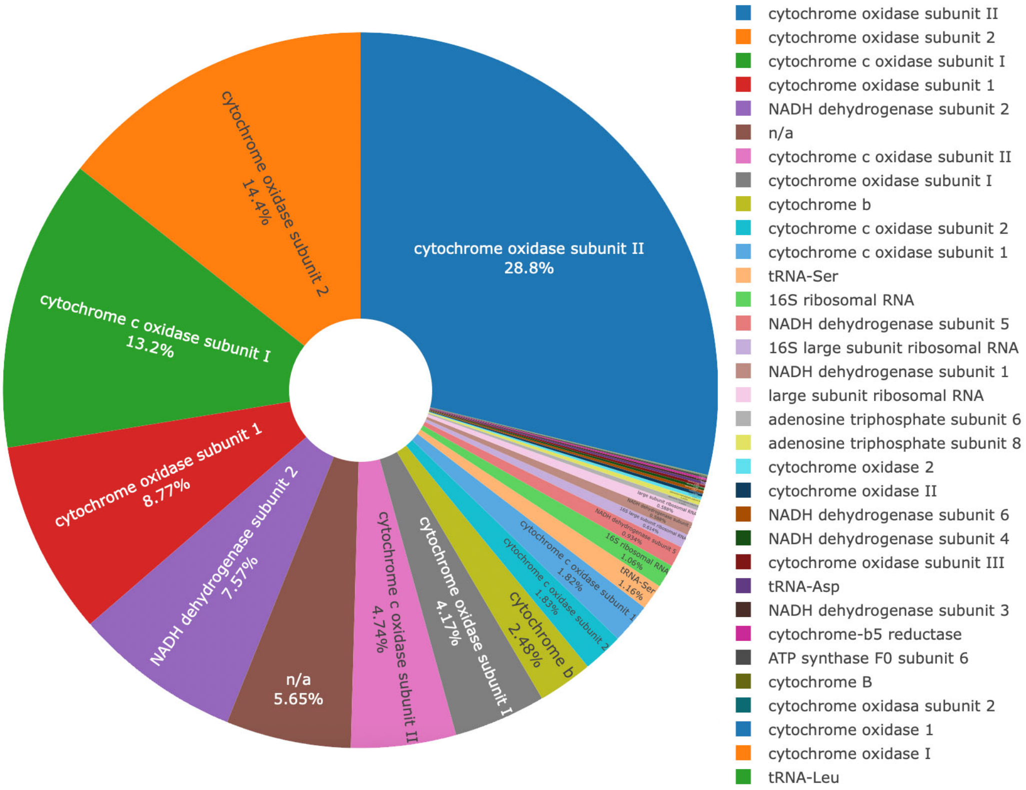
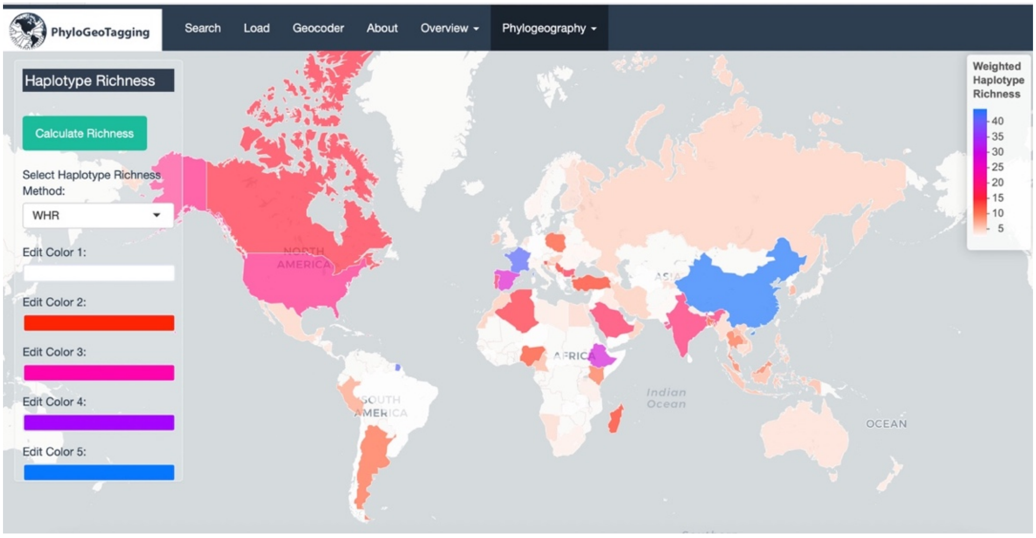


Figure 45. Donut chart representation of the sequence region information from the “Product” Feature in the Associated Records' Metadata.

Given the high number of subspecies in the *Apis* dataset, the *Taxonomic Pre-Grouping* setting is assigned to *Exact Match*. This method facilitates the examination of subspecies and ensures a fine level of taxonomic inspection. Keeping the pre-grouping parameter strict allows for less stringent settings in other parameters. Consequently, the k-mer and coverage threshold values are reduced to 90%. This combination of parameters results in 446 clusters, with sizes ranging from 1 to 1,142 sequences. The adoption of these specific settings is instrumental in achieving a delicate balance between specificity and sensitivity, allowing for a comprehensive yet precise analysis of the genetic diversity within the dataset. This approach underscores the flexibility and adaptability of the package in handling datasets with varied taxonomic complexities.

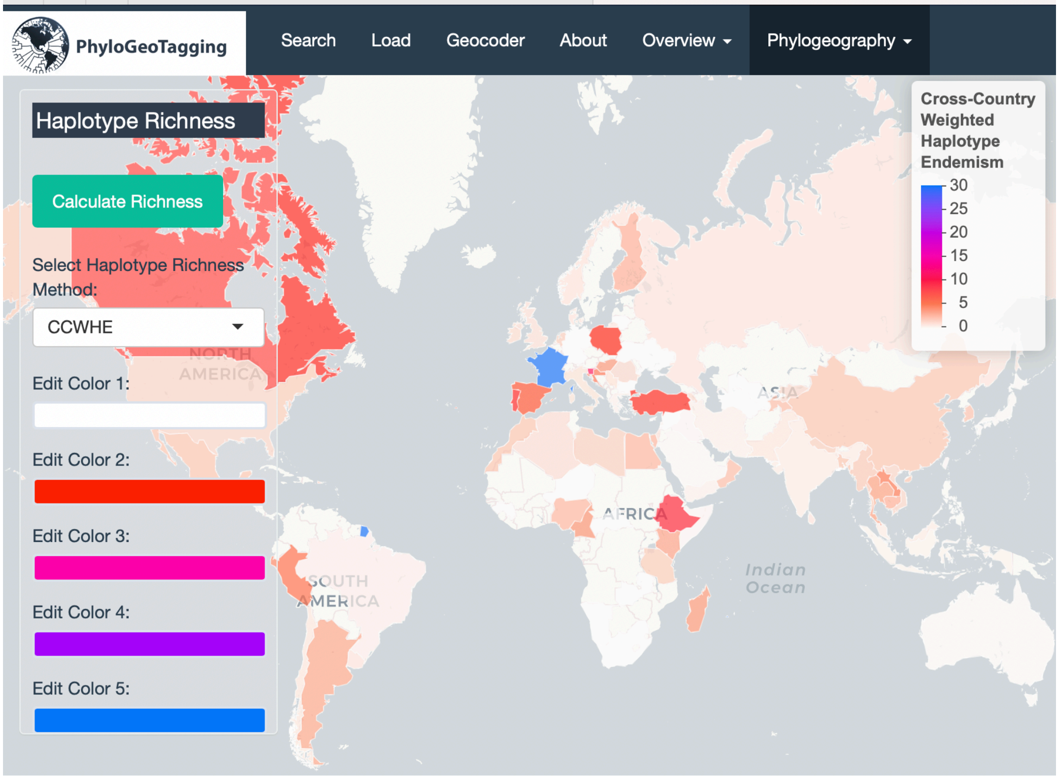
The *Apis* mitochondrion dataset exhibits extensive genetic variation compared to the cpDNA sequences analyzed in this study. Although the pre-grouping of sequences is restricted to the lowest taxonomic level, the level at which the subspecies are assigned to different clusters, the weighted haplotype richness metric in the dataset exceeds 40 (Figure 46). This notable level of genetic diversity within the *Apis* mitochondrion dataset underscores its potential as a valuable resource for studying phylogeographic patterns. The high haplotype richness metric particularly highlights the complexity and variability inherent in the mitochondrial DNA of the Honey Bee taxon.



Select Haplotype Richness Method: WHR

Figure 46. Weighted haplotype richness map for countries based on the *Apis* mitochondrion dataset.

