

Metagenomika i filogenetyka molekularna: Autotroficzne eugleniny w małych zbiornikach wodnych

Julia Smolik¹

¹ js406162@students.mimuw.edu.pl

29 maja 2023

1. Wstęp

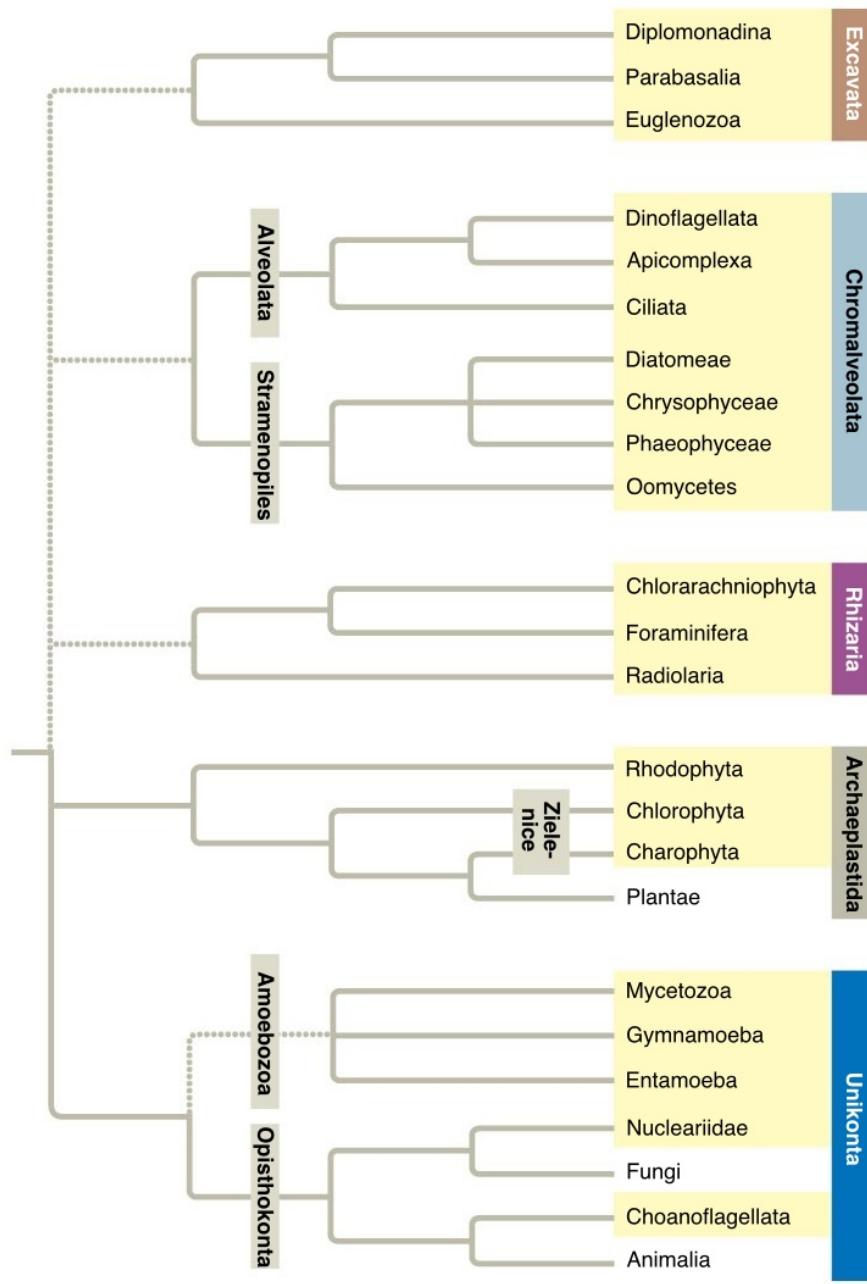
Pierwotniaki to nieformalnie nazywana, zróżnicowana, w większości jednokomórkowa grupa eukariontów [1]. Są one badane przez naukowców od ponad 300 lat. Wydawałoby się, że tak długie czas obserwacji pozwolił na odkrycie reprezentatywnej próby żyjących gatunków pierwotniaków. Jednak prowadzone przeszukiwanie genetyczne pozwoliły odkryć liczne, nieznane wcześniej pierwotniaki, a także zagłębić się w filogenezę tej grupy.

Wszystkie pierwotniaki były kiedyś przypisywane do jednego królestwa - Protista. Postępy w systematyce eukariontów pozwoliły na rozbicie tego królestwa. Okazało się, że Protista są w rzeczywistości polifiletyczne: niektóre pierwotniaki są bliżej spokrewnione z roślinami, grzybami lub zwierzętami niż z innymi pierwotniakami. W rezultacie królestwo Protista zostało zlikwidowane, a różne linie ewolucyjne pierwotniaków podnoszone są do rangi królestwa. Pierwotniaki, razem z roślinami, zwierzętami i grzybami są klasyfikowane jako eukarioty i tworzą one domenę Eukarya, jedną z trzech domen życia.

Ponieważ pierwotnie nazywana grupa Protista jest polifiletyczna, pierwotniakom można przypisać jedynie kilka wspólnych cech bez wymieniania wyjątków. W rzeczywistości pierwotniaki przejawiają największą różnorodność strukturalną i funkcjonalną ze wszystkich pozostałych grup eukariontów.

Większość pierwotniaków to organizmy jednokomórkowe, chociaż istnieją gatunki kolonijne i wielokomórkowe. Jednokomórkowe pierwotniaki są uważane za najprostsze eukarienty, jednak na poziomie komórkowym wiele pierwotniaków wykazuje bardzo złożoną budowę, wręcz najbardziej skomplikowaną ze wszystkich komórek. Pierwotniaki są najbardziej zróżnicowaną pod względem odżywiania grupą eukariontów. Niektóre z nich są fotoautotrofami i zawierają chloroplasty (wtórna endosymbioza [2, 3]). Inne są heterotrofami, które wchłaniają cząstki organiczne lub większe cząstki pokarmowe. Jeszcze inne są nazywane miksotrofami. Łącznie one fotosyntezę i odżywianie heterotroficzne. Pierwotniaki różnią się również sposobem rozmnażania i cyklem rozwojowym. Niektóre z nich są wyłącznie bezpłciowe, a inne mogą również rozmnażać się płciowo lub przynajmniej wykorzystywać procesy płciowe, takie jak mejoza i zapłodnienie.

Interesującą w tej analizie grupą pierwotniaków jest Excavata. Jest to klad, który został niedawno zaproponowany na podstawie wyników analiz morfologicznych cytoszkieletu. Niektóre organizmy należące do tej zróżnicowanej grupy mają "wyzłobiony" rowek pokarmowy z jednej strony ciała. Do Excavata zaliczane są Diplomonadina, Parabasalia i Euglenozoa. Dane molekularne wskazują, że każda z tych trzech grup jest monofiletyczna, ale dane te ani nie potwierdziły, ani nie wykluczyły hipotezy o monofiletyzmie całej supergrupy Excavata. Wsparcie dla kladu Excavata jest relatywnie słabe, co czyni go najbardziej kontrowersyjną supergrupą eukariontów (Rys. 1.1).



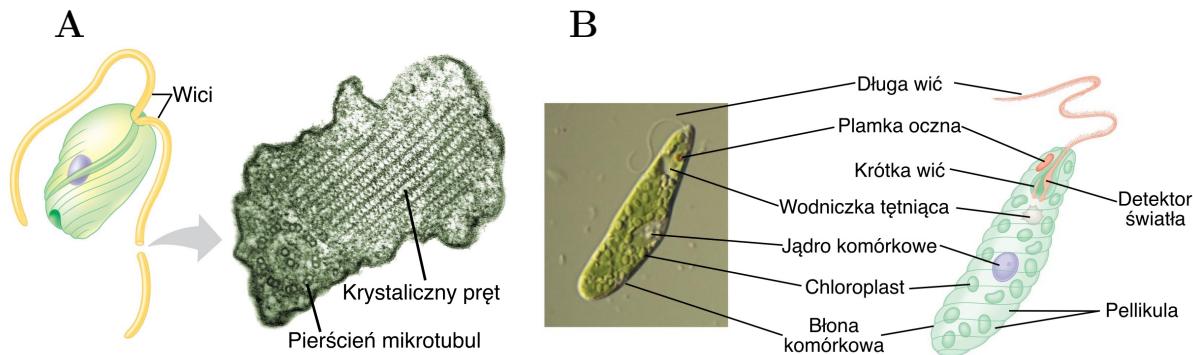
Rysunek 1.1: **Hipoteza filogenetyczna eukariontów.** Grupy eukariotyczne na szczytach gałęzi odnoszą się do supergrup, które są oznaczone pionowo po prawej stronie drzewa. Ze starego systemu klasyfikacji, opartego na pięciu królestwach, przetrwały królestwa Plantae (rośliny lądowe), Fungi (grzyby) i Animalia (zwierzęta). Grupy, które wcześniej klasyfikowano w królestwie Protista (pierwotniaki), są zaznaczone na żółto. Linie przerywane wskazują niepewne związki ewolucyjne, na temat których trwa dyskusja. Źródło: [1]

Główną cechą morfologiczną, która wyróżnia pierwotniaki należące do Euglenozoa, jest obecność wewnętrznych spiralnego lub krystalicznego pręcika o nieznanej funkcji. Dwie najlepiej poznane grupy Euglenozoa to Kinetoplastida i Euglenida.

Euglenidy (Excavata, Discoba, Euglenozoa, Euglenida) to grupa wolno żyjących, jednokomórkowych organizmów zaliczanych do nieformalnej grupy wiciowców żyjących w środowiskach wodnych [4]. Wspólną i unikalną cechą morfologiczną euglenidów jest obecność pokrycia komórkowego zwanego pellikulą. Jest to złożona struktura składająca się z białkowych pasków, które są pokryte błoną komórkową, a pod nią znajduje się system mikrotubul oraz cysterny siateczki endoplazmatycznej. Pierwotniaki należące do Euglenida na

jednym końcu ciała mają węlebienie (rezerwuar), z którego wychodzą jedna lub dwie wici [3] (Rys. 1.2).

U euglenidów obserwuje się kilka różnych sposobów odżywiania. Większość gatunków to heterotrofy, ale wśród tej grupy można znaleźć również organizmy fotoautotroficzne i miksoautotroficzne, zawierające plastidy. Zmiany trybu odżywiania pierwotniaków można śledzić na drzewach filogenetycznych. Wiele gatunków należących do Euglenida rodzaju *Euglena* (klejnotka) to miksoautotrofy: przy świetle słonecznym są autotrofami, ale przy braku dostępu światła stają się heterotrofami, absorbując organiczne substancje pokarmowe z otaczającego je środowiska. Wiele innych pierwotniaków należących do Euglenida wchłania ofiary poprzez fagocytozę. Większość wiedzy o grupie protistów zwanych euglenidami (Euglenida) pochodzi z szeroko zakrojonych badań fotosyntetycznego gatunku modelowego *Euglena gracilis*.



Rysunek 1.2: (A) Wić u Euglenozoa. Większość Euglenozoa ma krystaliczny pręt wewnętrz jednej z wici. Pręt biegnie wzdu 9+2 pierścieni mikrotubul występujących we wszystkich wiciach eukariotycznych. (B) *Euglena* (klejnotka). Przedstawiciel Euglenozoa powszechnie spotykany w stawach. Źródło: [1]

Różnorodność morfologiczna i behawioralna euglenidów dostarcza również przekonujących ilustracji głównych wydarzeń w ewolucji, takich jak efekty wtórnej endosymbiozy i zmiany w podstawowych mechanizmach rozwojowych [2].

Od 150 lat naukowcy badają *Euglenę* wyłącznie na podstawie cech morfologicznych, co pozwoliło na utworzenie setek opisów nowych taksonów i wielu sztucznych systemów klasyfikacji wewnętrzrodzajowej. Pomimo zaangażowania w odkrywanie filogenezy *Eugleny*, nadal pozostaje ona grupą polifiletyczną.

Pojawienie się filogenetyki molekularnej i połączenie danych molekularnych i morfologicznych wywarło ogromny wpływ na rozumienie pokrewieństw w obrębie rodzaju *Euglena* i całej grupy euglenidów. Porównawcze badania morfologiczne i molekularne wykazały, że wiele taksonów z rodzaju *Euglena* opisywanych w literaturze nie ma uzasadnienia takonomicznego. Dla wielu gatunków stworzono klucze do prawidłowej identyfikacji i uporządkowano skomplikowane kwestie nomenklaturowe oraz opisano dwa nowe gatunki (*E. pseudostellata*, *E. pseudochadefaudii*).

2. Cel analiz

W latach 2017 - 2019 prowadzono badania mające na celu przeanalizowanie składu taksonomicznego i różnorodności biologicznej autotroficznych euglenin w kilkunastu niewielkich zbiornikach wodnych zlokalizowanych w różnych regionach Polski.

Próbki wody z wybranych zbiorników były pobierane czterokrotnie w ciągu sezonu wegetacyjnego w latach 2017, 2018 i 2019. Pobrany materiał był osadzanym na filtrach, z których następnie izolowano całkowity DNA. Następnie przeprowadzano amplifikację rejonu V2 18S rDNA z użyciem starterów specyficznych dla euglenin. Otrzymane amplikony poddano sekwencjonowaniu wysokoprzepustowemu i otrzymano pliki w formacie fastq z odczytami przypisanymi do poszczególnych próbek.

