

Autotroficzne eugleniny w małych zbiornikach wodnych

Jakub J. Guzek

Uniwersytet Warszawski, Wydział Matematyki, Informatyki i Mechaniki, Metagenomika i filogenetyka molekularna,
Student ID: 456616

Formalnie, w tej pracy, eugleniny będą definiowane jako grupa organizmów przynależąca do supergrupy Excavata, i klasy Euglenida. Jedną z charakterystycznych cech organizmów z tej grupy jest obecność pelikuli – białkowej struktury otaczającej komórkę. Większość euglenin jest zdolna do fotosyntezy, a wielu przedstawicieli tej grupy ma zdolność do fagocytozy. Organizmy te występują często w słodkowodnych zbiornikach wodnych. W tej pracy przedstawiam analizę fragmentu zbioru danych metagenomicznych zebranych w trakcie badań prowadzonych w latach 2017-2019 w kilkunastu zbiornikach wodnych zlokalizowanych w różnych rejonach Polski.

Wstęp

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

Forward: CTGTGAATGGCTCCTTACATCAG

Reverse: CTSCCTCTCCGGAATCRAAC

Tak otrzymane amplikony poddano sekwencjonowaniu wysokoprzepustowemu i otrzymano pliki w formacie fastq z odczytami przypisanymi do poszczególnych próbek.

Poniżej prezentuję wyniki przeprowadzonej przeze mnie analizy danych i odpowiedzi na pytania z Zadania 1 z Metagnomiki i filogenetyki molekularnej.

Kontrola jakości danych

W analizowanym podzbiorze danych, ilość odczytów w próbkach wała się od ok. 75 000 do ok. 175 000 tys. (Rys. 1a); po roku w trakcie której pobrano próbki zdawała się nie mieć wpływu na ilość otrzymanych odczytów, ale liczba odczytów silnie wała się pomiędzy zbiornikami. Zbiorniki Izdebno 1 i Izdebno Nowe miały średnio najwięcej odczytów – tylko jedna próbka (198_INo) z tych dwóch zbiorników, nie miała większej ilości odczytów od próbek z pozostałych zbiorników (Rys. 1b). To miało przełożenie, na średnio większą ilość odczytów uzyskanych ze zbiorników położonych na Mazowszu (Izdebno 1, Izdebno Nowe i Załotnia) w porównaniu do tych położonych na Kaszubach (Choczweo, Rybno1 i Rybno2).

Większość uzyskanych odczytów była długości 250 nukleotydów, lub dłuższa (Rys. 1c), co jest raczej typową długością dla sekwencjonowania w technologią Illumina. Próbki zostały zsekwencjonowane trybem Paired End, co oznacza, że każda częsteczka kwasu nukleinowego z bibliotek została zsekwencjonowana najpierw od jednej strony a następnie z drugiej. W związku z tym dla

każdej próbki dostępne były dwa pliki .fastq – jeden z odczytami *forward* oraz jeden z odczytami *reverse*. Odczyty *forawrd* były średnio dłuższe od odczytów *reverse* (Rys. 1c), czego należy się spodziewać w sekwencjonowaniu technologią Illumina. Na początku większości odczytów we wszystkich próbkach występowały sekwencje starterów użytych do uzyskania sekwencjonowanych amplikonów (Rys. 1d).