Cross-Country Weighted Haplotype Endemism analysis reflects the only endemic haplotypes identified in studies that span multiple countries (Figure 47). This map aligns with general trends in genetic diversity, particularly for *Apis mellifera*, where high genetic diversity has been found in Africa and North America (Wallberg et al. 2014). Both weighted haplotype richness and haplotype endemism are markedly elevated in Europe, potentially enhanced by extensive apiculture activities and a broad spectrum of sequence submissions. Turkey is also characterized by a high number of endemic haplotypes, in accordance with previous findings (Kandemir et al., 2000).


Haplotype Richness

Select Haplotype Richness Method: CCWHE

Figure 47. Cross-Country Weighted Haplotype Endemism map for countries based on the *Apis* mitochondrion dataset.

This high genetic variation within species and subspecies appears to have been accompanied by high inter-population variations as well. AMOVA analysis conducted for population differentiation, return results for 125 clusters from 22 taxon. Eighty-seven of these clusters shows significant differentiation between population ($p<0.05$). Except for *Apis cerana indica, A. dorsata binghami*, *A. mellifera carpatica A. mellifera caucasica, A. mellifera meda*, and *A. nigrocincta*, samples, 16 taxon present significant population differentiation with one or more sequence regions involved, which represented different clusters (Figure 48).

This high genetic variation within species and subspecies appears to have been accompanied by significant inter-population variations as well. AMOVA analysis conducted for population differentiation returned results for 125 clusters from 22 taxa. Eighty-seven of these clusters show significant differentiation between populations ($p<0.05$). Except for *Apis cerana indica*, *A. dorsata binghami*, *A. mellifera carpatica*, *A. mellifera caucasica*, *A. mellifera meda*, and *A. nigrocincta* samples, 16 taxa demonstrate significant population differentiation with involvement of multiple sequence regions, each represented by distinct cluster. It is important to note, however, that taxa with a narrower distribution and/or fewer sequences submitted to GenBank are typically expected to show less and insignificant differentiation between populations, a trend which is illustrated in Figure 48. This substantial population differentiation among the majority of the taxa under study not only reinforces the genetic diversity within the *Apis* genus but also suggests a complex phylogeographic stratification shaped by various geographical factors.

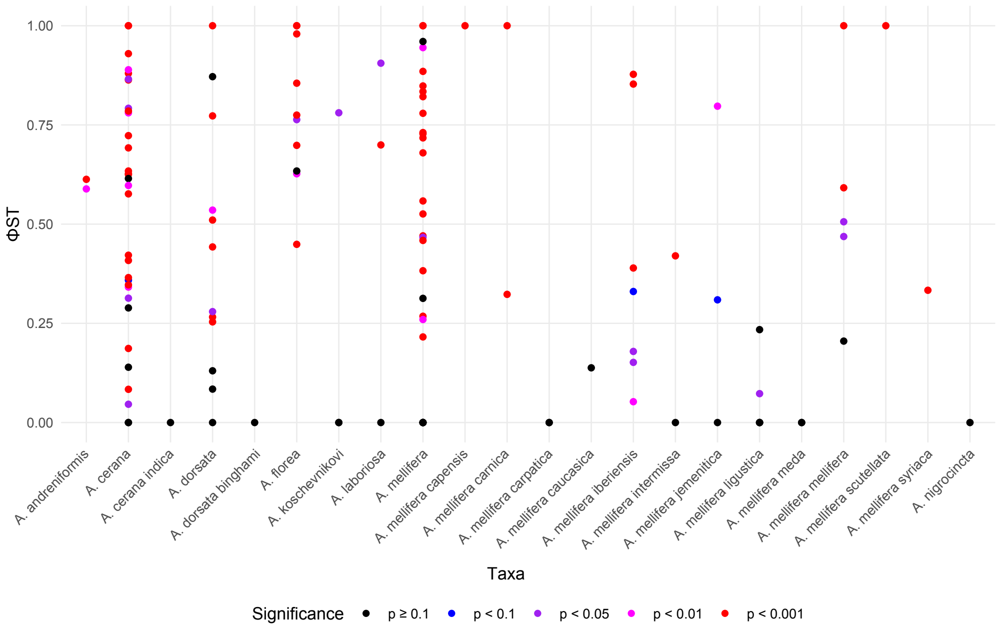


Figure 48. Colored dot plot of population differentiation Φ_{ST} across taxa. Each dot represents the population differentiation Φ_{ST} value for a cluster within a given taxon, color-coded by statistical significance. This visualization highlights the extent of

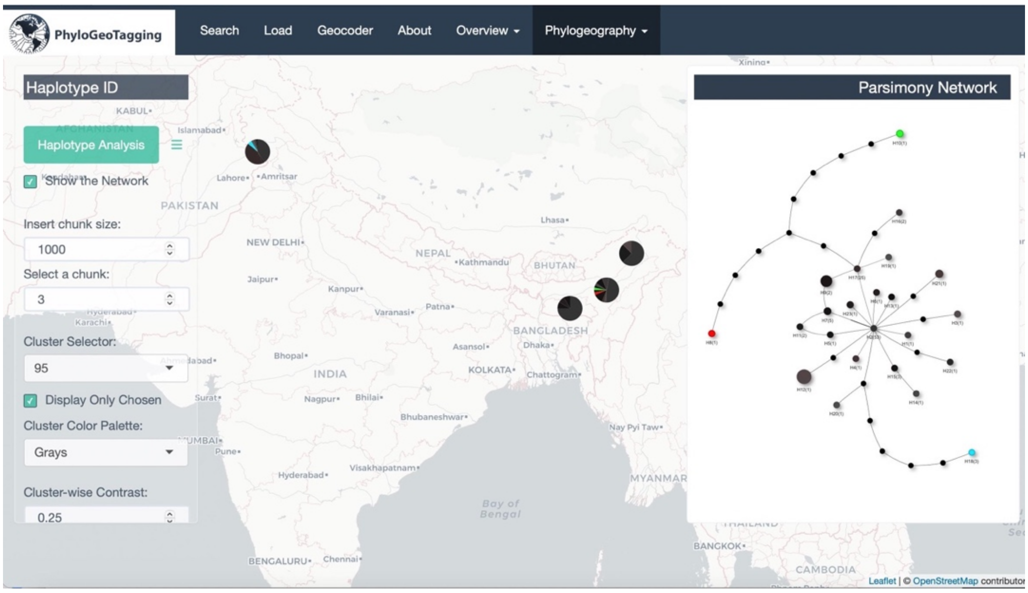
genetic differentiation within each taxon and the associated p-values indicating the strength of the statistical evidence.

The genetic diversity and differentiation within and among populations of *Apis* species and subspecies are remarkably high. Hapstep analysis, which could be applied to 59 clusters, revealed that 32 of these clusters resulted in a significant N_{ST} value, as shown in Table 4**.** This analysis illuminated diverse patterns of genetic differentiation across these clusters.

Table 4. Summary of Hapstep analysis for 59 clusters in *Apis* species and subspecies. The table details the diverse patterns and trends of genetic differentiation within sequence regions, each represented by a distinct cluster.