Celem pracy jest badanie i opis jakości danych, a następnie przeanalizowanie składu taksonomicznego za pomocą różnych statystyk, wykresów oraz analiz α -różnorodności i β -różnorodności.

3. Materiały i metody

Wykonanie analiz podzielono na dwa etapy. Pierwszy z nich polegał na manipulacji danymi tak, aby otrzymać przefiltrowane i przycięte odczyty, które następnie przyporządkowywano taksonomicznie. Wykorzystano w tym celu oprogramowanie QIIME [5]. Następnie wykonano analizę uzyskanych danych za pomocą skryptu napisanego w R. Przeprowadzono analizę α - i β -różnorodności, a także analizę składu taksonomicznego. Otrzymane wyniki przedstawiono za pomocą różnorodnych wykresów.

3.1. Analiza i opis jakości danych

Otrzymane dane składają się z 48 plików z odczytami przypisanymi do poszczególnych próbek. Pierwsza połowa plików (24 pliki) zawiera odczyty *forward*, a druga połowa (24 pliki) odczyty *reverse*. Próbki pobierano trzykrotnie w roku 2017 i trzykrotnie w roku 2019 z czterech zbiorników: Ceglów, Zabłotnia, Tały i Urwitał. Z każdego zbiornika zebrano łącznie po 6 próbek (każda próbka opisana jest przez odczyty *forward* i *reverse*), co w sumie dawało 24 próbki gotowe do analizy.

Ocena jakości danych przeprowadzona była za pomocą komendy korzystającej z narzędzia FastQC [6]:

```
fastqc nazwapliku.fastq
```

gdzie:

nazwapliku.fastq - pojedynczy plik z odczytami

Pliki wynikowe z działania programu FastQC łączone były w jedno wspólne podsumowanie za pomocą komendy korzystającej z narzędzia MultiQC [7]:

```
multiqc nazwa_folderu_z_wynikami --replace-names rename_samples.tsv
```

gdzie:

nazwa_folderu_z_wynikami - folder, w którym znajdowały się pliki wynikowe programu FastQC

rename_samples.tsv - nazwa pliku tsv ze zmienionymi nazwami próbek

3.2. Przygotowanie danych do analizy

Do zimportowania plików z odczytami do programu QIIME [5], należało przygotować plik z instrukcją importu (Rozdział 6: Dostępność kodu i danych). Zimportowanie plików do programu QIIME wykonano używając komendy:

```
qiime tools import \
--type 'SampleData[PairedEndSequencesWithQuality]' \
--input-format PairedEndFastqManifestPhred33 \
--input-path seqs_import.csv \
--output-path output_reads.qza
```

gdzie:

seqs_import.csv - nazwa pliku csv, w którym znajduje się instrukcja do importowania sekwencji

output_reads.qza - nazwa pliku wynikowego z sekwencjami

Następnie wykonano filtrowanie i przycinanie odczytów za pomocą komendy wykorzystującej oprogramowanie DADA2 [8]:

```
qiime dada2 denoise-paired \
--i-demultiplexed-seqs output_reads.qza \
--p-trunc-len-f liczba1 \
--p-trunc-len-r liczba2 \
--p-trim-left-f 23 \
--p-trim-left-r 20 \
--p-chimera-method consensus \
--p-trunc-q liczba3 \
```

```
--o-denoising-stats stats.qza \
--o-representative-sequences representatives.qza \
--o-table table.qza \
--verbose
```

gdzie:

output_reads.qza - nazwa pliku z importowanymi sekwencjami

liczba1 - liczba nukleotydów pozostała po przycięciu odczytów *forward*

liczba2 - liczba nukleotydów pozostała po przycięciu odczytów *reverse*

23 - długość starterów odczytów *forward*

20 - długość starterów odczytów *reverse*

consensus - metoda odrzucania sekwencji chimerowych

liczba3 - przycinanie na podstawie jakości

stats.qza - nazwa pliku wynikowego ze statystykami analizy

representatives.qza - nazwa pliku wynikowego ze przefiltrowanymi sekwencjami reprezentującymi

table.qza - nazwa pliku wynikowego z tabelą podsumowującą wyniki filtrowania

Oba typy odczytów zawierały sekwencje starterów (*forward*: CTGTGAATGGCTCCTACATCAG, *reverse*: CTSCCTCTCCGGAATCRAAC). Z początków sekwencji *forward* i *reverse* zdecydowano się zatem odciąć odpowiednio: 23 i 20 nukleotydów (długości starterów).

Na podstawie analizy jakości odczytów (Rozdział 4.1: Analiza i opis jakości danych, Rys. 4.2) rozpatrywano dla odczytów *forward* i *reverse* po dwie długości, do których należało przyciąć te odczyty.

Dla odczytów *forward* wybrano długości: 260, ponieważ od tej długości jakość odczytów zaczynała powoli maleć i 232, ponieważ dla tej wartości wszystkie odczyty miały jeszcze dość wysoki poziom jakości. Dla odczytów *reverse* wybrano długości: 242, ponieważ była to ostatnia wartość długości, gdzie wszystkie odczyty osiągały dobrą jakość i dodatkowo zapewniały najmniejszą stratę danych, oraz 212, ponieważ od tej długości jakość odczytów zaczynała maleć.

Sprawdzono również dwie wartości parametru współczynnika *q*, na podstawie którego przeprowadzane było przycinanie według jakości. Zdecydowano się na badanie wartości *q* = 3 oraz *q* = 7.

Przycinanie i filtrowanie odczytów wykonano osmiokrotnie, za każdym wykorzystując inną kombinację wybranych parametrów.

Zauważono, że zdecydowanie lepsze wyniki przycinania i filtrowania odczytów otrzymuje się przy niższych wartościach parametru *q* i przy większym przycinaniu sekwencji. Sprawdzono zatem, czy bardziej rygorystyczne przycięcie odczytów poprawi rezultaty i przeprowadzono jedno dodatkowo przycinanie i filtrowanie. Odczyty *forward* i *reverse* przycięto do długości odpowiednio: 232 i 197, a wartość parametru *q* ustawniono na 3.

Na podstawie wyników przycinania i filtrowania odczytów z użyciem różnych kombinacji parametrów (Rozdział 4.2: Przygotowanie danych do analizy), dalsze analizy zdecydowano się przeprowadzić na odczytach przyciętych do długości 232 i 212 (odpowiednio: *forward* i *reverse*) i parametrem *q* = 3.

3.3. Przyporządkowanie taksonomiczne

Następnym etapem analiz była klasyfikacja taksonomiczna sekwencji. Przyporządkowanie taksonomiczne wykonano za pomocą komendy korzystającej z oprogramowania QIIME i metody classify-sklearn [5,8–11]:

```
qiime feature-classifier classify-sklearn \
--i-classifier klasyfikator_Eu_metfilo.qza \
--i-reads przefiltrowane_sekwencje.qza \
--o-classification taxonomy.qza
```

gdzie:

klasyfikator_Eu_metfilo.qza - plik z klasyfikatorem

przefiltrowane_sekwencje.qza - plik z przefiltrowanymi sekwencjami reprezentatywnymi

taxonomy.qza - plik z wynikiem przyporządkowania taksonomicznego

Następnie przy użyciu otrzymanego przyporządkowania taksonomicznego i utworzonych metadanych (Rozdział 6: Dostępność kodu i danych) wykonano wykres słupkowy przedstawiający skład taksonomiczny w próbkach za pomocą komendy korzystającej z oprogramowania QIIME i paczki biomtable [5,8–11]:

```
qiime taxa barplot \  
--i-table table.qza \  
--i-taxonomy taxonomy.qza \  
--m-metadata-file sample_metadata.tsv  
--o-visualization barplot.qzv
```

gdzie:

table.qza - tabela powstała podczas filtrowania sekwencji
taxonomy.qza - plik z wynikiem przyporządkowania taksonomicznego
sample_metadata.tsv - plik z metadanymi dotyczącymi analizowanych próbek
barplot.qzv - plik z wykresem przedstawiającym skład gatunkowy w próbkach

Do dalszych etapów analiz niezbędne było wykonanie drzewa filogenetycznego reprezentatywnych (ASV - ang. *Amplicon Sequence Variant*). Najpierw jednak należało wykonać uliniowienie sekwencji. Wykorzystano komendę używającą oprogramowania QIIME [5] i metody mafft [12]:

```
qiime alignment mafft \  
--i-sequences representatives.qza \  
--o-alignment aligned.qza
```

gdzie:

representatives.qza - sekwencje reprezentatywne
aligned.qza - plik z wynikiem uliniowienia sekwencji

Wykonane uliniowienie następnie przyjęto:

```
qiime alignment mask \  
--i-alignment aligned.qza \  
--o-masked-alignment aligned_mask.qza
```

gdzie:

aligned.qza - plik z uliniowionymi sekwencjami
aligned_mask.qza - plik z wynikiem przycinania uliniowienia

Drzewo filogenetyczne wykonano wykorzystując komendę wykorzystującą oprogramowanie QIIME [5] i algorytm FastTree [13]:

```
qiime phylogeny fasttree \  
--i-alignment aligned_mask.qza \  
--o-tree tree.qza
```

gdzie:

aligned_mask.qza - plik z przyjętym uliniowieniem
tree.qza - plik z wynikowym drzewem

Utworzono drzewo ukorzeniono:

```
qiime phylogeny midpoint-root \  
--i-tree tree.qza \  
--o-rooted-tree rooted_tree.qza
```

gdzie:

tree.qza - plik z drzewem
rooted_tree.qza - plik z wynikiem ukorzeniania drzewa

3.4. Analiza danych w R

Sprawdzono podstawowe informacje o danych wczytanych do skryptu w R. Zbadano liczbę taksonów i próbek, maksymalną i minimalną liczbę odczytów w próbce oraz poziomy taksonomiczne za pomocą komend odpowiednio:

```
ntaxa(phyloseq)  
nsamples(phyloseq)  
max(sample_sums(phyloseq))  
min(sample_sums(phyloseq))  
rank_names(phyloseq)
```

gdzie:

phyloseq - zmienna, w której zapisany jest obiekt Phyloseq zawierający informację o klasyfikacji taksonomicznej, ukorzenionym drzewie filogenetycznym i metadanych

Następnie dane przefiltrowano, zostawiając tylko te odczyty, które należały do przedstawicieli Euglenida za pomocą komendy

```
subset_taxa(phyloseq, Order == "Euglenida")
```

gdzie:

phyloseq - zmienna, w której zapisany jest obiekt Phyloseq zawierający informację o klasyfikacji taksonomicznej, ukorzenionym drzewie filogenetycznym i metadanych

Po przefiltrowaniu odczytów ponownie sprawdzono podstawowe informacje o danych. Następnie sprawdzono wczytane przyporządkowanie taksonomiczne i zbadano ile przyporządkowań na poziomie gatunku jest unikatowych. Ponieważ niewiele z nich było nieznanych (mało wartości NaN) i było tylko 131 unikatowych gatunków (dużo się powtarzało), postanowiono połączyć ASV z tych samych gatunków. Pozwoliło to na usprawnienie obliczeń i umożliwiło bardziej czytelną wizualizację wyników, przy niewielkiej stracie danych. Połączenie ASV wykonano za pomocą komendy:

```
tax_glom(phyloseq_euglenida, taxrank = "Species")
```

gdzie:

phyloseq_euglenida - zmienna, w której znajdują się odczyty, które należały do przedstawicieli Euglenida
taxrank - poziom taksonomiczny, do którego ma być wykonane połączenie

Po wykonaniu łączenia sekwencji, ponownie sprawdzono podstawowe informacje o danych.