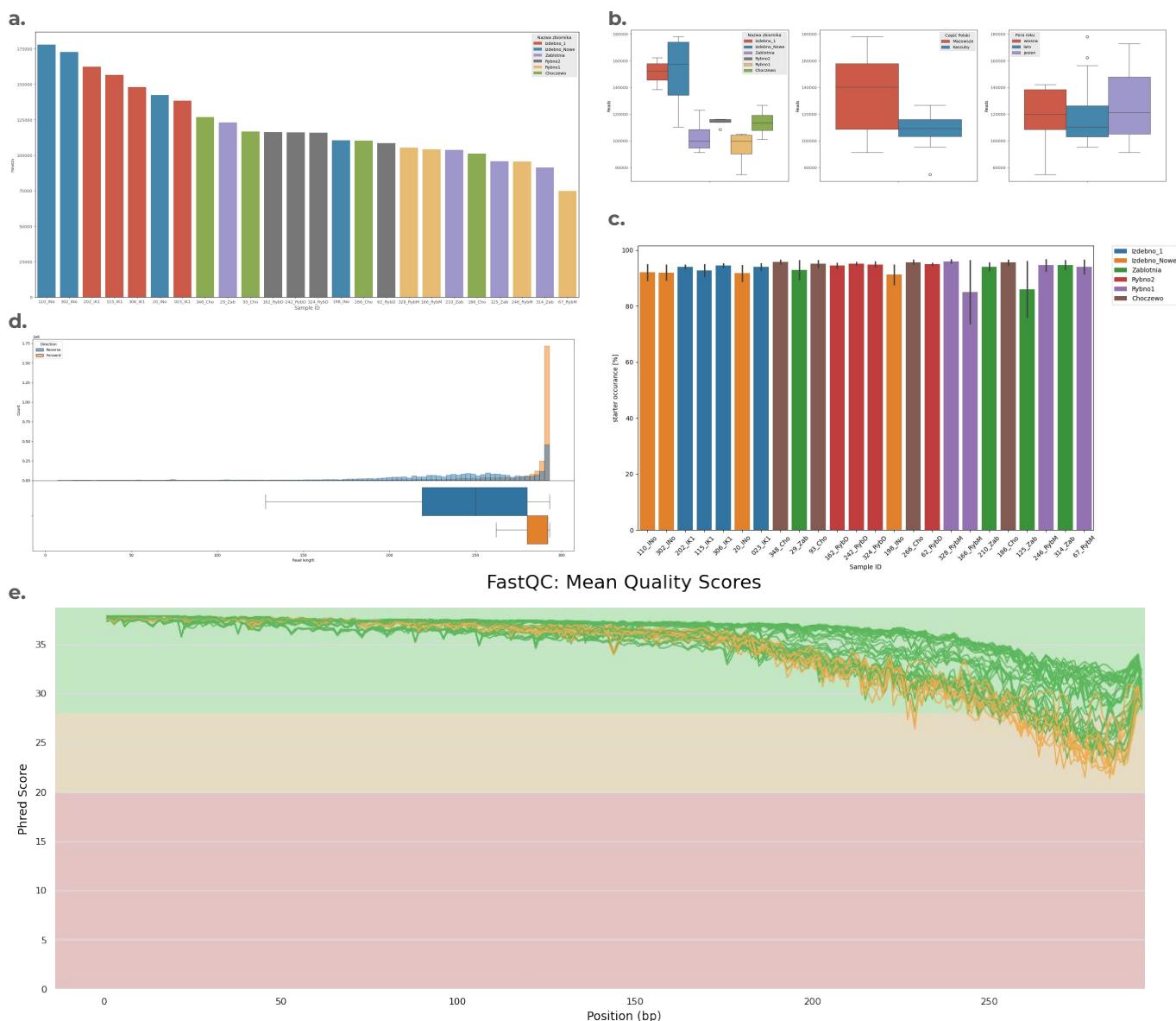
Uzyskane odczyty cechowały się dobrą jakością na prawie całej długości sekwencji, za wyjątkiem samych końców, gdzie widać wyraźny spadek jakości (Rys. 1e); spadek ten jest silniejszy i zaczyna się wcześniej, ok. 200-220 nukleotydów, dla odczytów *reverse* (Rys. uzup. 1), w stosunku do odczytów *forawrd*, dla których zaczyna się ok 240 nukleotydów.

Na podstawie uzyskanych wyników, dobrane zostały następujące parametry filtrowania i przycinania surowych odczytów: (1) przycięcie odczytów do długości 240, 220, 200 dla odczytów *forward* i 220, 200, 180 dla odczytów *reverse*, (2) przycięcie 15 lub 40 nukleotydów z początku odczytów. (3) odrzucenie odczytów, w których jakość dla dowolnej pozycji była poniżej 15, 12 lub 10 phred score.