Cluster	Taxa	nhap	npop	harmean	h_T	v_T	v_T-pval	G_{ST}	N_{ST}	N_{ST}-pval
83	*A. andreniformis*	5	2	3.429	1	1.25	0.817	0.5	0.667	0.379
68	*A. cerana*	7	2	3.429	1	1.247	1	0	0.269	0.053.
95	*A. cerana*	23	4	15.575	0.685	0.352	0.005**	0.096	0.08	0.562
110	*A. cerana*	70	16	3.735	1	0.959	0.019*	0	0.105	0.001***
122	*A. cerana*	15	4	9.071	0.947	0.812	0.203	0.477	0.703	0.001***
158	*A. cerana*	35	4	6.314	1	0.973	0.548	0.31	0.571	0.075.
204	*A. cerana*	63	12	4.643	1	0.922	0.005**	0	0.411	0.001***
216	*A. cerana*	28	8	3.713	0.991	1.076	0.769	0.103	0.524	0.001***
257	*A. cerana*	60	9	5.657	1	1.121	0.967	0	0.675	0.001***
356	*A. cerana*	2	2	5.333	0.375	0.375	1	0.348	0.348	1
366	*A. cerana*	10	3	3.857	0.963	1.332	0.995	0.019	0.92	0.001***
382	*A. cerana*	13	16	9.44	0.486	0.232	0.002**	0.612	0.496	0.967
394	*A. cerana*	15	2	6.545	1	1.131	0.924	0.056	0.371	0.003**
418	*A. cerana*	9	16	9.5	0.762	0.724	0.321	0.567	0.663	0.016*
439	*A. cerana*	14	4	34.688	0.314	0.177	0.013*	0.097	0.036	0.971
444	*A. cerana*	9	3	3.6	0.981	1.07	0.783	0.117	0.42	0.01**
31	*A. dorsata*	34	7	14.116	0.819	0.505	0.008**	0.408	0.561	0.001***
174	*A. dorsata*	21	6	7.048	0.863	0.62	0.054.	0.334	0.444	0.052.
323	*A. dorsata*	22	5	4.803	0.92	0.703	0.001***	0.174	0.379	0.002**
353	*A. dorsata*	7	2	7.619	1	1.075	0.618	0.575	0.823	0.032*
370	*A. dorsata*	9	2	8	0.72	0.532	0.058.	-0.031	0.005	0.174
383	*A. dorsata*	23	3	5	1	1.112	0.92	0	0.259	0.002**
410	*A. dorsata*	9	2	4.2	0.952	1.062	0.729	-0.05	0.012	0.112
417	*A. dorsata*	12	4	3.2	0.981	0.952	0.243	-0.019	0.069	0.178
438	*A. dorsata*	8	2	3.75	1	1.353	1	0	0.595	0.02*
281	*A. florea*	4	5	3.818	0.673	0.528	0.222	0.558	0.51	0.836
354	*A. florea*	9	3	4.615	1	1.204	0.986	0.311	0.716	0.002**
376	*A. florea*	3	2	3.429	1	1.312	1	0.75	0.857	0.358
403	*A. florea*	23	7	10.231	0.804	0.587	0.086.	0.385	0.665	0.001***
453	*A. koschevnikovi*	9	3	3	1	1.014	0.907	0	0.053	0.12
1	*A. mellifera*	43	11	5.185	0.971	0.873	0.051.	0.036	0.241	0.001***
6	*A. mellifera*	25	4	6.91	0.99	1.149	0.946	0.084	0.578	0.001***
47	*A. mellifera*	45	13	5.433	0.984	0.884	0.128	0.122	0.507	0.001***
75	*A. mellifera*	10	4	9.655	0.83	0.517	0.099.	0.386	0.271	0.654
186	*A. mellifera*	86	55	5.347	0.964	0.798	0.016*	0.319	0.46	0.005**
241	*A. mellifera*	18	2	6.222	1	2.148	1	0	0.857	0.001***
295	*A. mellifera*	13	4	5.255	0.901	0.828	0.14	-0.033	-0.034	0.447
297	*A. mellifera*	32	3	11.2	1	1.416	1	0.182	0.878	0.001***
324	*A. mellifera*	6	2	3	1	1.336	1	0	0.628	0.112
327	*A. mellifera*	4	4	8.102	0.729	0.568	0.206	0.363	0.273	0.674
330	*A. mellifera*	5	4	4.898	0.833	0.88	0.646	0.56	0.432	0.847
331	*A. mellifera*	30	2	23.303	1	1.274	0.924	0.28	0.792	0.001***
334	*A. mellifera*	29	7	3.978	1	1.001	0.535	0.014	0.209	0.001***
335	*A. mellifera*	25	2	8.276	0.967	0.739	0.011*	-0.034	-0.035	0.463
362	*A. mellifera*	24	19	7.042	0.87	0.803	0.323	0.466	0.793	0.001***
433	*A. mellifera*	12	3	4	1	1.306	1	0.056	0.837	0.001***
340	*A. mellifera capensis*	2	3	3	0.667	0.667	1	1	1	1
299	*A. mellifera carnica*	30	54	5.276	0.647	0.392	0.018*	0.296	0.375	0.006**
435	*A. mellifera carpatica*	8	2	4.8	0.917	0.8	0.193	0.017	-0.041	0.66
7	*A. mellifera iberiensis*	66	2	31.061	1	1.047	0.966	0	0.05	0.006**
78	*A. mellifera iberiensis*	21	2	10.476	1	1.114	0.968	0	0.172	0.044*
191	*A. mellifera iberiensis*	49	2	24.245	1	1.285	1	0	0.372	0.001***
268	*A. mellifera iberiensis*	10	4	5.161	0.535	0.341	0.091.	0.066	0.087	0.253
406	*A. mellifera iberiensis*	15	2	5.867	1	1.17	0.98	0	0.09	0.081.
8	*A. mellifera intermissa*	18	2	10.839	1	0.966	0.515	0.1	0.438	0.001***
307	*A. mellifera jemenitica*	19	2	6.316	1	1.129	0.865	0	-0.075	0.89
301	*A. mellifera ligustica*	3	2	3.429	0.5	0.409	0.166	-0.167	-0.056	0.67
360	*A. mellifera mellifera*	8	3	3.857	0.938	1.001	0.807	0.023	0.471	0.001***
350	*A. mellifera scutellata*	2	3	3	0.667	0.667	1	1	1	1

This analysis also brought to light various trends in genetic diversity across these clusters. In 28 of the clusters, h_T was greater than v_T; in three clusters, h_T equaled v_T; and in the remaining 28 clusters, h_T was less than v_T. The most significant discrepancy, where h_T exceeded v_T, was observed in Cluster-95. In this cluster, certain haplotypes have newly emerged and demonstrated notable differentiation from other haplotypes found in the population, as illustrated in Figure 49. Specifically, haplotype H8 is shown in red, H10 in green, and H18 in cyan. This differentiation leads to a reduction in the v_T value, indicating ongoing diversification within the populations.



☑ Show the Network

Insert chunk size: 1000

Select a chunk: 3

Cluster Selector: 95

☑ Display Only Chosen

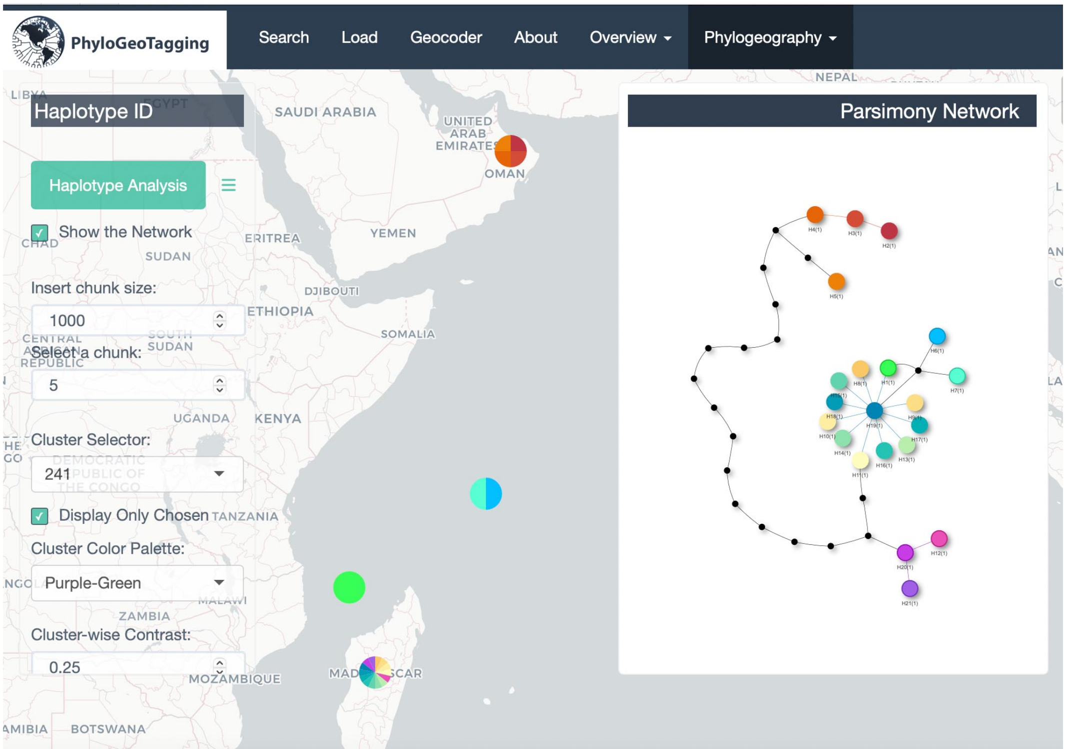
Cluster Color Palette: Grays

Cluster-wise Contrast: 0.25

Leaflet | © OpenStreetMap contributors

Figure 49. Haplotype differentiation in Cluster-95. Showcases the emergence and notable differentiation of haplotypes (H8 in red, H10 in green, H18 in cyan), emphasizing the diversification within populations that results in the low v_T value.