Następne kroki analiz wymagały wyrównania liczebności odczytów w próbkach. Aby zwizualizować liczebność odczytów w każdej próbce, postanowiono wykonać wykres krzywej wysycenia. Posłużyono się komendą:

```
tab = data.frame(t(otu_table(phyloseq_glom)))  
rarecurve(tab, step=50, cex=0.9, label=FALSE, col=rainbow(24), lwd = 3)
```

gdzie:

phyloseq_glom - zmienna, w której znajdują się połączone ASV z tych samych gatunków

tab - transponowana tabela z liczbą odczytów w próbkach

step - liczba kroków

cex - rozmiar czcionki

label - wyświetlenie etykiet z nazwami próbek

col - kolor linii

lwd - grubość linii

Na podstawie wykresu krzywej wysycenia postanowiono przyciąć liczebność odczytów w każdej z próbek do wartości najmniej licznej próbki, a następnie ponownie sprawdzić podstawowe informacje o danych. Ograniczanie liczebności odczytów wykonano za pomocą komendy:

```
rarefy_even_depth(phyloseq_glom, sample.size=min(sample_sums((phyloseq_glom))), replace=F)
```

gdzie:

phyloseq_glom - zmienna, w której znajdują się połączone ASV z tych samych gatunków

sample.size - liczba kroków (liczba odczytów po przycięciu)

replace - informacja, czy losowanie ma się odbywać ze zwracaniem (wybrano opcję bez zwracania)

3.5. Analiza α -różnorodności

Sprawdzono trzy różne wskaźniki α -różnorodności: obserwowane OTU (Observed) - jakościowa miara bogactwa zespołu, wskaźnik różnorodności Shannona (ilościowa miara różnorodności zespołu), który bierze pod uwagę liczbę gatunków i równomierność ich liczebności, oraz wskaźnik różnorodności Simpsona, który nie uwzględnia liczebności gatunków. Wykorzystano w tym celu komendę:

```
diversity <- estimate_richness(phyloseq_rare, measures = c("Observed", "Shannon", "Simpson"))
data_alpha <- cbind(sample_data(phyloseq_rare), diversity)
```

gdzie:

`phyloseq_rare` - zmienna, w której znajdują się próbki, z wyrównaną liczebnością odczytów
`measures` - wybrane wskaźniki α -różnorodności

Wizualizację wybranych wskaźników α -różnorodności wykonano za pomocą komendy:

```
plot_richness(phyloseq_rare, measures = c("Observed", "Shannon", "Simpson"),
x = "rok", color = "zbiornik") + geom_boxplot(aes(fill = region), alpha=0.2)
+ scale_color_manual(values = c25) + scale_fill_manual(values = c25)
+ scale_x_continuous(breaks = scales::pretty_breaks(n = 2), expand=c(0.01, 0))
+ ggtitle("Alfa-różnorodność")
+ theme(plot.title = element_text(size = 20, face = "bold"),
text = element_text(size=20),
axis.text.x = element_text(angle=45, hjust=0.5))
```

gdzie:

`phyloseq_rare` - zmienna, w której znajdują się próbki, z wyrównaną liczebnością odczytów
`measures` - wybrane wskaźniki α -różnorodności

`x` - zmienna, według której wykonane jest łączenie próbek

`color` - kolorowanie według zmiennej

`geom_boxplot` - dodanie wykresu pudełkowego

`scale_color_manual` - kolor punktów i linii

`scale_fill_manual` - kolor wypełnień

`scale_x_continuous` - zmiana wyglądu osi OX

`ggtitle` - dodanie tytułu wykresu

`theme` - wygląd elementów wykresu

Następnie wykonano analizę statystyczną różnic między próbками z różnych lat, zbiorników i regionów Polski w zależności od wybranego wskaźnika α -różnorodności. Wykorzystano w tym celu komendę:

```
summary(aov(wskaźnik ~ zmienna, data = data_alpha))
```

gdzie:

`wskaźnik` - nazwa wybranego wskaźnika α -różnorodności

`zmienna` - zmienna grupująca

`data_alpha` - zmienna, w której zapisana jest analiza α -różnorodności wybranymi wskaźnikami

Analizę statystyczną wykonano dla każdej kombinacji danego wskaźnika α -różnorodności z wybranymi zmiennymi grupującymi.

3.6. Analiza β -różnorodności

W celu zbadania β -różnorodności, najpierw należało policzyć macierz odległości mierzącą różnice między próbками. Wybrano obliczanie odległości metodą Bray-Curtisa, Unifraq i Jaccarda. Posłużono się komendami odpowiednio:

```
distance(phyloseq_rare, method="bray")
distance(phyloseq_rare, method="unifrac")
distance(phyloseq_rare, method="jaccard", binary = TRUE)
```

gdzie:

phyloseq_rare - zmienna, w której znajdują się próbki, z wyrównaną liczebnością odczytów
method - nazwa wybranej metody do obliczenia macierzy odległości
binary - informacja o binarności

Każda z macierzy odległości następnie zwizualizowano jako mapę ciepła. Wykorzystano komendę:

```
plot_dist_as_heatmap(odleglosc, title = "Mapa ciepła macierzy odległości")
```

gdzie:

odleglosc - zmienna, w której zapisana jest macierz odległości
title - tytuł wykresu

Na podstawie macierzy odległości wykonano ordynacje metodą NMDS (ang. *Nonmetric MultiDimensional Scaling*), które później zwizualizowano. Wykorzystano komendy:

```
ord = ordinate(phyloseq_rare, method = "NMDS", distance = "nazwa_metody")
plot_ordination(phyloseq_rare, ord, color="region", label = "rok")
+ scale_color_manual(values = c25)+ scale_fill_manual(values = c25)
+ theme_light()
+ geom_point(aes(color = region, shape = pora_roku), alpha = 0.9, size = 3.5)
+ stat_ellipse(geom = "polygon", aes(fill = zbiornik), alpha = 0.2, linetype =0)
+ ggtitle("Ordynacja NMDS")
```

gdzie:

phyloseq_rare - zmienna, w której znajdują się próbki, z wyrównaną liczebnością odczytów
method - nazwa metody do wykonania ordynacji
distance - nazwa metody używanej do obliczenia macierzy odległości
color - zmienna, po której będą pokolorowane punkty
label - zmienna, po której będą podpisane punkty
scale_color_manual - kolor punktów i linii
scale_fill_manual - kolor wypełnień
theme_light - zapewnia jasny motyw wykresu
geom_point - zmienia wygląd punktów na wykresie
stat_ellipse - dodaje elipsy

Istotność statystyczną różnic na wykresach ordynacji zbadano używając dwóch metod: ANOSIM i PERMANOVA.

Test ANOSIM wykonano komendą:

```
zmienna = get_variable(phyloseq_rare, "nazwa_zmiennej")
anosim(odleglosc, grouping = zmienna)
```

gdzie:

phyloseq_rare - zmienna, w której znajdują się próbki, z wyrównaną liczebnością odczytów
nazwa_zmiennej - nazwa zmiennej, o której ma być przechowywana informacja
odleglosc - zmienna, w której zapisana jest macierz odległości

Test PERMANOVA wykonano komendą:

```
metadane <- data.frame(sample_data(phyloseq_rare))
adonis(odleglosc ~ zmienna, data = metadane)$aov.tab
```

gdzie:

phyloseq_rare - zmienna, w której znajdują się próbki, z wyrównaną liczebnością odczytów
odleglosc - zmienna, w której zapisana jest macierz odległości
zmienna - zmienna grupująca

3.7. Analiza składu taksonomicznego

Analizę składu taksonomicznego przedstawiono w sposób graficzny za pomocą 3 rodzajów wykresów: wykresu słupkowego, mapy ciepła oraz drzewa filogenetycznego z zaznaczonymi na gałęziach informacjami o próbkach.

Wykres słupkowy wykonano komendą:

```
plot_bar(phyloseq_rare, fill = "Genus", x="rok")
+ geom_bar(stat="identity")
+ theme_bw()
+ facet_wrap("zbiornik", scales="free")
+ scale_fill_manual(values=c25)
+ scale_x_continuous(breaks = scales::pretty_breaks(n = 3), expand=c(0.01, 0))
+ ggtitle("Wykres słupkowy składu taksonomicznego")
+ theme(plot.title = element_text(size = 20, face = "bold"),
text = element_text(size=20),
axis.text.x = element_text(angle=0, hjust=0.5))
```

gdzie:

`phyloseq_rare` - zmienna, w której znajdują się próbki, z wyrównaną liczebnością odczytów
`fill` - zmienna, wobec której wykonane jest kolorowanie

`x` - zmienna, wobec której wykonane jest łączenie próbek

`geom_bar` - dodaje warstwę z ładnym wykresem

`theme_bw` - zapewnienie szarego motywu wykresu

`facet_wrap` - grupowanie próbek na podstawie zmiennej

`scale_color_manual` - kolor punktów i linii

`scale_fill_manual` - kolor wypełnień

`scale_x_continuous` - zmiana wyglądu osi *OX*

`ggtitle` - dodanie tytułu wykresu

`theme` - wygląd elementów wykresu

Mapy ciepła wykonano komendą:

```
edited = phyloseq_rare
tax_table(edited)[,"Species"] <- gsub("D_10__", "", tax_table(edited)[,"Species"])
plot_heatmap(edited, taxa.label="Species", low ="blue", high="red",
na.value="white",taxa.order="Species")
+ facet_wrap("zbiornik", scales="free")
+ ggtitle("Mapa ciepła składu taksonomicznego")
+ theme(plot.title = element_text(size = 20, face = "bold"),
text = element_text(size=20),
axis.text.x = element_text(angle=45, hjust=0.5))
```

gdzie:

`phyloseq_rare` - zmienna, w której znajdują się próbki, z wyrównaną liczebnością odczytów

`tax_table(edited)[,"Species"]` - nazwy gatunków zamienione, aby nie zawierały fragmentu "D_10__"

`taxa.label` - podpisy taksonów

`low, high, na.value` - skala kolorów

`taxa.order` - kolejność taksonów

`facet_wrap` - grupowanie próbek na podstawie zmiennej

`ggtitle` - dodanie tytułu wykresu

`theme` - wygląd elementów wykresu

Drzewo filogenetyczne z zaznaczonymi na gałęziach informacjami o próbkach wykonano komendą:

```
plot_tree(phyloseq_rare, ladderize="left", color="samples", shape = "zbiornik",
size="abundance", label.tips="Genus", text.size=2, base.spacing=0.04)
+ ggtitle("Drzewo z kropkami składu taksonomicznego")
+ theme(plot.title = element_text(size = 20, face = "bold"), text = element_text(size=20), axis.text.x =
```

gdzie:

`phyloseq_rare` - zmienna, w której znajdują się próbki, z wyrównaną liczebnością odczytów
`ladderize` - sposób "drabinkowania" drzewa
`color` - zmienna, wobec której kolorowane są punkty
`shape` - zmienna, wobec której wybierany jest kształt punktów
`size` - rozmiar na podstawie liczebności
`label.tips` - etykiety gałęzi
`text.size` - rozmiar czcionki
`base.spacing` - odległość między punktami
`ggtitle` - dodanie tytułu wykresu
`theme` - wygląd elementów wykresu

4. Wyniki

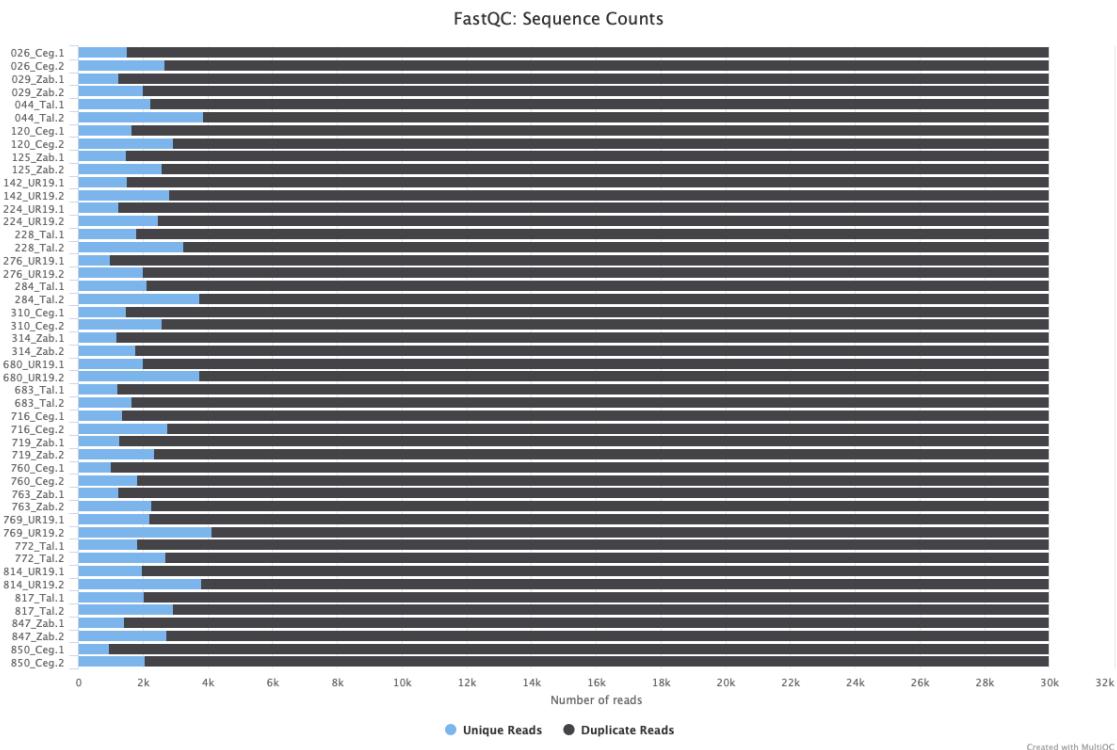
4.1. Analiza i opis jakości danych

Otrzymane dane przeanalizowano i przeprowadzono badanie ich jakości. W Tabeli 4.1 przedstawione są ogólne statystyki odczytów. Procent duplikatów w każdym z plików przyjmował wartości na poziomie 86,2%-96,8%. We wszystkich próbkach zawartość %GC wahała się w zakresie 48%-58%. Próbki różniły się między sobą zakresem długości odczytów, przy czym z danej próbki odczyty *reverse* były zawsze krótsze. Odczyty *forward* miały długość 267-283 pz, a *reverse* 221-258 pz.

W każdej próbce (zarówno odczytach *forward*, jak i *reverse*) znajdowało się 30000 odczytów (Rys. 4.1), które były sekwencjonowane metodą Sanger/Illumina 1.9. Każda z próbek różniła się liczbą unikalnych i powtarzonych odczytów.