Procesowanie danych

Decyzja o przycięciu pewnej ilości nukleotydów z początku odczytów została podjęta na podstawie dużej powtarzalności początków sekwencji we wszystkich próbkach (Rys. uzup. 2). Wszystkie kombinacje wymienionych wcześniej parametrów zostały przetestowane (Rys. uzup. 3) i ostatecznie wybrana została kombinacja zachowująca średnio największą ilość odczytów po usunięciu chimer: (1) przycięcie *forawrd* do 240 nt. (2) przycięcie *reverse* do 200 nt. (3) usunięcie 40 nukleotydów z początku wszystkich odczytów oraz (4) odrzucenie wszystkich odczytów o jakości poniżej 10. Dla tych parametrów po usunięciu chimer zachowane zostało od 23.71% (166_RybM) do 58.96% (93_Cho) odczytów (mediana dla próbek: ~40%), co przekładało się na 3 166 sekwencji reprezentatywnych (ASV – ang. amplicon sequence variants). Wszystkie dalsze analizy były prowadzone przy wykorzystaniu ASV uzyskanych z tymi parametrami.

Wszystkie kombinacje parametrów zachowywały relatywnie dużo odczytów po wstępny filtrowaniu, jednak niektóre kombinacje prowadziły do bardzo małych ilości złączonych i niechimerycznych odczytów – w skrajnych



Rysunek 1 | Wyniki analizy jakości danych. a. Wykres słupkowy przedstawiający ilość odczytów w poszczególnych plikach; kolory oznaczają zbiornik z którego została pobrana dana próbka. b. Wykresy pudełkowe przedstawiające rozproszenie i kształt rozkładu ilości odczytów pogrupowanej po: zbiornikach, rejonach polski i porach roku. c. Histogram i wykres pudełkowy przedstawiające długości odczytów, pogrupowane po odczytach forward i reverse. d. Procent odczytów, w których obecna była sekwencja startera, dla wszystkich próbek; słupki błędów reprezentują różnice między odczytami forward i reverse dla danej próbki. e. Średnia jakość odczytów na danej pozycji w odczycie, dla wszystkich próbek.

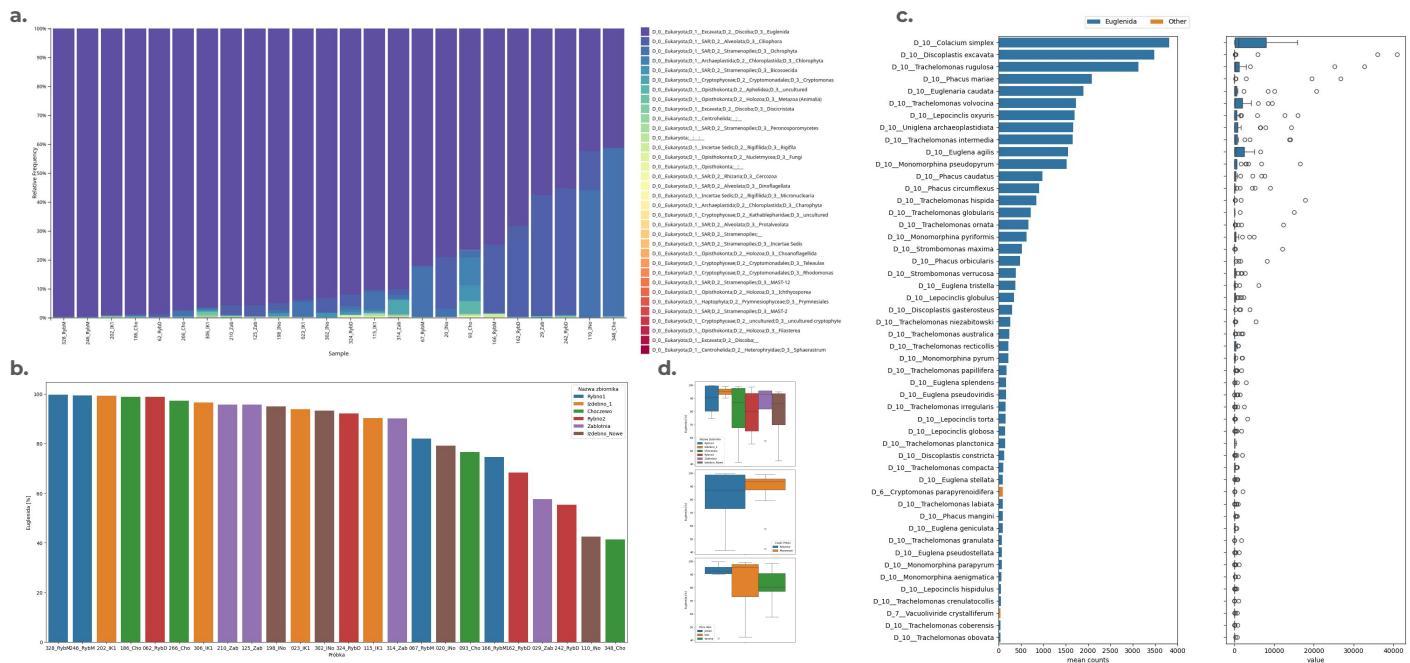
wypadkach do bliskich zeru w niektórych próbkach (Rys. uzu. 3). Te wyjątkowo niskie wyniki były obserwownalne głównie dla kombinacji parametrów, w których zarówno odczyty *forward* jak i *reverse* były skrajnie przycinane (200 nt. i 180 nt.), co zapewne prowadziło do tego, że obszar nałożenia się między nimi był niewystarczająco długi do poprawnego złączenia dwóch końców sekwencji amplikonu (fragment V2 rDNA ma 250 nt.).