In contrast, Cluster-241, which includes *A. mellifera* cytochrome oxidase subunit II sequences, exhibited the largest difference with v_T exceeding h_T. This cluster is characterized by many variants with less within population differentiations. This pattern is most evident in the largest population representing all of Madagascar. It appears that Madagascar has been recently colonized by *A. mellifera* individuals carrying the H19 haplotype (Figure 52). This haplotype is central to the haplotype network and has generated many new variants. During the initial diversification on the main island, individuals with the H19 haplotype seem to have colonized a northwest island, and the H19 haplotype gave rise to the H1 haplotype, which might become dominant. A probable migration from this island to another in the northeast led to a new mutation in carriers of the H1 haplotype. This mutation resulted in an extinct or unsampled haplotype, indicated by small black nodes in the network, which gave rise to H6 and H7, currently populating this small island. The haplotype network shows a loop connecting these nodes, suggestive of homoplastic mutations. Consequently, due to this loop, it remains unclear whether the ancestors of H6 and H7 are directly descended from H19 or H1. Additionally, it is worth mentioning that the H12, H20, and H22 haplotype group, another lineage that colonized Madagascar, is connected to the main H19 group via the newly formed H11 haplotype. This connection is likely a result of homoplasy as well. Homoplasy may create shorter links in evolutionary networks than actual genealogical connections. However, this aspect requires a more detailed examination of the sequence data, which is not done here as it is beyond the scope of this study. A high level of population differentiation, facilitated by between-island and south-north separation, and the recent diversification of haplotypes, primarily within populations, has resulted in N_{ST} being significant and greater than G_{ST}. The high number of newly emerged haplotypes contributes to an increase in both h_T and v_T values. However, these newly emerged haplotypes have not accumulated many mutations, leading to v_T being greater than h_T. Unfortunately, the data from this cluster is the only information available from Madagascar in the *Apis* set, preventing a comprehensive comparative phylogeographic analysis.



☑ Show the Network

Insert chunk size: 1000

Select a chunk: 5

Cluster Selector: 241

☑ Display Only Chosen

Cluster Color Palette: Purple-Green

Cluster-wise Contrast: 0.25

Figure 50. Haplotypes map and parsimony network for Cluster-241. This figure demonstrates high population differentiation driven by geographic separations, both between islands and in a south-north direction, with a notable N_{ST} value exceeding G_{ST}.It also shows recent haplotype diversification within populations, characterized by newly emerged haplotypes with limited mutations, leading to a higher v_T compared to h_T.

Finally, due to the rich variation in DNA sequences within the *Apis* dataset, ΦST Barriers have been observed to emerge around the world (Figure 51). When the ΦST Barriers map is examined, the most notable feature in the genus *Apis*, in contrast to plants, is the reduced Φ_{ST} values in regions where the sampling points (populations) are densely clustered. On the map, blue colors are prominently surrounding the green population points. This indicates, as expected, that there is a higher level of gene flow between geographically proximate populations, leading to weakened genetic barriers. In the *Apis* dataset, it is observed that the barriers are predominantly situated over deserts and oceans. These expansive and inhospitable terrains act as natural obstructions to gene flow, reinforcing the formation of genetic barriers. The widespread appearance of barriers around the world can be attributed to the comprehensive sampling of the *Apis* dataset from diverse geographies. This dataset includes world-wide common species and subspecies (e.g., *A. mellifera*) compared to plant samples, which has allowed for widespread phylogeographic analysis on a global scale.

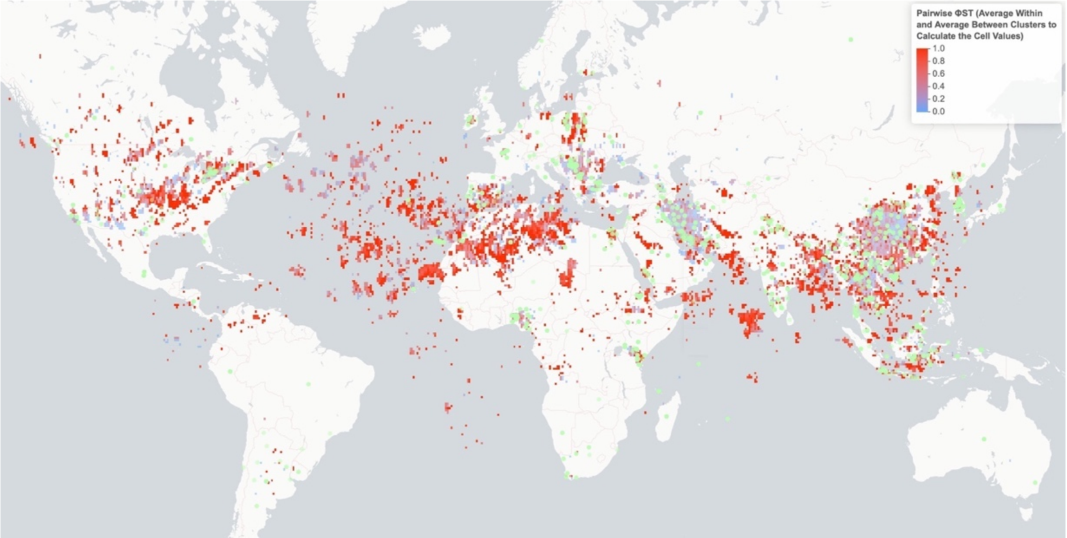

Pairwise ΦST (Average Within and Average Between Clusters to Calculate the Cell Values)
1.0
0.8
0.6
0.4
0.2
0.0

Figure 51. Global ΦST Barriers in the *Apis* dataset. This map highlights the emergence of ΦST Barriers worldwide, with a notable observation of reduced Φ_{ST} values in densely sampled regions. The blue colors surrounding the green population points signify a higher level of gene flow between geographically proximate populations, resulting in weakened genetic barriers, a distinctive feature in the genus *Apis* compared to plants examined in the examples.

3.2. Limitations and Future Prospects

This PhD thesis presents the complete version of the novel package **PhyloGeoTagging**; however, it is more appropriately regarded as an introductory phase to a continuing research study. The procedures introduced herein constitute the initial version of the software, and it is noteworthy that new ideas for its improvement have emerged even during the writing of this thesis. Inevitably, a number of limitations have appeared, attributable both to this package being the first version and to certain inherent constraints within some of the methods introduced, despite their utility. Consequently, it will be beneficial to outline some expectations and limitations regarding the outcomes of various methods encompassed within the package and to discuss future development in general terms.

GenBank is a comprehensive public DNA sequence database that contains approximately 3 billion nucleotide sequences (Sayers et al., 2023). Although GenBank offers a wealth of resources, it is not particularly designed for specific purposes such as storing sequence data collected for phylogeographic analysis. In 2005, GenBank introduced a latitude-longitude field in their stored data, which allows for the association of sequence data with geographical information. However, by the end of 2023, available geographic coordinate data is still limited, especially for plants. In many sequence records, even address information is not available to allow geocoding. For instance, a check of the address feature availability for flowering plants cpDNA records, using the search term "Angiosperms [ORGANISM] AND plastid[filter] AND src country[PROP]," reveals that only 46% have available data in the country feature, indicating that more than half of the data lacks address information. A similar search using "Metazoa [ORGANISM] AND mitochondrion[filter] AND src country[PROP]" returns that 70% of Animal mtDNA records have address information, more suitable for geocoding but still not fully comprehensive for the records. This is just the beginning of the story. As shown in the examples in the previous sections, the address information available may be quite generic, such as just the country names, resulting in lower analysis resolutions. It is evident that over time, GenBank metadata will become richer. Another mission with this and similar tools is to uncover the utility of the metadata, which will provide more encouragement for people who upload sequences to GenBank to include enriched metadata. In the future, more flexible data gathering for geocoding within records and the addition of more geocoding tools beside Nominatim, could help partially overcome these problems. Other features of the metadata, not currently disclosed, can also be made more accessible by enhancing the flexibility of the parser algorithm and the Parser tab. It would also be beneficial to provide the option for users to manipulate both the parser and the Geocoder tables. This would enable users to insert additional metadata, such as address information, or to make corrections.

Clustering is an indispensable part of the automated analysis procedure in this package. It is simple yet efficient. Still, there is room for development. Development in the parser process may add new dimensions to pre-grouping methods, especially the potential to explore reliable gene information of sequences. In the following updates, it is planned that pre-grouping can be done according to other features in the Geocoder table. This will be particularly useful for those who want to pre-group sequences by criteria such as country, title, journal, or gene. For example, by using the title feature, the clusters will consist only of sequences possibly obtained from the same study and submitted to GenBank together.

Another useful modification of the package will be adding the table presentation option for IBD, AMOVA, and Hapstep analysis. These tables will not only tidy the output but also serve as a filter for subsetting the dataset. This will be particularly helpful for both simplifying and emphasizing chosen aspects of the data (e.g., a subset of data showing phylogeographic structure). It will be possible to filter the main dataset repeatedly with different filters at different analyses. This also allows researchers to make deductions; with the available tools, one can detect phylogenetic signals in a sequence group, then identify its associated scientific reference and delve deeper into the original study. Improving such filters at every step also strengthens the deductive side of the tool. To make in-depth research easier, interactive buttons can be integrated into the Shiny application, providing direct web links to access the corresponding GenBank records.