Tabela 4.1: Opis danych. Długość odczytów, procent duplikatów i zawartość par GC w próbkach.

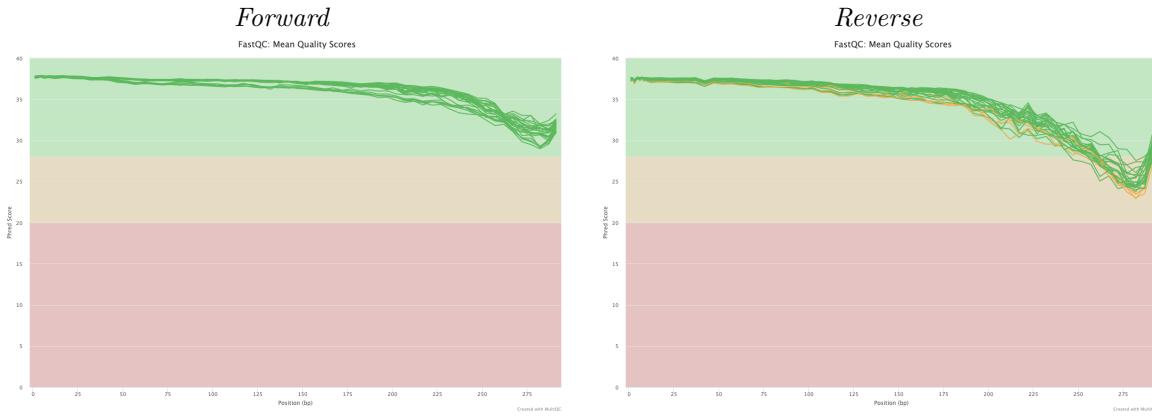
Zbiornik	Rok	Pora roku	Próbka	Długość odczytów (pz)		% Duplikatów		% GC	
				Forward	Reverse	Forward	Reverse	Forward	Reverse
Ceglów	2017	wiosna	026_Ceg	273 pz	237 pz	95%	91,1%	57%	57%
		lato	120_Ceg	277 pz	226 pz	94,5%	90,2%	56%	56%
		jesień	310_Ceg	277 pz	221 pz	95,0%	91,3%	56%	57%
	2019	wiosna	716_Ceg	275 pz	233 pz	95,5%	90,8%	58%	59%
		lato	760_Ceg	280 pz	256 pz	96,6%	93,9%	48%	49%
		jesień	850_Ceg	279 pz	244 pz	96,8%	93,1%	56%	57%
Zabłotnia	2017	wiosna	029_Zab	283 pz	253 pz	95,8%	93,3%	48%	49%
		lato	125_Zab	278 pz	227 pz	95,1%	91,3%	57%	57%
		jesień	314_Zab	282 pz	243 pz	96,0%	94,1%	53%	54%
	2019	wiosna	719_Zab	280 pz	242 pz	95,8%	92,2%	53%	53%
		lato	763_Zab	281 pz	232 pz	95,8%	92,5%	56%	57%
		jesień	847_Zab	282 pz	237 pz	95,2%	90,9%	54%	55%
Tały	2017	wiosna	044_Tal	272 pz	246 pz	92,5%	87,1%	56%	57%
		lato	228_Tal	267 pz	238 pz	94,0%	89,1%	56%	57%
		jesień	284_Tal	273 pz	247 pz	93,0%	87,5%	54%	55%
	2019	wiosna	683_Tal	279 pz	258 pz	95,9%	94,5%	51%	54%
		lato	772_Tal	270 pz	245 pz	93,9%	91,0%	57%	57%
		jesień	817_Tal	276 pz	250 pz	93,2%	90,2%	55%	56%
Urwitałt	2017	wiosna	276_UR19	280 pz	249 pz	96,7%	93,3%	54%	55%
		lato	142_UR19	279 pz	246 pz	94,9%	90,6%	54%	55%
		jesień	224_UR19	279 pz	248 pz	95,8%	91,8%	54%	55%
	2019	wiosna	680_UR19	278 pz	244 pz	93,3%	87,5%	55%	55%
		lato	769_UR19	276 pz	242 pz	92,6%	86,2%	56%	56%
		jesień	814_UR19	279 pz	244 pz	93,4%	87,3%	55%	56%



Rysunek 4.1: Zliczenia sekwencji w próbkach. Na osi OX widoczna jest liczba odczytów, natomiast na osi OY przedstawiona jest nazwa próbki. Odczyty *forward* i *reverse* oznaczone są jako odpowiednio 1 i 2. Kolorem niebieskim i czarnym zaznaczono frakcję odczytów odpowiednio unikalnych i zduplikowanych.

Następnie oceniono jakość odczytów. Na Rysunku 4.2 przedstawiono dla każdej z próbek wykres średniej wartości jakości dla każdej pozycji nukleotydowej. Ze względu na kumulację błędów wraz z wydłużaniem się odczytów, ich jakość maleje na końcach. Ze wszystkich 48 próbek tylko 4 wykazywały średnią jakość (żółty kolor linii) i były to wyłącznie odczyty *reverse*. Pozostałe 44 próbki opisane były jako próbki o dobrej jakości (zielony kolor linii).

Dla odczytów *forward* wybrano długości: 260, ponieważ od tej długości jakość odczytów zaczyna powoli maleć i 232, ponieważ dla tej wartości wszystkie odczyty mają jeszcze dość wysoki poziom jakości. Dla odczytów *reverse* wybrano długości: 242, ponieważ była to ostatnia wartość długości, gdzie wszystkie odczyty osiągały dobrą jakość i dodatkowo była zapewniona najmniejsza strata danych, oraz 212, ponieważ od tej długości jakość odczytów zaczynała maleć.



Rysunek 4.2: Wykres średniej wartości jakości dla każdej analizowanej próbki. Lewy panel przedstawia wykres jakości dla odczytów *forward*, a prawy dla odczytów *reverse*. Na osi *OX* widoczna jest pozycja nukleotydów (liczona w parach zasad), natomiast na osi *OY* przedstawiona jest wartość Phred opisująca jakość. Kolor zielony, żółty i czerwony odpowiada odpowiednio dobrej, średniej i złej jakości.

4.2. Przygotowanie danych do analizy

Odczyty odfiltrowano i przycięto na podstawie wykresów analizy jakości. W Tabeli 4.2 widoczne są wyniki przycinania odczytów: informacja o minimalnym i maksymalnym procencie odczytów przefiltrowanych oraz minimalnym i maksymalnym procencie odczytów bez chimer w każdej próbce w zależności od przyjętych parametrów. Sprawdzono wszystkie kombinacje długości odczytów *forward* i *reverse*, a także wartości parametru q .

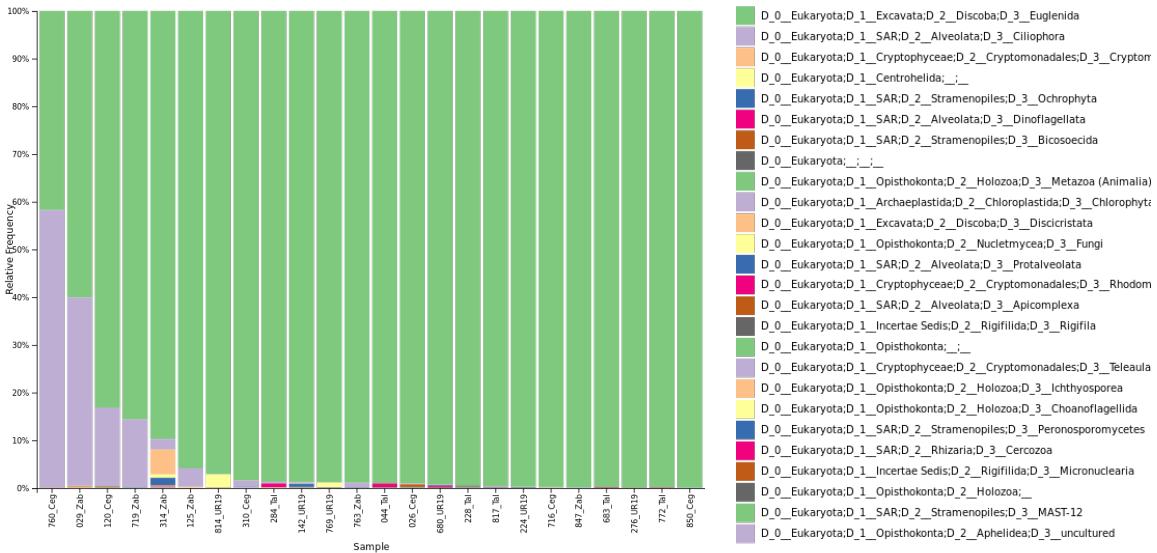
Tabela 4.2: Parametry użyte w przygotowaniu danych do analiz. Kolorem żółtym zaznaczono ostatecznie wybrane parametry, których użyto do dalszych kroków analiz.

Długość forward	Długość reverse	Parametr q	Min % przefiltrowanych	Max % przefiltrowanych	Min % bez chimer	Max % bez chimer
260	242	3	34,89%	72,25%	27,31%	66,57%
		7	26,13%	69,7%	20,9%	62,35%
260	212	3	61,42%	83,72%	32,83%	79,19%
		7	56,61%	82,44%	32,65%	78,1%
232	212	3	62,62%	86,58%	33,99%	82,14%
		7	58,44%	86,58%	34,01%	82,14%
232	242	3	35,14%	74,42%	26,45%	67,93%
		7	26,44%	71,98%	21,18%	64,33%

Przyglądając się tabeli można zauważyc, że lepsze wyniki otrzymuje się przy mniejszych wartościach parametru q i przy krótszych długościach odczytów. Sprawdzono dodatkowo, czy użycie odczytów o jeszcze mniejszej długości niż pierwotnie zakładano, powoduje powstanie lepszych wyników. Wykonano przycinanie i filtrowanie z parametrami: długość *forward* = 232, długość *reverse* = 197, q = 3. Okazało się, że minimalny procent przefiltrowanych odczytów osiągał wyższe wartości, ale jednocześnie malał procent odczytów bez chimer. Postanowiono zatem dalsze analizy przeprowadzić na odczytach przyciętych do długości 232 i 212 (odpowiednio: *forward* i *reverse*) i parametrem q = 3, które w Tabeli 4.2 zaznaczone są kolorem żółtym.

4.3. Przyporządkowanie taksonomiczne

Po wykonaniu klasyfikacji plik wynikowy przekonwertowano do wyświetlenia. Wszystkie przypisania taksonomiczne były wykonane z pewnością na poziomie ~70%-100%. Następnie, z wykorzystaniem metadanych, wykonano wykres słupkowy przedstawiający skład taksonomiczny w próbce. Rysunek 4.3 przedstawia wykres słupkowy, w którym organizmy przedstawione są do poziomu gromad (ang. *Phylum*). Zdecydowana większość odczytów pochodziła od Euglenida. Inne gromady, które pojawiały się w wynikach to np. Ciliophora, Ochrophyta, Dinoflagellata. Najbardziej zróżnicowaną taksonomicznie próbka była próbka 314_Zab, czyli próbka pobrana jesienią 2017 z Zabłotni, a najmniej zróżnicowaną 850_Ceg, czyli próbka pobrana jesienią 2019 z Cegłów. Najliczniej występującą grupą była Euglenida, którą zidentyfikowano 21827 razy w próbce 142_UR19, czyli próbce pobranej w lato 2017 z Urwitału.