In utilizing sequence data from repositories such as GenBank for phylogeographic analysis, it is essential to consider the variability arising from the data submission patterns to the database. This variability can introduce a notable challenge in ensuring the representativeness of the data. This issue is particularly pertinent in richness analyses, where the data's comprehensiveness is important. While sampling biases are effectively adjusted for by weighting endemic haplotypes in relation to the distribution of sequence samples across countries in the implemented method, a notable limitation is introduced by the nature of sequence submissions. Specifically, the accuracy and representativeness of the analysis are contingent upon the sequences submitted being a random and unbiased sample of the genetic diversity within each country. However, if predominantly polymorphic variants are submitted from certain countries, this could inversely affect the analysis. In such instances, an artificially inflated count of unique or endemic haplotypes may appear to be present in countries contributing a higher proportion of polymorphic sequences. This phenomenon could potentially distort the true picture of haplotype richness and distribution. Therefore, caution must be exercised in interpreting the results, particularly concerning the nature and representativeness of the sequence data, while significant insights into the endemic genetic variation are offered by the methodology.

Although this package was initially created for phylogeographic studies, it can also be utilized to examine differences between species, particularly by employing AMOVA. However, it is essential to note that AMOVA is not specifically designed for species differentiation studies. This method partitions genetic variance within and among groups—such as populations or species—using molecular data and quantifies the proportion of total genetic variance attributable to different hierarchical factors. While invaluable for understanding the distribution of genetic variation, the results of AMOVA must be interpreted with caution when applied to species differentiation, since the method is more focused on partitioning variance rather than on testing hypotheses about species boundaries or speciation events. For example, even if a Φ_{ST} value of 1 is observed for species differentiation, it does not confirm the complete separation of all species; it simply indicates that the existing variance is primarily due to differences between some species. This is particularly true if there is no variation within species, as a Φ_{ST} value of 1 can still be obtained even in the absence of interspecific variance. A high Φ_{ST} value (approaching 1) at the species level suggests that a considerable portion of the genetic variance is due to differences between species. However, this does not automatically imply complete genetic isolation or distinctiveness among all species within the dataset. It is important to remember that Φ_{ST} is a relative measure, indicating the proportion of total variance that is between groups compared to within groups. Therefore, if the within-group variation is minimal or nonexistent, Φ_{ST} will be high, which can potentially give a misleading impression of differentiation. Nevertheless, the current model design of AMOVA in this package still offers a general overview of the differences between species, particularly in cases of allopatric speciation. AMOVA is useful for exploring various combinations of genetic variation, accommodating analyses both within and between taxa, and within and between populations. While the package is currently capable of examining population differentiation within the same taxa, future versions will be enhanced to facilitate various model designs, such as comparisons between taxa within populations.

Although organelle DNA is predominantly utilized in phylogeographic studies and is generally not employed for direct phylogenetic inference, researchers may also wish to conduct phylogenetic analyses with organelle or nuclear DNA sequences. For this reason, and to strengthen the phylogenetic and molecular taxonomy aspects of this dataset, phylogenetic methods will also be incorporated in future updates.

The **PhyloGeoTagging** package will be submitted to a major R repository, and its corresponding Shiny application will be deployed on a public server, both becoming publicly accessible following the thesis publication.

CHAPTER 4

CONCLUSION

This thesis represents a significant step forward in the field of phylogeographic research through the development of an R package and Shiny web application: **PhyloGeoTagging**. This package, specifically designed for phylogeographic studies, simplifies and accelerates the process of data analysis using GenBank's extensive database. Addressing several challenges inherent in phylogeographic research, such as the time-consuming retrieval and preparation of data from GenBank and the often limited geographical data available, this package introduces a user-friendly interface that guides users through a comprehensive workflow. This workflow includes searching with the ESearch utility, downloading sequences and preparing associated metadata using EFetch and XML parsing, and enhancing geospatial data through geocoding with the Nominatim API.

The package's functionalities extend to automating crucial tasks in phylogeographic studies, such as clustering homologous sequences, performing sequence alignment, and conducting various phylogeographic analyses like IBD, AMOVA, and Hapstep. These processes, traditionally demanding in terms of time and expertise, are now streamlined, making comprehensive phylogeographic studies more accessible to a wider research audience. Furthermore, the application's capabilities in identifying haplotypes, representing haplotype networks, calculating richness, and mapping genetic differentiation enhance the understanding and visualization of genetic data. Dynamic maps and infographics, integrated into the application, transform complex datasets into engaging and comprehensible formats, aiding in the interpretation of the phylogeographic patterns.

This package enables the analysis of large datasets across any taxonomic scale, incorporating multiple gene regions available within a taxonomic group in GenBank. Designed for phylogeographic procedures, it categorizes and clusters homologous sequences—referred to as 'clusters' in this study—ensuring analyses are conducted within homogenous groups. This capability allows users to input diverse datasets—effectively handling both “apples” and “oranges”—and ensures that comparisons are made accurately, with “apples compared with apples, and oranges with oranges”. The program facilitates the analysis of the entire dataset in a single step, offering the option to display combined or separate results based on sequence homology, as per user request. Furthermore, the richness calculation algorithms introduced in this thesis are specifically formulated for large-scale meta-analysis of sequence sets, significantly enhancing the package's utility for comprehensive phylogeographic studies.

Significant upgrades have been made by me to my previously published R package, **haplotypes**, which has been established as the core dependency for the **PhyloGeoTagging** package developed during this thesis.

In the context of this thesis, the case studies of the Fagaceae family and the *Salvia* and *Apis* genera stand as a testament to the transformative potential of the **PhyloGeoTagging** package in phylogeographic research. The meticulous analysis of over 28,000 chloroplast DNA and over 10,000 mitochondrial DNA sequences from GenBank, culminating in the successful geocoding of over 22,000 records, underscores the package's proficiency in managing extensive datasets and enhancing geographical data precision on a global scale. This comprehensive approach yielded insightful findings, highlighting significant genetic differentiation and the identification of specific phylogeographic structures across various species and regions.

Overall, this thesis marks an important contribution to phylogeographic research by facilitating large-scale meta-analyses. It repurposes sequences from GenBank, submitted for diverse studies, and uses them specifically for phylogeographic inference, thereby enriching our understanding of global biodiversity patterns at any chosen scale. The launch of this package, accompanied by detailed technical documentation, is set to be a significant addition to the phylogeography community and the broader field of open-data science. It underscores the potential of combining advanced computational techniques with biological research, setting the stage for future innovations in phylogeography studies.

REFERENCES

Aktas, C. (2020). haplotypes: Manipulating DNA Sequences and Estimating Unambiguous Haplotype Network with Statistical Parsimony (Version 1.1.3) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=haplotypes

Al-Jumaili, A. S., & Hanotte, O. (2022). The usefulness of maternally inherited genetic markers for phylogeographic studies in village chicken. *Animal Biotechnology, 34*(4),863–881. https://doi.org/10.1080/10495398.2021.2000429

Albery, G. F., Eskew, E. A., Ross, N., & Olival, K. J. (2020). Predicting the global mammalian viral sharing network using phylogeography. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-16153-4

Almende, B. V., et al. (2022). visNetwork: Network Visualization using 'vis.js' Library (Version 2.1.2) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=visNetwork

Amandita, F. Y., Rembold, K., Vornam, B., Rahayu, S., Siregar, I. Z., Kreft, H., & Finkeldey, R. (2019). DNA barcoding of flowering plants in Sumatra, Indonesia. *Ecology and Evolution*, *9*(4), 1858–1868. https://doi.org/10.1002/ece3.4875

Arbogast, B. S. (2001). Phylogeography: The history and formation of species. *American Zoologist, 41*(1), 134-135. https://doi.org/10.1668/0003-1569(2001)041[0134:BR]2.0.CO;2

Avise, J. (2000). *Phylogeography: The History and Formation of Species*. President and Fellows of Harvard College. ISBN 978-0-674-66638-2.

Avise, J. C. (1998). The history and purview of phylogeography: A personal reflection. *Molecular Ecology, 7*(4), 371–379. https://doi.org/10.1046/j.1365-294x.1998.00391.x

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A., & Saunders, N. C. (1987). Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics, 18*, 489–522.