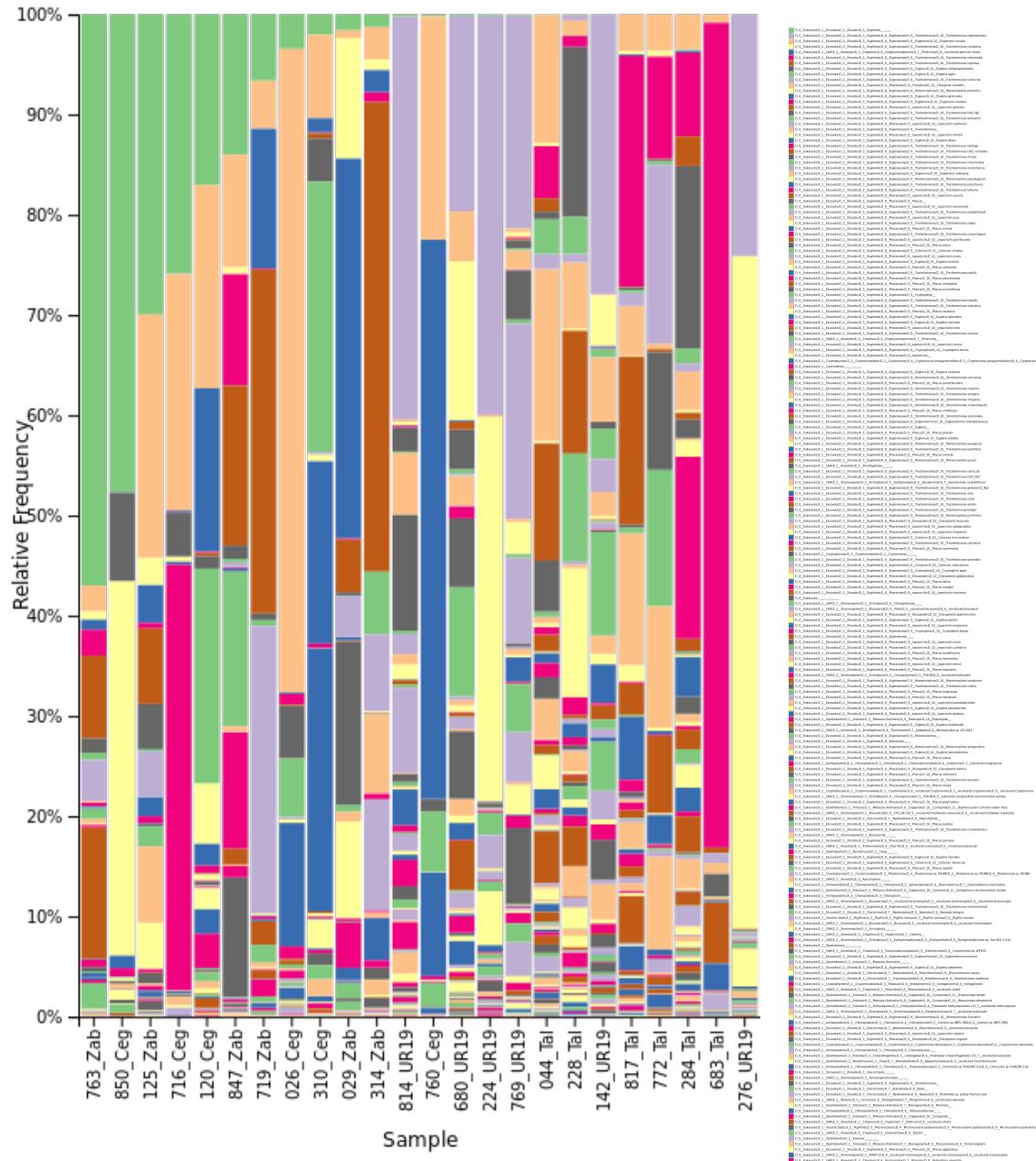
Rysunek 4.3: Wykres słupkowy przedstawiający skład taksonomiczny w próbkach do poziomu gromady. Na osi OX przedstawiona jest nazwa próbki, a na osi OY relatywny procent występowania danego taksonu. Kolory słupków odpowiadają nazwom taksonów.

Sprawdzono następnie, jaki procent odczytów z badanych próbek należał do przedstawicieli Euglenida. W Tabeli 4.3 przedstawiono procent odczytów spełniający te kryteria dla każdej z analizowanej próbki osobno. Najmniejszy procent odczytów (41.719%) należących do przedstawicieli Euglenida zidentyfikowano dla próbki 760_Ceg, czyli próbki pobranej latem 2019 z Ceglowa, a największy (99.947%) dla 850_Ceg, czyli próbki pobranej jesienią 2019 z Ceglowa.

Tabela 4.3: Procent odczytów należących do Euglenida w poszczególnych próbkach

Próbka	Procent odczytów
760_Ceg	41.719%
029_Zab	60.057%
120_Ceg	83.143%
719_Zab	85.593%
314_Zab	89.697%
125_Zab	95.820%
814_UR19	97.076%
310_Ceg	98.391%
284_Tal	98.670%
142_UR19	98.751%
769_UR19	98.786%
763_Zab	98.888%
044_Tal	98.999%
026_Ceg	99.044%
680_UR19	99.242%
228_Tal	99.502%
817_Tal	99.705%
224_UR19	99.732%
716_Ceg	99.782%
847_Zab	99.844%
683_Tal	99.876%
276_UR19	99.901%
772_Tal	99.912%
850_Ceg	99.947%

Rysunek 4.4 przedstawia wykres słupkowy, w którym organizmy przedstawione są do poziomu gatunków (ang. *Species*). Najczęściej występującymi gatunkami były *Trachelomonas intermedia*, *Trachelomonas compacta*, *Trachelomonas rugulosa*. *Trachelomonas intermedia* został zidentyfikowany 15260 razy w próbce 683_Tal (Tały, wiosna 2019), *Trachelomonas compacta* 14304 razy w próbce 276_UR19 (Urwitał, wiosna 2017), a *Trachelomonas rugulosa* 10191 razy w próbce 314_Zab (Zabłotnia, jesień 2017).



Rysunek 4.4: Wykres słupkowy przedstawiający skład taksonomiczny w próbkach do poziomu gatunku. Na osi *OX* przedstawiona jest nazwa próbki, a na osi *OY* relatywny procent występowania danego taksonu. Kolory słupków odpowiadają nazwom taksonów.

4.4. Analiza danych w R

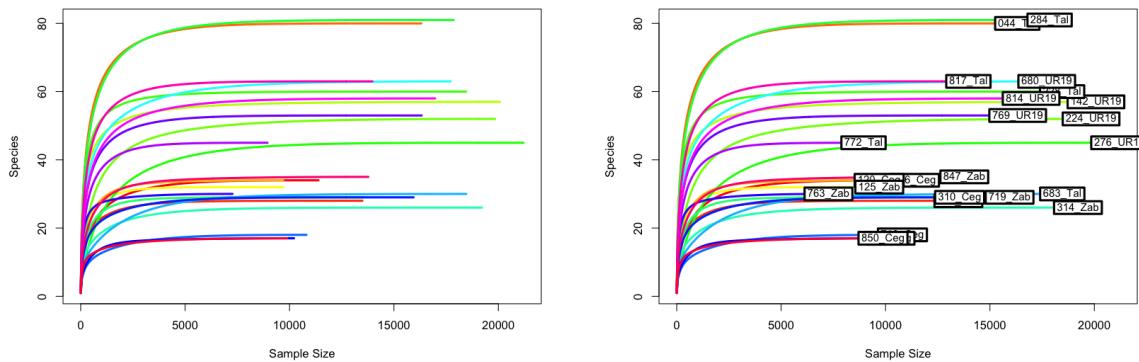
W analizowanych danych znajdowały się 24 próbki, w których zidentyfikowano 1731 taksonów. Najmniej liczna próbka zawierała 10197 odczytów, a najbardziej liczna 24641 odczytów. Poziomy taksonomiczne do-

stępne w danych to: Królestwo, Gromada, Klasa, Rząd, Rodzina, Rodzaj, Gatunek (ang. *Kingdom, Phylum, Class, Order, Family, Genus, Species*).

W wyniku filtrowania odczytów do tych, których przedstawicielem jest Euglenida, spowodowało ograniczenie liczebności danych. Zawierały one mniej taksonów i miały mniejszą minimalną i maksymalną liczbę odczytów w próbce. Poziomy taksonomiczne pozostały bez zmian. W przefiltrowanych danych znajdowały się 24 próbki, w których zidentyfikowano 1609 taksonów. Najmniej liczna próbka zawierała 10188 odczytów, a najbardziej liczna 21827 odczytów. Poziomy taksonomiczne dostępne w danych to: Królestwo, Gromada, Klasa, Rząd, Rodzina, Rodzaj, Gatunek (ang. *Kingdom, Phylum, Class, Order, Family, Genus, Species*). Do dalszych analiz używano odczytów, które należały do Euglenida.

Analizując przyporządkowanie taksonomiczne na poziomie gatunku zauważono, że niewiele przyporządkowań jest nieznanych (mało wartości NaN). Dodatkowo, zidentyfikowano tylko 131 unikatowych gatunków. Zdecydowano się zatem na połączenie ASV z tych samych gatunków. Połączenie ASV z tych samych gatunków, zgodnie z oczekiwaniemi, zmniejszyło liczbę taksonów, ale w większości były to taksony, które się powtarzały. Nieznacznie zmieniły się również minimalna i maksymalna liczba odczytów w próbce. Nie wydaje się jednak, żeby była to znacząca różnica w porównaniu z danymi niepołączonymi. Nie zmieniła się liczba próbek, ani poziomy taksonomiczne. W połączonych danych znajdowały się 24 próbki, w których zidentyfikowano 130 taksonów. Najmniej liczna próbka zawierała 7268 odczytów, a najbardziej liczna 21208 odczytów. Poziomy taksonomiczne dostępne w danych to: Królestwo, Gromada, Klasa, Rząd, Rodzina, Rodzaj, Gatunek (ang. *Kingdom, Phylum, Class, Order, Family, Genus, Species*). Połączenie ASV pozwoliło na usprawnienie obliczeń i umożliwiło bardziej czytelną wizualizację wyników, przy jednoczesnej niewielkiej stracie danych. Dalsze analizy przeprowadzono na ASV połączonych dla tych samych gatunków.

Wykonano wykres krzywej wysycenia dla danych z połączonymi ASV dla tych samych gatunków, który widoczny jest na Rysunku 4.5. Na podstawie krzywej można stwierdzić, że liczba odczytów w najmniej licznej próbce wydaje się wystarczająca, żeby można było do tej wartości ograniczyć liczbę odczytów we wszystkich próbkach, bez znaczącej straty informacji o bogactwie gatunkowym i różnorodności biologicznej. Najmniej liczna próbka zawierała 7268 odczytów. Było to mniej więcej 1/3 liczebności odczytów w najbardziej licznej próbce, ale w porównaniu z resztą próbek ta wartość nie odbiegała znacząco. Najprawdopodobniej pewne dane zostały utracone, ale warto było ograniczyć liczbę odczytów, ponieważ usprawniło to obliczenia. Sama strata danych była natomiast niekrytyczna, ponieważ ograniczenie liczby odczytów we wszystkich próbkach do najmniej licznej próbki nie powodowało znaczącej utraty różnych gatunków w żadnej z próbek. Wszystkie próbki zdecydowano się zatem ograniczyć do 7268 odczytów.



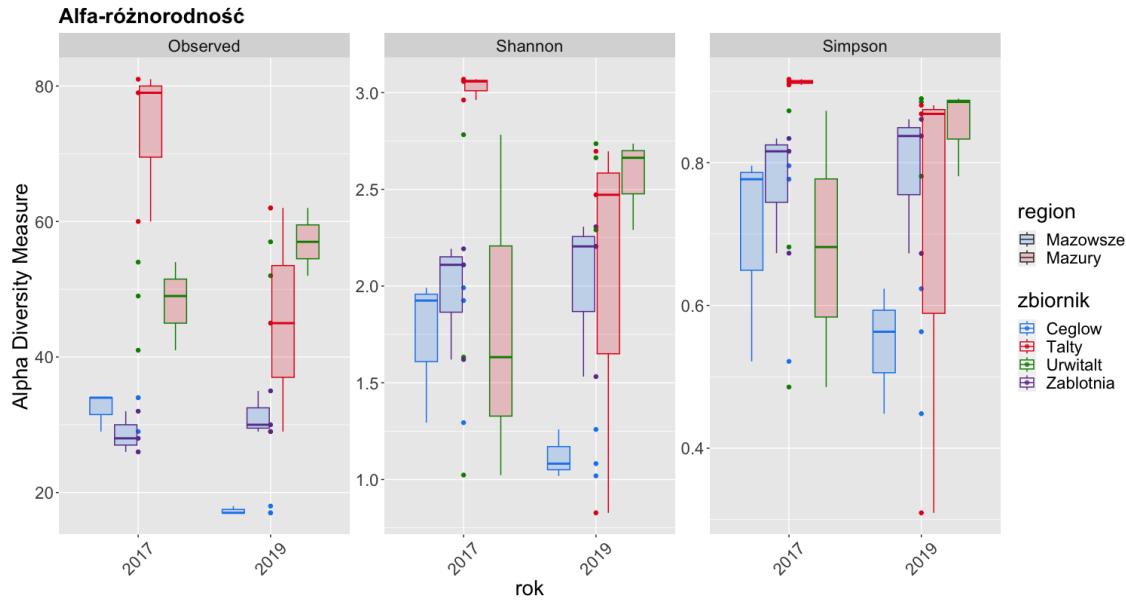
Rysunek 4.5: Wykres krzywej wysycenia dla każdej analizowanej próbki. Lewy panel przedstawia wykres krzywej wysycenia bez podpisów linii, a prawy z podpisami odpowiadającymi nazwom próbek. Na osi *OX* widoczny jest rozmiar próbki, natomiast na osi *OY* przedstawiona jest liczba różnych gatunków. Kolor linii odpowiada każdej z analizowanej próbek z osobna.

W wyniku ograniczenia danych nie utracono żadnych taksonów, ani żadnej próbki. Wydaje się, że ograniczenie odczytów w próbkach nie zaburzyło znacząco danych i ich różnorodności. W ograniczonych danych znajdowały się 24 próbki, w których zidentyfikowano 130 taksonów. Wszystkie próbki zawierały 7268 odczytów. Poziomy taksonomiczne dostępne w danych to: Królestwo, Gromada, Klasa, Rząd, Rodzina, Rodzaj, Gatunek (ang. *Kingdom, Phylum, Class, Order, Family, Genus, Species*).

Gdyby liczba odczytów w najmniej licznej próbce znacząco odbiegała od pozostałych próbek lub ograniczenie liczebności odczytów w próbkach spowodowałoby znaczne obniżenie różnorodności gatunkowej, warte rozważenia byłoby odrzucenie analizowanej próbki z dalszych kroków badań. Co prawda spowodowałoby to utratę danych, w tym pewnych unikatowych gatunków, ale byłaby to mniej znacząca utrata danych niż w przypadku ograniczenia różnorodności bardzo licznych próbek.

4.5. Analiza α -różnorodności

Na Rysunku 4.6 przedstawiono wykres α -różnorodności z wykorzystaniem wybranych wskaźników: Observed, Shannona i Simpsona.



Rysunek 4.6: Wykres α -różnorodności z wykorzystaniem wybranych wskaźników: Observed, Shannon, Simpson. Na osi OX widoczny jest rok poboru próbki, a na osi OY miara α -różnorodności. Niebieski, czerwony, zielony i fioletowy kolor punktów i konturów pudełek odpowiada zbiornikowi odpowiednio: Cegłów, Tały, Urwitałt, Zabłotnia. Niebieski i czerwony kolor wypełnienia pudełek odpowiada regionowi odpowiednio: Mazowsza i Mazur.

Dla obserwowanego bogactwa gatunkowego widać, że odpowiadające sobie dane z różnych zbiorników i regionów Polski nie różnią się znacznie między latami. W 2017 roku miara wskaźnika α -różnorodności dla zbiornika Tały przyjmowała wartości w zakresie ok. 60-80, podczas gdy w roku 2019 było to ok. 30-60. Analogicznie dla zbiorników Urwitałt, Zabłotnia i Ceglów w roku 2017 były to zakresy w okolicach odpowiednio: 40-55, 28-32 i 29-35. Natomiast w roku 2019 dla zbiorników Urwitałt, Zabłotnia i Ceglów zaobserwowano miarę wskaźnika α -różnorodności na przybliżonym poziomie odpowiednio: 52-62, 29-37 i 15-18.

Dla wskaźnika α -różnorodności Shannona można zauważać, że odpowiadające sobie dane z różnych zbiorników i regionów Polski nie różnią się znacznie między latami w większości przypadków. Jedynie zauważalne różnice można zidentyfikować dla zbiornika Tały i Urwitałt. Miara wskaźnika α -różnorodności w 2017 roku w zbiorniku Tały przyjmowała wtedy wartości w zakresie ok. 2.9-3.1, podczas gdy w roku 2019 było to ok. 0.4-2.6. Dla zbiornika Urwitałt było to odpowiednio: 1.1-2.7 i 2.3-2.7. Miara wskaźnika α -różnorodności w Zabłotni i Ceglów w roku 2017 mieściła się w zakresach odpowiednio: 1.6-2.6 i 1.2-2.0. Natomiast w roku 2019 dla tych zbiorników zaobserwowano miarę wskaźnika α -różnorodności na przybliżonym poziomie odpowiednio: 1.51-2.22 i 1.01-1.2.

Dla wskaźnika α -różnorodności Simpsona największe różnice pomiędzy latami można dostrzec dla zbiornika Tały. W 2017 roku miara α -różnorodności przyjmowała wartości na poziomie 0.92, a w 2019 roku w zakresie 0.31-0.88. Dla pozostałych zbiorników zmiana między badanymi latami jest niewielka. W roku 2017 w zbiornikach Ceglów, Urwitałt i Zabłotnia zaobserwowano miarę α -różnorodności w zakresie odpowiednio: 0.52-0.79, 0.48-0.88 i 0.68-0.83. Natomiast w roku 2019 było to odpowiednio: 0.45-0.62, 0.78-0.88 i 0.68-0.85.

Przyglądając się wykresom można zauważyc, że niezależnie od badanego wskaźnika α -różnorodności, nie występują zbyt duże różnice pomiędzy latami 2017 i 2019. Zarówno dla obserwowanego bogactwa gatunkowego, jak i wskaźnika Shannona widać, że dane dla różnych regionów (Mazowsze i Mazury) oraz dla różnych zbiorników (Ceglów, Tały, Urwitał i Zabłotnia) różnią się od siebie. Dla żadnej ze zmiennych nie widać dużych różnic przy analizowaniu wskaźnika Simpsona.

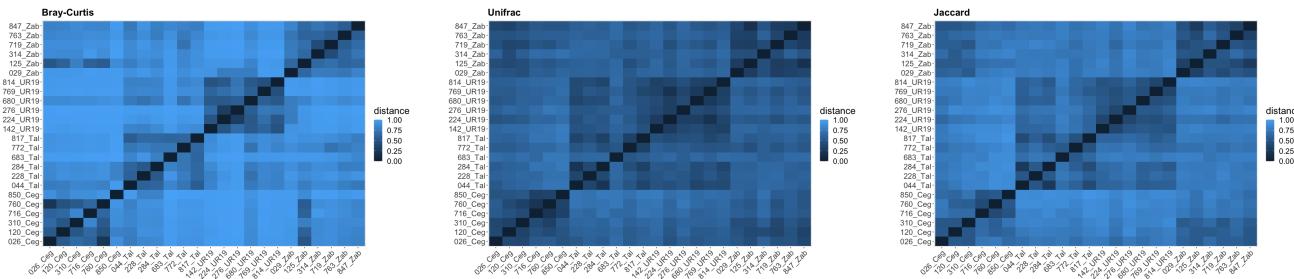
Sprawdzono następnie, czy obserwowane na wykresach różnice są istotne statystycznie. W Tabeli 4.4 przedstawiono wyniki testu ANOVA. Dla żadnego z badanych wskaźników α -różnorodności nie zidentyfikowano istotnych statystycznie różnic dla dwóch różnych lat pobierania próbek. Istotnie statystycznie różnice występują w zależności od zbiornika i regionu dla wskaźnika α -różnorodności Shannona i obserwowanego bogactwa gatunkowego. P-wartość dla obserwowanego bogactwa gatunkowego przyjmuje wartość 6.79e-05 przy analizowaniu danych z różnych zbiorników i 3.77e-06 przy badaniu różnych regionów. Oznacza to, że bardziej istotna różnica występuje między próbami w zależności od regionu niż zbiornika. P-wartości dla wskaźnika Shannona dla odpowiadających zmiennych przyjmują niższe wartości, odpowiednio: 0.0438 i 0.0225. Oznacza to, że wskaźnik Shannona znajduje mniej istotne statystycznie różnice dla tych zmiennych niż przy użyciu bogactwa gatunkowego jako wskaźnika α -różnorodności. Dla wskaźnika α -różnorodności Simpsona żadna zmienna nie przyjmuje różnic istotnych statystycznie.

Tabela 4.4: Wyniki badania istotności statystycznej różnic pomiędzy próbami z różnych lat, zbiorników, regionów Polski w zależności od używanego wskaźnika α -różnorodności. Gwiazdką i kolorem żółtym oznaczono wyniki istotne statystycznie (p -wartość ≤ 0.05)

Zmienna	Wskaźnik α -różnorodności					
	Observed		Shannon		Simpson	
	F-wartość	p-wartość	F-wartość	p-wartość	F-wartość	p-wartość
Rok	1.098	0.306	0.543	0.469	0.463	0.503
Zbiornik	12.81	6.79e-05 *	3.241	0.0438 *	1.453	0.257
Region	37.33	3.77e-06 *	6.025	0.0225 *	1.364	0.255

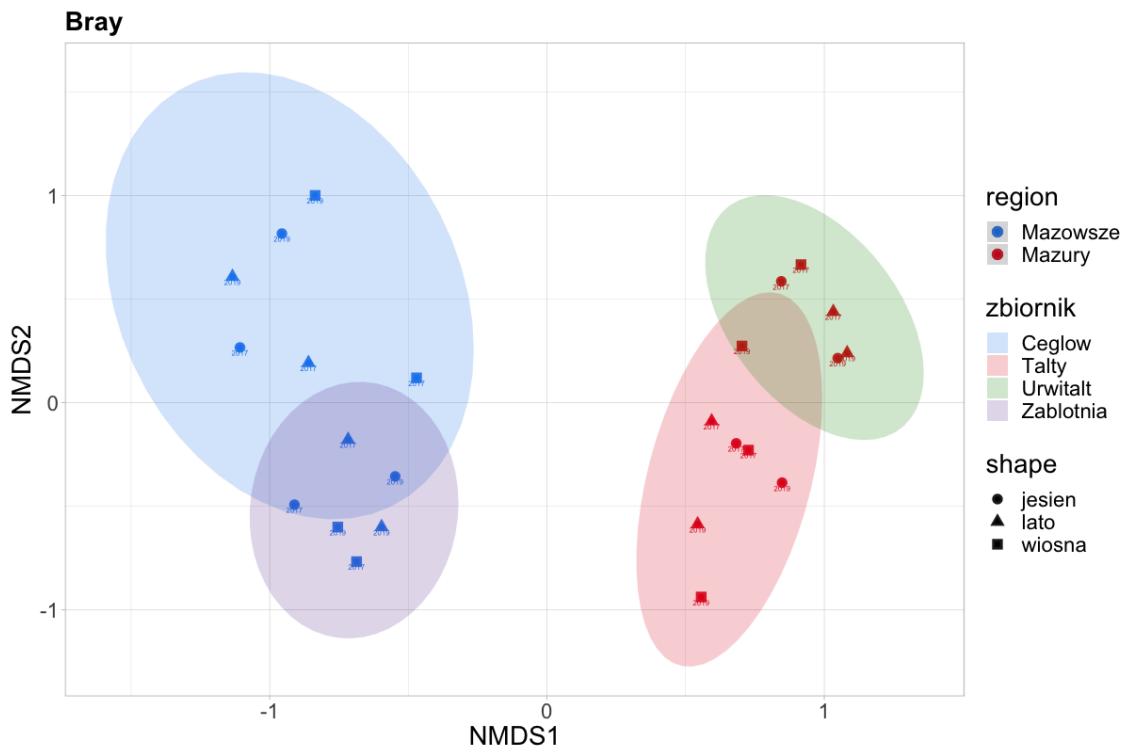
4.6. Analiza β -różnorodności

Rysunek 4.7 przedstawia mapy ciepła, które są wizualizacją macierzy odległości mierzonych wybranymi metodami: Bray-Curtisa, Unifraaq i Jaccarda. Każda z metod pomimo tego samego celu (obliczenie podobieństwa między próbami), powoduje otrzymanie różnych wyników. Jest to spowodowane różnymi założeniami w tych metodach. Odległość Jaccarda informuje, jak podobne są do siebie 2 próbki na podstawie braku lub obecności taksonów. Odległość Bray-Curtisa sprawdza liczbę gatunków. Metoda Unifrac bierze dodatkowo pod uwagę informacje filogenetyczne. Na podstawie map ciepła widać, że przy użyciu metody Unifrac odległość między dwoma próbami jest o wiele mniejsza niż odległość między tymi samymi dwoma próbami mierzona metodą Bray-Curtisa, ale tylko nieznacznie mniejsza niż metodą Jaccarda. Macierze odległości mierzone metodą Unifraq i Jaccarda są do siebie bardziej podobne.



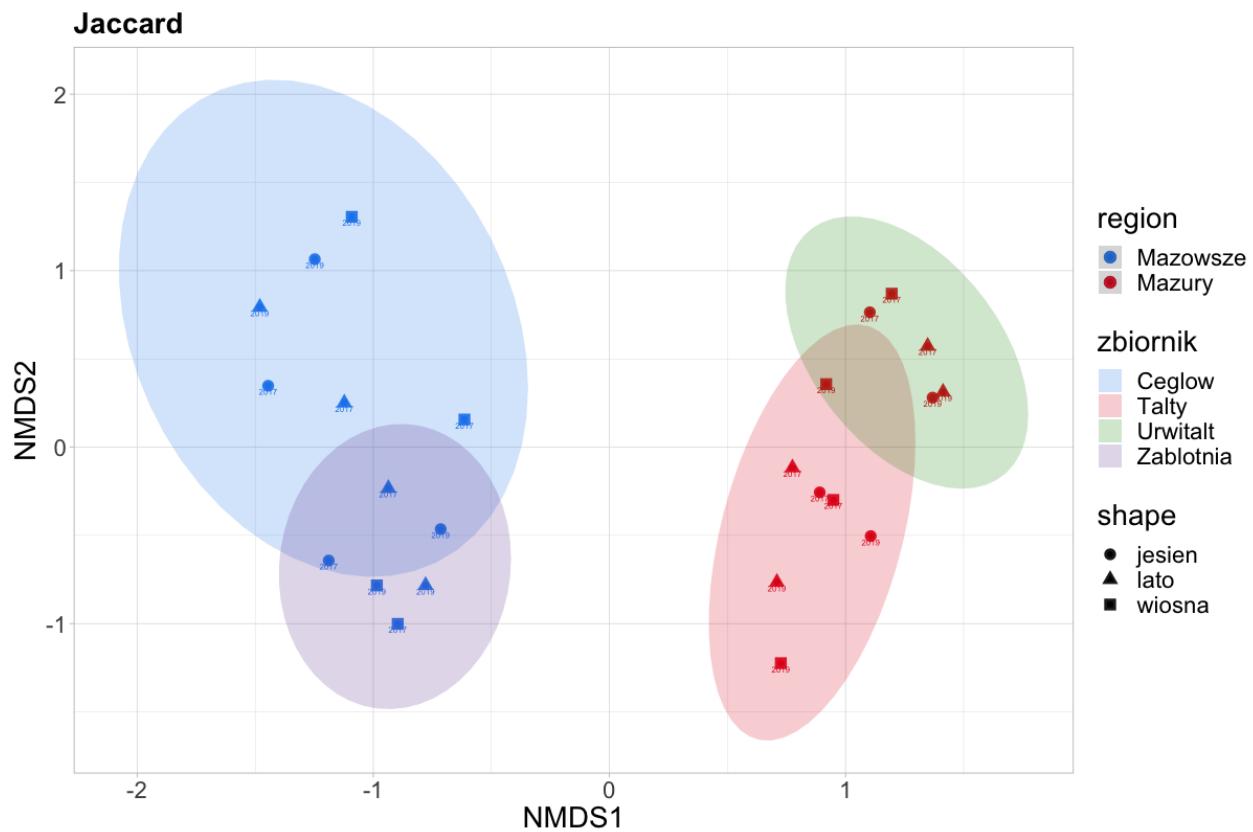
Rysunek 4.7: Mapy ciepła przedstawiające macierze odległości obliczone wybranymi metodami: Bray-Curtis, Unifrac, Jaccard. Na osiach OX i OY przedstawione są nazwy próbek. Skala kolorów odpowiada odległości między próbami: im ciemniejszy kolor tym mniejsza odległość między próbami (bardziej podobne próbki).