Avise, J. C., Bowen, B. W., & Ayala, F. J. (2016). In the light of evolution X: Comparative phylogeography. *Proceedings of the National Academy of Sciences*, *113*(29), 7957–7961. https://doi.org/10.1073/PNAS.1604338113

Axelrod, D. I. (1983). Biogeography of Oaks in the Arcto-Tertiary Province. *Annals of the Missouri Botanical Garden*, *70*(4), 629. https://doi.org/10.2307/2398982

Azara S.I., M., & Yakubu, A. In-silico molecular analysis of rabies virus across regions. *Computational Molecular Biology*. https://doi.org/10.5376/cmb.2014.04.0008

Azara S.I. M., & Yakubu, A. (2014). In-silico molecular analysis of rabies virus across regions. *Computational Molecular Biology*. https://doi.org/10.5376/cmb.2014.04.0008

Benn Torres, J. (2016). A history of you, me, and humanity: mitochondrial DNA in anthropological research. *Genetics and Evolution*, *03*(02), 146–156. https://doi.org/10.3934/genet.2016.2.146

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., & Sayers, E. W. (2017). GenBank. *Nucleic Acids Research*, *46*(D1), D41–D47. https://doi.org/10.1093/nar/gkx1094

Benson, D., Lipman, D. J., & Ostell, J. (1993). GenBank. *Nucleic Acids Research*, *21*(13), 2963–2965. https://doi.org/10.1093/nar/21.13.2963

Bermingham, E., & Moritz, C. (1998). Comparative phylogeography: concepts and applications. *Molecular Ecology*, *7*(4), 367–369. https://doi.org/10.1046/j.1365-294x.1998.00424.x

Bi, Y., Zhang, M.F., Xue, J., Dong, R., Du, Y. P., & Zhang, X. H. (2018). Chloroplast genomic resources for phylogeny and DNA barcoding: A case study on Fritillaria. *Scientific Reports, 8*(1). https://doi.org/10.1038/s41598-018-19591-9

Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C., & Hochreiter, S. (2015). msa: An R package for multiple sequence alignment. *Bioinformatics, 31(24)*, 3997-3999. https://doi.org/10.1093/bioinformatics/btv494

Bowen, B. W., & Karl, S. A. (2007). Population genetics and phylogeography of sea turtles. *Molecular Ecology*, *16*(23), 4886–4907. https://doi.org/10.1111/j.1365-294x.2007.03542.x

Briggs, J. C. (2009). Darwin's biogeography. *Journal of Biogeography, 36*, 1011-1017. https://doi.org/10.1111/j.1365-2699.2008.02076.x

Carstens, B. C., Morales, A. E., Field, K., & Pelletier, T. A. (2018). A global analysis of bats using automated comparative phylogeography uncovers a surprising impact of Pleistocene glaciation. *Journal of Biogeography*, *45*(8), 1795–1805. https://doi.org/10.1111/jbi.13382

Celep, F., & Doğan, M. (2023). The Genus Salvia in Turkey: Morphology, Ecology, Phytogeograpy, Endemism and Threat Categories. *Medicinal and Aromatic Plants of the World*, 107–120. https://doi.org/10.1007/978-3-031-43312-2_5

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2022). *shiny: Web Application Framework for R* (Version 1.7.3. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=shiny

Cheng, J., Schloerke, B., Karambelkar, B., & Xie, Y. (2023). leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library (Version 2.2.0) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=leaflet

Costa, J. (2013). Synonymy and its discontents: Alfred Russel Wallace's nomenclatural proposals from the 'Species Notebook' of 1855–1859. *The Bulletin of Zoological Nomenclature, 70*, 131-148. https://doi.org/10.21805/BZN.V70I2.A5

Cottrell, J. E., et al. (2005). Postglacial migration of Populus nigra L.: Lessons learnt from chloroplast DNA. *Forest Ecology and Management, 206*, 71–90. https://doi.org/10.1016/j.foreco.2004.10.052.

Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sunderland, MA: Sinauer Associates.

Curtu, A. L., Gailing, O., & Finkeldey, R. (2007). Evidence for hybridization and introgression within a species-rich oak (Quercus spp.) community. *BMC Evolutionary Biology, 7*, 218. https://doi.org/10.1186/1471-2148-7-218

Cutter, A. D. (2013). Integrating phylogenetics, phylogeography, and population genetics through genomes and evolutionary theory. *Molecular Phylogenetics and Evolution*, *69*(3), 1172–1185 https://doi.org/10.1016/J.YMPEV.2013.06.006

da Silva Ferrette, B. L., Coelho, R., Peddemors, V. M., Ovenden, J. R., De Franco, B. A., Oliveira, C., Foresti, F., & Mendonça, F. F. (2021). Global phylogeography of the smooth hammerhead shark: Glacial refugia and historical migration patterns. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *31*(9), 2348–2368. https://doi.org/10.1002/aqc.3629

Darwin, C. (1964). *On the Origin of Species: A Facsimile of the First Edition*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9xp5

de Queiroz, A. (2005). The resurrection of oceanic dispersal in historical biogeography. *Trends in Ecology & Evolution, 20*(2), 68–73. https://doi.org/10.1016/j.tree.2004.11.006

Der, J. P., Thomson, J. A., Stratford, J. K., & Wolf, P. G. (2009). Global chloroplast phylogeny and biogeography of bracken (Pteridium;

Dennstaedtiaceae). *American Journal of Botany*, *96*(5), 1041–1049. https://doi.org/10.3732/ajb.0800333

Dizkirici, A., Celep, F., Kansu, C., Kahraman, A., Dogan, M., & Kaya, Z. (2015). A molecular phylogeny of Salvia euphratica sensu lato (Salvia L., Lamiaceae) and its closely related species with a focus on the section Hymenosphace. *Plant Systematics and Evolution*, *301*(10), 2313–2323. https://doi.org/10.1007/s00606-015-1230-1

Doan, K. et al. (2021). Phylogenetics and phylogeography of red deer mtDNA lineages during the last 50 000 years in Eurasia. *Zoological Journal of the Linnean Society, 194(2)*, 431–456. https://doi.org/10.1093/zoolinnean/zlab025

Douda, J., Doudová, J., Drašnarová, A., Kuneš, P., Hadincová, V., Krak, K., Zákravský, P., & Mandák, B. (2014). Migration patterns of subgenus Alnus in Europe since the Last Glacial Maximum: A systematic review. *PLoS ONE, 9*, e88709. https://doi.org/10.1371/journal.pone.0088709

Edwards, R. A., et al. (2019). Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nature Microbiology*, *4*(10), 1727–1736. https://doi.org/10.1038/s41564-019-0494-6

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, *131*(2), 479–491. https://doi.org/10.1093/genetics/131.2.479

Fedorov, V., Goropashnaya, A., Jarrell, G. H., & Fredga, K. (1999). Phylogeographic structure and mitochondrial DNA variation in true lemmings (Lemmus) from the Eurasian Arctic. *Biological Journal of The Linnean Society, 66*(3), 357–371. https://doi.org/10.1111/J.1095-8312.1999.TB01896.X

Ficetola, G., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., Taberlet, P., & Pompanon, F. (2010). An In silico approach for the evaluation of DNA barcodes. *BMC Genomics*, *11*(1), 434. https://doi.org/10.1186/1471-2164-11-434

Gagneux, S., & Small, P. M. (2007). Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. *The Lancet Infectious Diseases*, *7*(5), 328–337. https://doi.org/10.1016/s1473-3099(07)70108-1

Ghanavi, H. R., Twort, V., Hartman, T. J., Zahiri, R., & Wahlberg, N. (2022). The (non) accuracy of mitochondrial genomes for family-level phylogenetics in Erebidae (Lepidoptera). *Zoologica Scripta*, *51*(6), 695–707.https://doi.org/10.1111/zsc.12559

Gill, F. B., Mostrom, A. M., & Mack, A. L. (1993). Speciation in North American chickadees: I. Patterns of mtDNA genetic divergence. *Evolution, 47(1)*, 195–212. https://doi.org/10.1111/j.1558-5646.1993.tb01210.x

Gohel, D., & Skintzos, P. (2023). *ggiraph: Make 'ggplot2' Graphics Interactive* (Version 0.8.7) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=ggiraph

Gömöry, D., Zhelev, P., & Brus, R. (2020). The Balkans: A genetic hotspot but not a universal colonization source for trees. *Plant Systematics and Evolution, 306*(1). https://doi.org/10.1007/s00606-020-01647-x

Gratton, P., Marta, S., Bocksberger, G., Winter, M., Trucchi, E., & Kühl, H. (2016). A world of sequences: can we use georeferenced nucleotide databases for a robust automated phylogeography? *Journal of Biogeography*, *44*(2), 475–486. https://doi.org/10.1111/jbi.12786

Held, C. (2014). Chapter 10.5: Phylogeography and population genetics. In C. DeBroyer, et al. (Eds.), *Biogeographic Atlas of the Southern Ocean,* 4 p., ISBN: 978-0-948277-28-3.