Na podstawie macierzy odległości wykonano ordynację metodą NMDS (ang. *Nonmetric MultiDimensional Scaling*).



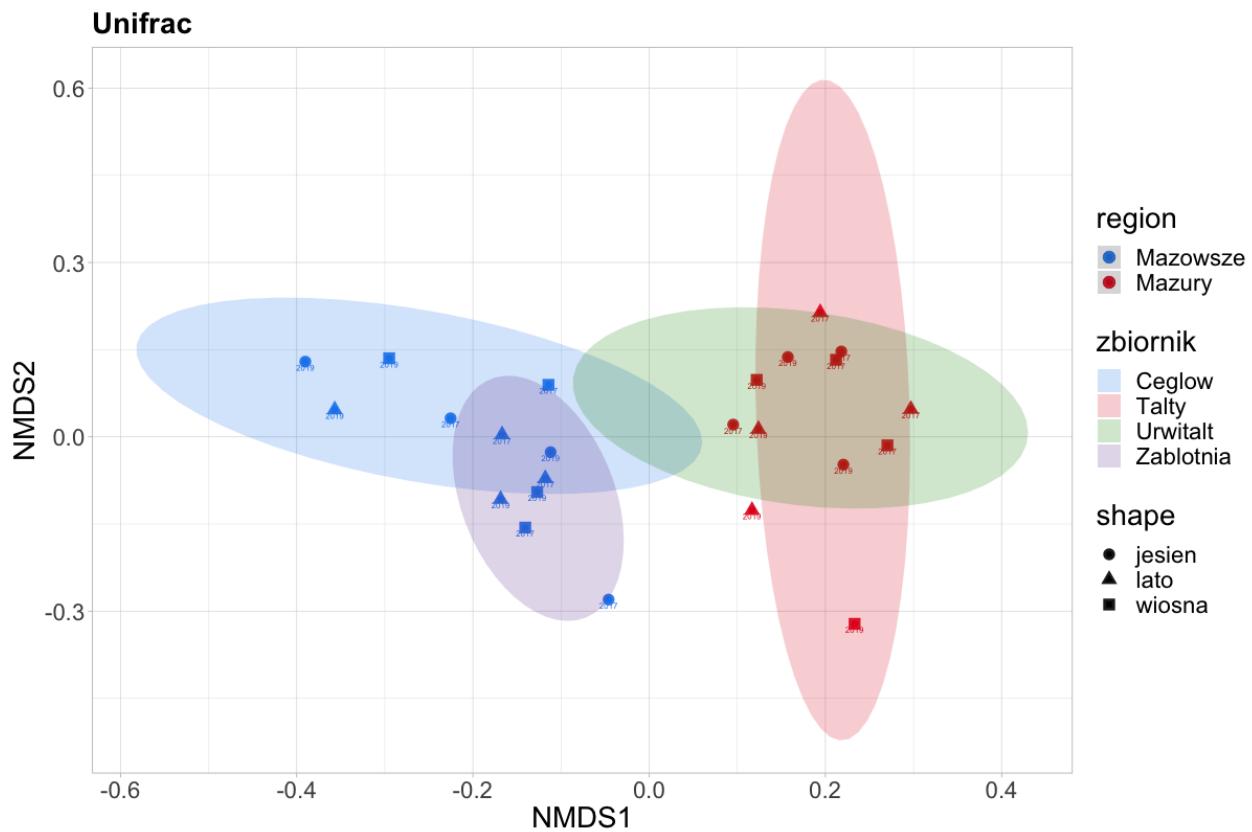
Rysunek 4.8: Wykres β -różnorodności dla odległości Bray-Curtisa. Niebieskie i czerwone punkty odpowiadają regionowi odpowiednio Mazowsza i Mazur. Niebieskie, czerwone, zielone i fioletowe elipsy reprezentują zbiorniki odpowiednio: Ceglów, Tały, Urwitałt, Zabłotnia. Punkty o kształcie koła, trójkąta i kwadratu reprezentują pory roku odpowiednio: jesień, lato i wiosnę. Punkty na wykresie podpisane są w zależności od roku pobrania próbki.

Na Rysunku 4.8 przedstawiony jest wykres β -różnorodności dla odległości Bray-Curtisa. Można zauważyć, że próbki są dobrze pogrupowane ze względu na zbiornik i region, z którego pochodzą. Gorzej jednak klastrują się próbki pochodzące z różnych zbiorników, ponieważ elipsy nachodzą na siebie. Próbki pochodzące z tych samych zbiorników i regionów są położone bardzo blisko siebie, co świadczy o ich podobieństwie. Znacznie gorzej pogrupowane są próbki z różnych lat. Świadczy to o braku jednoznacznego podobieństwa próbek ze względu na rok ich pobrania.



Rysunek 4.9: Wykres β -różnorodności dla odległości Jaccarda. Niebieskie i czerwone punkty odpowiadają regionowi odpowiednio Mazowsza i Mazur. Niebieskie, czerwone, zielone i fioletowe elipsy reprezentują zbiorniki odpowiednio: Ceglów, Tały, Urwitałt, Zabłotnia. Punkty o kształcie koła, trójkąta i kwadratu reprezentują pory roku odpowiednio: jesień, lato i wiosnę. Punkty na wykresie podpisane są w zależności od roku pobrania próbki.

Rysunek 4.9 przedstawia jest wykres β -różnorodności dla odległości Jaccarda. Tak jak w wykresie na Rys. 4.8, można zauważyć, że próbki są dobrze pogrupowane ze względu na zbiornik i region, z którego pochodzą. Gorzej jednak klastrują się próbki pochodzące z różnych zbiorników, ponieważ elipsy nachodzą na siebie. Próbki pochodzące z tych samych zbiorników i regionów są położone bardzo blisko siebie, co świadczy o ich podobieństwie. Tak jak na wykresie na Rys. 4.8, tutaj również nie zidentyfikowano jednoznacznego podobieństwa próbek ze względu na rok ich pobrania.



Rysunek 4.10: Wykres β -różnorodności dla odległości Unifraq. Niebieskie i czerwone punkty odpowiadają regionowi odpowiednio Mazowsza i Mazur. Niebieskie, czerwone, zielone i fioletowe elipsy reprezentują zbiorniki odpowiednio: Ceglów, Tały, Urwitałt, Zabłotnia. Punkty o kształcie koła, trójkąta i kwadratu reprezentują pory roku odpowiednio: jesień, lato i wiosnę. Punkty na wykresie podpisane są w zależności od roku pobrania próbki.

Na Rysunku 4.10 przedstawiony jest wykres β -różnorodności dla odległości Unifraq. Ten wykres różni się znacząco od dwóch pozostałych wykresów (Rys. 4.8, Rys. 4.9). Próbki są gorzej pogrupowane ze względu na zbiornik, z którego pochodzą - elipsy w o wiele większym stopniu na siebie nachodzą. Nadal jednak można zauważać dobre klastrowanie ze względu na region próbek - próbki z Mazowsza znajdują się od siebie w bliskiej odległości i są oddalone od próbek pochodzących z Mazur. Świadczy to o podobieństwie próbek ze względu na region ich pobrania. Podobnie jak na Rys. 4.8 i Rys. 4.9, tutaj także nie zidentyfikowano grupowania ze względu na rok pobrania próbki, co sugeruje brak jednoznacznego podobieństwa ze względu na tę zmienną.

Sprawdzono następnie, czy obserwowane na wykresach różnice są istotne statystycznie. W Tabeli 4.5 przedstawiono wyniki testów ANOSIM i PERMANOVA dla trzech wybranych metod liczenia macierzy odległości. Dla żadnej z badanych metod nie zidentyfikowano istotnych statystycznie różnic dla dwóch różnych lat pobierania próbek. Istotnie statystycznie różnice występują w zależności od zbiornika i regionu dla wszystkich badanych metod i użytych testów. P-wartość jest na poziomie 0.001, niezależnie od użytego testu i metody. Oznacza to, że wszystkie różnice są tak samo istotne statystycznie.

Dla testu ANOSIM im większa wartość R, tym bardziej próbki się od siebie różnią. Na podstawie Tabeli 4.5 widać, że najbardziej różnią się od siebie próbki pochodzące z różnych zbiorników, gdzie R przyjmuje wartości 0.8747, 0.7946, 0.8703 dla metod odpowiednio: Bray-Curtisa, Unifraq i Jaccarda. Najmniejsze różnice obserwowane są dla badanych lat, gdzie R przyjmuje wartości -0.01589, 0.01305, 0.01989 dla metod odpowiednio: Bray-Curtisa, Unifraq i Jaccarda. Świadczy to o dużym podobieństwie między próbками dla tej zmiennej.

Dla testu PERMANOVA im większa wartość R², tym większy procent zmienności tłumaczony jest zmienną grupującą. Największe wartości R² również przyjmowane są dla próbek pochodzących z różnych zbiorników i wynoszą 0.48952, 0.52178, 0.51327 dla metod odpowiednio: Bray-Curtisa, Unifraq i Jaccarda. Najmniejszy procent zmienności tłumaczony jest przez rok poboru próbki, gdzie R² przyjmuje wartości

0.03931, 0.05226, 0.0512 dla metod odpowiednio: Bray-Curtisa, Unifraq i Jaccarda.

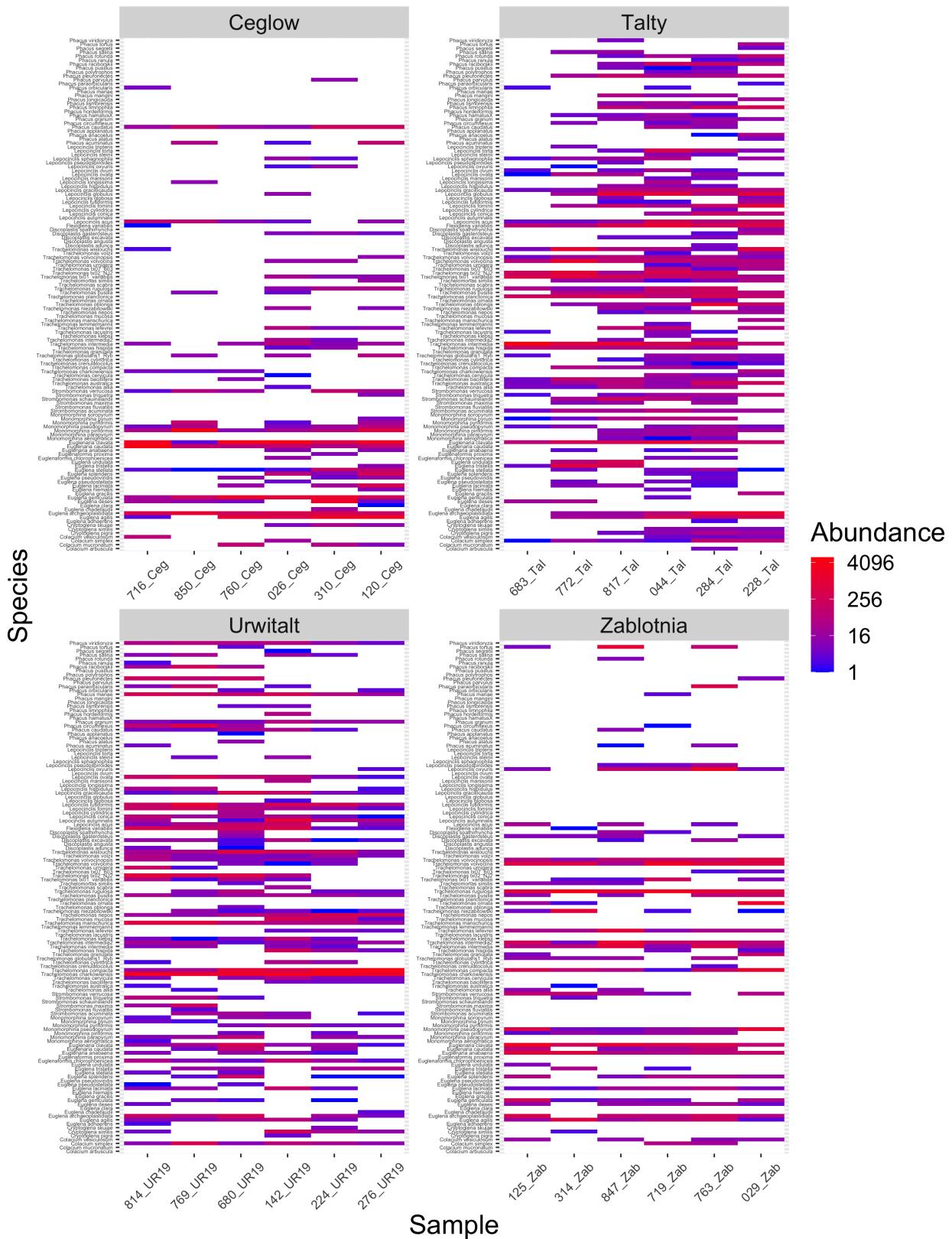
Tabela 4.5: Wyniki badania istotności statystycznej różnic pomiędzy próbками z różnych lat, zbiorników, regionów Polski w zależności od używanej metody obliczania macierzy odległości i rodzaju testu statystycznego. Gwiazdką i kolorem żółtym oznaczono wyniki istotne statystycznie (p -wartość ≤ 0.05)