Hernangómez, D. (2023). nominatimlite: Interface with Nominatim API Service. https://doi.org/10.5281/zenodo.5113195, https://dieghernan.github.io/nominatimlite/

Hickerson, M. J. (2016). *Biogeography, Evolutionary Theories in*. Elsevier eBooks. https://doi.org/10.1016/b978-0-12-800049-6.00105-0

Jackson, N. D., Morales, A. E., Carstens, B. C., & O'Meara, B. C. (2017). Phrapl: Phylogeographic inference using approximate likelihoods. *Systematic Biology, 66*(6), 1045-1053. https://doi.org/10.1093/sysbio/syx001

Jensen, M. P., FitzSimmons, N. N., Bourjea, J., Hamabata, T., Reece, J., & Dutton, P. H. (2019). The evolutionary history and global phylogeography of the green turtle (Chelonia mydas). *Journal of Biogeography*, *46*(5), 860–870. https://doi.org/10.1111/jbi.13483

Johnson, M. E., & Baarli, B. G. (2015). Charles Darwin in the Cape Verde and Galápagos archipelagos: The role of serendipity in the development of theories on the ups and downs of oceanic islands. *Earth Sciences History*, *34*(2), 220–242. https://doi.org/10.17704/1944-6187-34.2.220

Kandemir, I., Kence, M., & Kence, A. (2000). Genetic and morphometric variation in honeybee (Apis mellifera L.) populations of Turkey. *Apidologie*, *31*(3), 343–356. https://doi.org/10.1051/apido:2000126

Kholodova, M. V. (2009). Comparative phylogeography: Molecular methods, ecological interpretation. *Molecular Biology, 43*(5), 847-854. https://doi.org/10.1134/S002

Kraft, S., Pérez-Álvarez, M., Olavarría, C., & Poulin, E. (2020). Global phylogeography and genetic diversity of the long-finned pilot whale Globicephala melas, with new data from the southeastern Pacific. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-58532-3

Lang, D. T. (2022). XML: Tools for Parsing and Generating XML Within R and S-Plus (Version 3.99-0.12) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=XML

Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences*, *116*(45), 22651–22656. https://doi.org/10.1073/pnas.1911714116

López de Heredia, U., Carrión, J. S., Jiménez, P., Collada, C., & Gil, L. (2007). Molecular and palaeoecological evidence for multiple glacial refugia for evergreen oaks on the Iberian Peninsula. *Journal of Biogeography, 34*, 1505-1517. https://doi.org/10.1111/j.1365-2699.2007.01715.x

Losos, J.B., & Ricklefs, E. R. (2009). Adaptation and diversification on islands. *Nature, 457*, 830–836. https://doi.org/10.1038/nature07893

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res*. 27: 209-220.

Mantooth, S. J., & Riddle, B. R. (2011). Molecular biogeography: The intersection between geographic and molecular variation. *Geography Compass, 5*, 1-20. https://doi.org/10.1111/j.1749-8198.2009.00297.x

Marques, A. C., Maronna, M. M., & Collins, A. G. (2013). Putting GenBank Data on the Map. *Science*, *341*(6152), 1341–1341. https://doi.org/10.1126/science.341.6152.1341-a

McCauley, D. E., Sunby, A. K., Bailey, M. F., & Welch, M. E. (2007). Inheritance of chloroplast DNA is not strictly maternal in Silene vulgaris (Caryophyllaceae): Evidence from experimental crosses and natural populations. *American Journal of Botany*, *94*(8), 1333–1337. https://doi.org/10.3732/ajb.94.8.1333

Miraldo, A., Li, S., Borregaard, M. K., Flórez-Rodríguez, A., Gopalakrishnan, S., Rizvanovic, M., Wang, Z., Rahbek, C., Marske, K. A., & Nogués-Bravo, D. (2016). An Anthropocene map of genetic diversity. *Science*, *353*(6307), 1532–1535. https://doi.org/10.1126/science.aaf4381

Nei, M., & Li, W. H. (1979, October). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, *76*(10), 5269–5273. https://doi.org/10.1073/pnas.76.10.5269

Nevill, P. G., Bossinger, G., & Ades, P. K. (2010). Phylogeography of the world's tallest angiosperm Eucalyptus regnans: Evidence for multiple isolated quaternary refugia. *Journal of Biogeography, 37*, 179–192. https://doi.org/10.1111/j.1365-2699.2009.02193.x

Oksanen, J., et al. (2022). vegan: Community Ecology Package (Version 2.6-4) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=vegan

P. Lumogdang, L., Llameg, M. B., S. Antonio, E., L. Labrador, K., & Jeriel I. Bersaldo, M. (2022). DNA barcoding based on 16S mitochondrial DNA (mtDNA) Molecular Marker of Mangrove Clams from the Selected Sites of Davao Region, Philippines. *Asian Journal of Fisheries and Aquatic Research*, 40–47. https://doi.org/10.9734/ajfar/2022/v16i630391

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal, 10(1)*, 439-446. https://doi.org/10.32614/RJ-2018-009

Pelletier, T. A., Parsons, D. J., Decker, S. K., Crouch, S., Franz, E., Ohrstrom, J., & Carstens, B. C. (2022). phylogatR: Phylogeographic data aggregation and repurposing. *Molecular Ecology Resources*, *22*(8), 2830–2842. https://doi.org/10.1111/1755-0998.13673

Peterson, A. T., Asase, A., Canhos, D., de Souza, S., & Wieczorek, J. (2018). Data Leakage and Loss in Biodiversity Informatics. *Biodiversity Data Journal*, *6*. https://doi.org/10.3897/bdj.6.e26826

Petit, R. J., & Vendramin, G. G. (2007). Plant phylogeography based on organelle genes: An introduction. In S. Weiss & N. Ferrand (Eds.), *Phylogeography of southern European refugia* (pp. 23–97). Springer, Dordrecht, The Netherlands.

Petit, R. J., Csaikl, U., Bordacs, S., et al. (2002b). Chloroplast DNA variation in European white oaks: Phylogeography and patterns of diversity based on data from over 2600 populations. *Forest Ecology and Management*, *176*(1–3), 595–599. https://doi.org/10.1016/s0378-1127(02)00558-3

Petit, R. J., et al. (2002a). Identification of refugia and postglacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *Forest Ecology and Management, 156*, 49-71. https://doi.org/10.1016/s0378-1127(01)00634-x

Pettenkofer, T., Burkardt, K., Ammer, C., et al. (2019). Genetic diversity and differentiation of introduced red oak (Quercus rubra) in Germany in comparison with reference native populations. *European Journal of Forest Research*, *138*(2), 275–285. https://doi.org/10.1007/s10342-019-01167-5

Pons, O., & Petit, R. J. (1996). Measuring and Testing Genetic Differentiation With Ordered Versus Unordered Alleles. *Genetics*, *144*(3), 1237–1245. https://doi.org/10.1093/genetics/144.3.1237

Porter, T. M., & Hajibabaei, M. (2018). Over 2.5 million COI sequences in GenBank and growing. *Plos One*, *13*(9), e0200177. https://doi.org/10.1371/journal.pone.0200177

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Radosavljević, I., Satovic, Z., di Pietro, R., Jug Dujaković, M., Varga, F., Škrtić, D., & Liber, Z. (2022). Phylogeographic structure of common sage (Salvia officinalis L.) reveals microrefugia throughout the Balkans and colonizations of the Apennines. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-20055-4

Riddle, B. R. (2016). Comparative phylogeography clarifies the complexity and problems of continental distribution that drove A. R. Wallace to favor islands. *Proceedings of the National Academy of Sciences*, *113*(29), 7970–7977. https://doi.org/10.1073/pnas.1601072113

Riddle, B. R., & Hafner, D. J. (2010). Integrating pattern with process at biogeographic boundaries: The legacy of Wallace. *Ecography, 33*, 321-325. https://doi.org/10.1111/j.1600-0587.2010.06544.x

Riddle, B. R., Dawson, M. N., Hadly, E. A., Hafner, D. J., Hickerson, M. J., Mantooth, S. J., & Yoder, A. D. (2008). The role of molecular genetics in sculpting the future of integrative biogeography. *Progress in Physical Geography: Earth and Environment, 32(2)*, 173-202. https://doi.org/10.1177/0309133308093822

Rieseberg, L. H., & Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants, 5*, 65–84.

Rusin, M. (2023). Phylogeography of the Western Populations of Stylodipus telum (Rodentia, Dipodidae) based on Mitochondrial DNA. *Zoodiversity*,*57*(1), 13–18. https://doi.org/10.15407/zoo2023.01.013

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Sherry, S., Yankie, L., & Karsch-Mizrachi, I. (2022b). GenBank 2023 update. *Nucleic Acids Research*, *51*(D1), D141–D144. https://doi.org/10.1093/nar/gkac1012

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Sherry, S. T., Yankie, L., & Karsch-Mizrachi, I. (2023). GenBank 2024 Update. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkad903

Sayers, E. W., O'Sullivan, C., & Karsch-Mizrachi, I. (2022a). Using GenBank and SRA. *Plant Bioinformatics*, 1–25. https://doi.org/10.1007/978-1-0716-2067-0_1

Semerikova, S. A., & Semerikov, V. L. (2014). Mitochondrial DNA Variation and Reticulate Evolution of the Genus Abies. *Russian Journal of Genetics*. *50*(4), 366–377. https://doi.org/10.1134/S1022795414040139

Shepherd, L. D., de Lange, P. J., Perrie, L. R., & Heenan, P. B. (2017). Chloroplast phylogeography of New Zealand Sophora trees (Fabaceae): Extensive hybridization and widespread Last Glacial Maximum survival. *Journal of Biogeography, 44*, 1640–1651. https://doi.org/10.1111/jbi.12963

Sidlauskas, B., Ganapathy, G., Hazkani-Covo, E., Jenkins, K. P., Lapp, H., McCall, L. W., Price, S., Scherle, R., Spaeth, P. A., & Kidd, D. M. (2009). Linking Big: The Continuing Promise Of Evolutionary Synthesis. *Evolution*, *64*(4), 871–880. https://doi.org/10.1111/j.1558-5646.2009.00892.x

Simeone, M. C., Grimm, G. W., Papini, A., Vessella, F., Cardoni, S., Tordoni, E., Piredda, R., Franc, A., & Denk, T. (2016). Plastome data reveal multiple geographic origins of Quercus Group Ilex. *PeerJ*, *4*, e1897. https://doi.org/10.7717/peerj.1897

Simeone, M. C., Piredda, R., Papini, A., Vessella, F., & Schirone, B. (2013). Application of plastid and nuclear markers to DNA barcoding of Euro-Mediterranean oaks (Quercus, Fagaceae): Problems, prospects and phylogenetic implications. *Botanical Journal of the Linnean Society, 172*, 478-499. https://doi.org/10.1111/boj.12059

Simmons, M. P., & Ochoterena, H. (2000). Gaps as Characters in Sequence-Based Phylogenetic Analyses. *Systematic Biology*, *49*(2), 369–381. https://doi.org/10.1093/sysbio/49.2.369

Sulloway, F. J. (1984). Darwin and the Galápagos. *Biological Journal of the Linnean Society, 21*, 29-59. https://doi.org/10.1111/j.1095-8312.1984.tb02052.x

Tamaki, I., & Okada, M. (2014). Genetic admixing of two evergreen oaks, Quercus acuta and Q. sessilifolia (subgenus Cyclobalanopsis), is the result of interspecific introgressive hybridization. *Tree Genetics & Genomes, 10*, 989–999. https://doi.org/10.1007/s11295-014-0737-x

Tekpinar, A. D., Aktaş, C., Kansu, I., Duman, H., & Kaya, Z. (2021). Phylogeography and phylogeny of genus Quercus L. (Fagaceae) in Turkey implied by variations of trnT(UGU)-L(UAA)-F (GAA) chloroplast DNA region. *Tree Genetics & Genomes*, *17*(5). https://doi.org/10.1007/s11295-021-01522-x

Tiffney, B. H. (2008). Phylogeography, fossils, and northern hemisphere biogeography: The role of physiological uniformitarianism. *Annals of the Missouri Botanical Garden*, *95*(1), 135–143. https://doi.org/10.3417/2006199

Torroni, A., Achilli, A., Olivieri, A., & Semino, O. (2020). Haplogroups and the history of human evolution through mtDNA. *The Human Mitochondrial Genome*, 111–129. https://doi.org/10.1016/B978-0-12-819656-4.00005-X

Trewick, S. (2017). Plate tectonics in biogeography. In D. Richardson, N. Castree, M. F. Goodchild, A. Kobayashi, W. Liu, & R. A. Marston (Eds.), *International Encyclopedia of Geography: People, the Earth, Environment, and Technology*. https://doi.org/10.1002/9781118786352.wbieg0638

Vidya, T., Sukumar, R., & Melnick, D. J. (2008). Range-wide mtDNA phylogeography yields insights into the origins of Asian elephants. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1658), 893–902. https://doi.org/10.1098/rspb.2008.1494

Vinga, S., & Almeida, J. (2003). Alignment-free sequence comparison—a review. *Bioinformatics, 19(4)*, 513–523. https://doi.org/10.1093/bioinformatics/btg005

Vitelli, M., Vessella, F., Cardoni, S., Pollegioni, P., Denk, T., Grimm, G. W., & Simeone, M. C. (2017). Phylogeographic structuring of plastome diversity in Mediterranean oaks (Quercus Group Ilex, Fagaceae). *Tree Genetics & Genomes*, *13*(1). https://doi.org/10.1007/s11295-016-1086-8

Wallace, A. (2011). *The Geographical Distribution of Animals: With a Study of the Relations of Living and Extinct Faunas as Elucidating the Past Changes of the Earth's Surface* (Cambridge Library Collection - Zoology). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139097116

Wallberg, A., Han, F., Wellhagen, G., Dahle, B., Kawata, M., Haddad, N., Simões, Z. L. P., Allsopp, M. H., Kandemir, I., De la Rúa, P., Pirk, C. W., & Webster, M. T. (2014). A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee Apis mellifera. *Nature Genetics*, *46*(10), 1081–1088. https://doi.org/10.1038/ng.3077

Weider, L. J., Hobæk, A., Colbourne, J. K., Crease, T. J., Dufresne, F., & Hebert, P. D. N. (1999). Holarctic phylogeography of an asexual species complex I. Mitochondrial DNA variation in Arctic Daphnia. *Evolution*, *53*(3), 777–792. https://doi.org/10.1111/J.1558-5646.1999.TB05372.X

Wilkinson, S.P. (2018) kmer: an R package for fast alignment-free clustering of biological sequences (Version 1.0.0) [R package] https://cran.r-project.org/package=kmer

Winter, D. J. (2017). rentrez: An R package for the NCBI eUtils API. *R Journal, 9*, 520.

Wöhrmann, T., Michalak, I., Zizka, G., & Weising, K. (2020). Strong genetic differentiation among populations of Fosterella rusbyi (Bromeliaceae) in Bolivia. *Botanical Journal of the Linnean Society, 192(4)*, 744–759.

Xie, Y., Cheng, J., & Tan, X. (2023). DT: A Wrapper of the JavaScript Library 'DataTables' (Version 0.28) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=DT

Zahiri, R., Christian Schmidt, B., Schintlmeister, A., Yakovlev, R. V., & Rindoš, M. (2019). Global phylogeography reveals the origin and the evolutionary history of the gypsy moth (Lepidoptera, Erebidae). *Molecular Phylogenetics and Evolution*, *137*, 1–13. https://doi.org/10.1016/j.ympev.2019.04.021

Zhang, R., Hipp, A. L., & Gailing, O. (2015). Sharing of chloroplast haplotypes among red oak species suggests interspecific gene flow between neighboring populations. *Botany*, *93*(10), 691–700. https://doi.org/10.1139/cjb-2014-0261

Zheng, S., Li, Y., Yang, X., Chen, J., Hua, J., & Gao, Y. (2018). DNA barcoding identification of Pseudococcidae (Hemiptera: Coccoidea) using the mitochondrial COI gene. *Mitochondrial DNA Part B*, *3*(1), 419–423 https://doi.org/10.1080/23802359.2018.1457988

Zhong, K. L., Song, X. H., Choi, H. G., Satoshi, S., Weinberger, F., Draisma, S. G. A., Duan, D. L., & Hu, Z. M. (2020).MtDNA-Based Phylogeography of the Red Alga Agarophyton vermiculophyllum (Gigartinales, Rhodophyta) in the Native Northwest Pacific. *Frontiers in Marine Science,7*. https://doi.org/10.3389/fmars.2020.00366

Zimmerer, K. (1994). Human geography and the "new ecology": The prospect and promise of integration. *Annals of the Association of American Geographers, 84*(1), 108-125. https://doi.org/10.1111/j.1467-8306.1994.tb01731.x

CURRICULUM VITAE

Surname, Name: Aktaş, Caner

EDUCATION

Degree	**Institution**	**Year of Graduation**
MS	METU Biology	2010
BS	Hacettepe University Biology	2006
High School	TED Ankara College Foundation Schools, Ankara	1999

FOREIGN LANGUAGES

English

PUBLICATIONS

1. Aktas, C. (2020). haplotypes: Manipulating DNA Sequences and Estimating Unambiguous Haplotype Network with Statistical Parsimony (Version 1.1.3) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=haplotypes

2. Tekpinar, A. D., Aktaş, C., Kansu, I., Duman, H., & Kaya, Z. (2021). Phylogeography and phylogeny of genus Quercus L. (Fagaceae) in Turkey implied by variations of trnT(UGU)-L(UAA)-F (GAA) chloroplast DNA region. *Tree Genetics & Genomes*, *17*(5). https://doi.org/10.1007/s11295-021-01522-x