Zmienna	Rodzaj testu									
	ANOSIM				PERMANOVA					
	Bray-Curtis		Unifraq		Jaccard		Bray-Curtis		Unifraq	
	R	p-wartość	R	p-wartość	R	p-wartość	R2	p-wartość	R2	p-wartość
Rok	-0.01589	0.538	0.01305	0.335	0.01989	0.291	0.03931	0.533	0.05226	0.249
Zbiornik	0.8747	0.001 *	0.7946	0.001 *	0.8703	0.001 *	0.48952	0.001 *	0.52178	0.001 *
Region	0.5792	0.001 *	0.791	0.001 *	0.8556	0.001 *	0.20003	0.001 *	0.3144	0.001 *

4.7. Analiza składu taksonomicznego

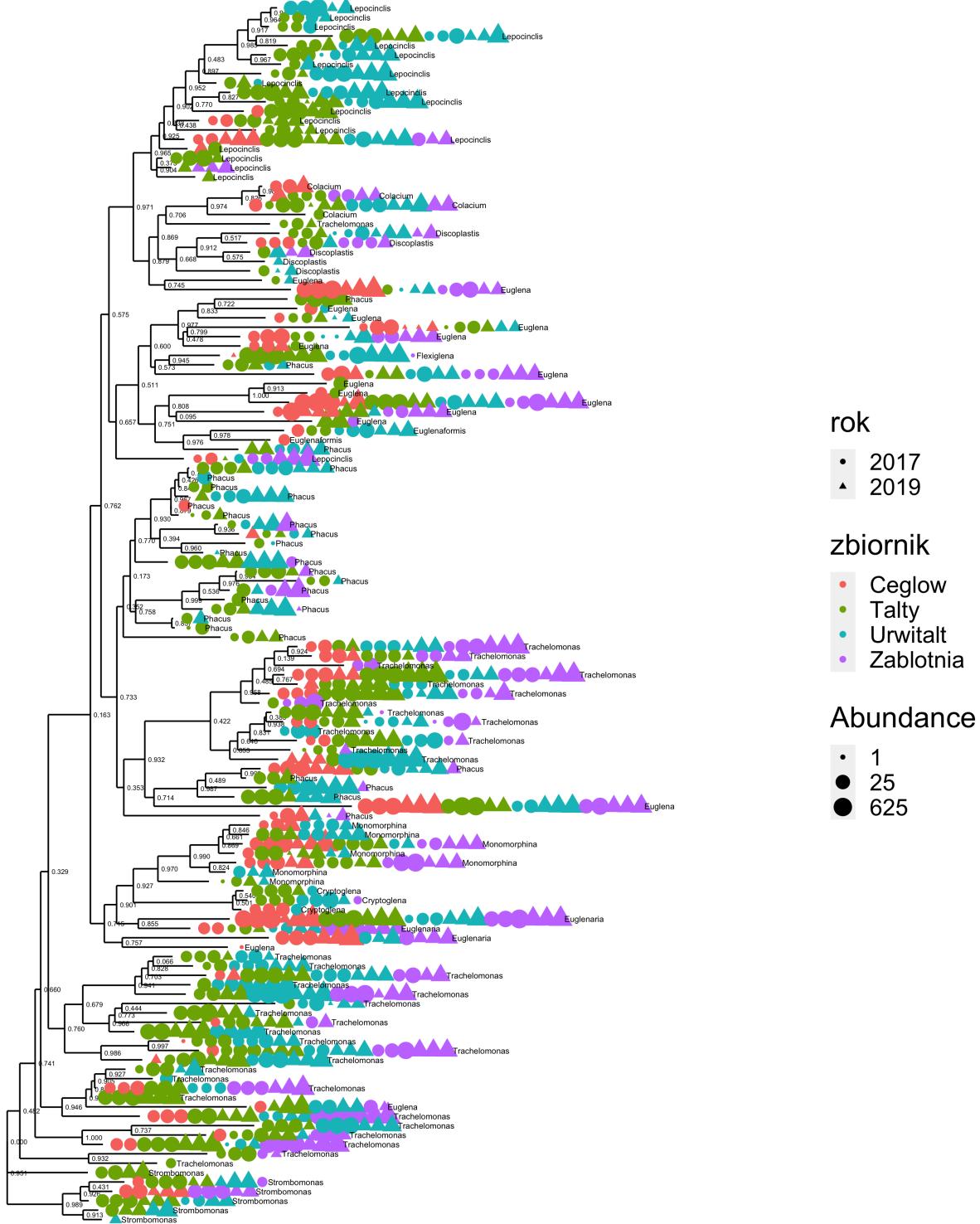
Analizę składu taksonomicznego przedstawiono w sposób graficzny za pomocą 3 rodzajów wykresów: mapy ciepła, drzewa filogenetycznego z zaznaczonymi na gałęziach informacjami o próbkach oraz wykresu słupkowego.

Mapa ciepła składu taksonomicznego

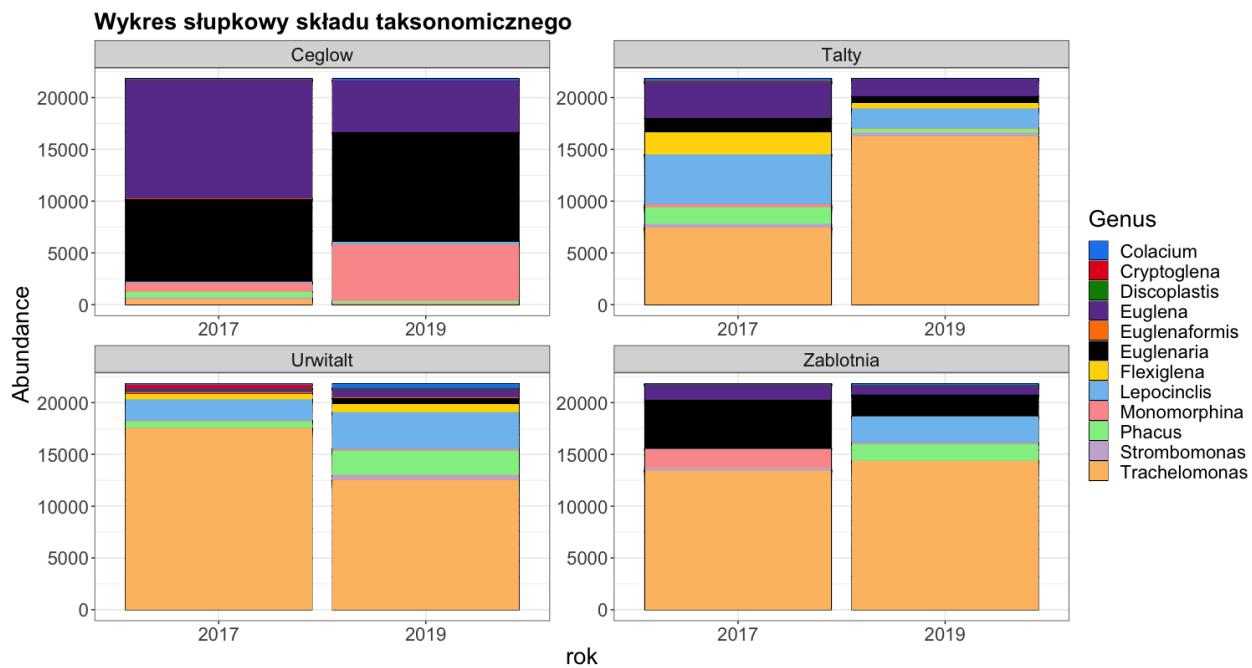


Rysunek 4.11: Mapa ciepła przedstawiająca skład taksonomiczny badanych odczytów. Na osi *OX* przedstawiona jest nazwa próbki, a na osi *OY* nazwa gatunku. Próbki podzielone są na cztery panele, każdy odpowiada zbiornikowi, z którego pochodziły. Skala kolorów odpowiada bogactwu danego gatunku w badanej próbce. Czerwony kolor odpowiada wysokim wartościom, a niebieski niskim.

Drzewo z kropkami składu taksonomicznego



Rysunek 4.12: Drzewo filogenetyczne z zaznaczonymi na gałęziach informacjami o próbkach. Na gałęziach widoczna jest nazwa rodzaju. Kolorem czerwonym, zielonym, niebieskim i fioletowym zaznaczono zbiorniki odpowiednio: Ceglów, Tały, Urwitałt i Zabłotnia. Punkty w kształcie koła i trójkąta odpowiadają próbkom pobranym w roku odpowiednio 2017 i 2019. Rozmiar punktów reprezentuje liczebność grupy - im większy punkt, tym większe bogactwo danego rodzaju.



Rysunek 4.13: Wykres słupkowy przedstawiający skład taksonomiczny badanych odczytów. Na osi OX przedstawiony jest rok pobrania próbki, a na osi OY liczebność danego rodzaju. Próbki podzielone są na cztery panele, każdy odpowiada zbiornikowi, z którego pochodziły. Skala kolorów odpowiada rodzinie organizmu.

5. Podsumowanie

Od wielu lat Euglenida jest obiektem badań naukowców. Pojawienie się filogenetyki molekularnej i połączenie danych molekularnych i morfologicznych wywarło ogromny wpływ na rozumienie pokrewieństw całej grupy euglenidów.

Celem pracy było badanie i opis jakości danych, a następnie przeanalizowanie składu taksonomicznego za pomocą różnych statystyk, wykresów oraz wykonanie analiz α -różnorodności i β -różnorodności.

Na podstawie otrzymanych wyników można zauważyc, że analizowane czynniki mają wpływ na społeczności autotroficznych euglenin występujące w małych zbiornikach wodnych. Analiza α -różnorodności i β -różnorodności pozwoliła na wyciągnięcie zgodnych wniosków, że najbardziej różnicującymi zmiennymi były zbiornik i region, z którego pochodziły próbki. Różnice te były istotne statystycznie, niezależnie od wybranego testu i wykorzystywanej metody. Jednocześnie zaobserwowano, że próbki pochodzące z roku 2017 i 2019 nie różnią się od siebie. Potwierdziły to testy statystyczne, które wykazały, że dla tej zmiennej p-wartość osiągała wartość powyżej progu równego 0.05.

6. Dostępność kodu i danych

Wszystkie dane użyte do analiz dostępne są na stronie Github (<https://github.com/juliasmolik/MiFM>). Znajdują się tam otrzymane pliki z sekwencjonowania, jak również skrypty w Pythonie służące do generowania dodatkowych danych, które również umieszczone są w repozytorium. Na stronie Github dostępny jest także skrypt w R, którego użycie do analizy danych z sekwencjonowania.

Literatura

- [1] Campbell, N. A. et al. *Biology. Eighth Edition* (Pearson Education, Inc., 2008).
- [2] Leander, B. S., Lax, G., Karnkowska, A. & Simpson, A. G. B. *Euglenida*, 1047–1088 (Springer International Publishing, 2017).

- [3] Kostygov, A. Y. *et al.* Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses. *Open Biol* **11**, 200407 (2021).
- [4] Zakryś, B., Milanowski, R. & Karnkowska, A. Evolutionary origin of euglena. *Adv. Exp. Med. Biol.* **979**, 3–17 (2017).
- [5] Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature Biotechnology* **37**, 852–857 (2019).
- [6] Andrews, S. *et al.* FastQC: a quality control tool for high throughput sequence data. Babraham Institute (2010). URL <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- [7] Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
- [8] Callahan, B. J. *et al.* Dada2: high-resolution sample inference from illumina amplicon data. *Nature methods* **13**, 581 (2016).
- [9] Bokulich, N. A. *et al.* Optimizing taxonomic classification of marker-gene amplicon sequences with qiime 2’s q2-feature-classifier plugin. *Microbiome* **6**, 90 (2018).
- [10] McKinney, W. Data structures for statistical computing in python. In van der Walt, S. & Millman, J. (eds.) *Proceedings of the 9th Python in Science Conference*, 51 – 56 (2010).
- [11] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *Journal of machine learning research* **12**, 2825–2830 (2011).
- [12] Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
- [13] Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* **26**, 1641–1650 (2009).