Przyporządkowanie taksonomiczne

Po filtrowaniu i połączeniu amplikonów, uzyskane ASV zostały przyporządkowane do znanej taksonomii. Do klasyfikacji użyty został gotowy wytrenowany klasyfikator (**klasyfikatorEu_04.04.24.qza**). Następnie na podstawie uzyskanej klasyfikacji sprawdzone zostały składy taksonomiczne poszczególnych próbek. Najliczniejszymi grupami taksonomicznymi we wszystkich próbkach były: Euglenida,

Ciliophora, Ochrophyta, Chlorophyta i Bicosoecida (Rys. 2a). Odczyty zaklasyfikowane do euglenin stanowiły ponad 50% we wszystkich próbach oprócz dwóch (Rys. 2b). Średnio najczęstsze były ASV przypisane do gatunków: (1) *Colacium simplex*, (2) *Discoplastis excavata* i (3) *Trachelomonas rugulosa* (Rys. 2c).

Największy procentowy udział euglenin miały zbiorniki Rybno1 i Izdebno 1. Udział procentowy wydaje się być nieznacznie mniejszy w próbkach zebranych wiosną w porównaniu do tych zebranych latem i jesienią. Natomiast rejon Polski nie wydaje się mieć znaczącego wpływu na ilość sekwencji euglenin w próbkach¹ (Rys. 2d).



Rysunek 2 | Próbki i poszczególne zbiorniki znacząco różnią się składami taksonomicznymi, oraz udziałem odczytów przypisanych do euglenin. a. Wykres słupkowy przedstawiający skład taksonomiczny poszczególnych próbek jako procentowy udział wybranych grup taksonomicznych. b. Wykres słupkowy przedstawiający procentowy udział ASV zaklasyfikowanych do grupy Euglenida w poszczególnych próbkach. c. Wykres słupkowy (lewo) pokazujący średnią ilość odczytów przypisanych do danego gatunku i wykres pułapkowy (prawo) przedstawiający rozproszenie i rozkład ilości odczytów przypisanych do danego gatunku w próbkach. d. Wykresy pułapkowe pokazujące rozproszenie udziału procentowego ASV przypisanych do Euglenida w różnych zbiornikach, rejonach Polski i porach roku.

Skład taksonomiczny

W celu przeprowadzenia właściwych analiz dotyczących bioróżnorodności euglenin w otrzymanych próbkach, dane otrzymane w poprzednich krokach (ASV, tabela z liczebnością ASV przyporządkowanych do taksonów w próbkach, drzew filogenetyczne zaklasyfikowanych ASV oraz metadane), zostały wczytane do obiektu *phyloseq* przy pomocy biblioteki *qiime2R*. Na tym etapie wszystkie operacje przeprowadzane były na poziomie pojedynczych ASV, stanowiących teraz OTU (*Operational Taxonomic Unit*). Otrzymany obiekt *phyloseq* przefiltrowano tak aby usunąć z niego wszystkie OTU nienależące do grupy Euglenida, a następnie zapisano go do pliku, w formacie *rds*, do późniejszych analiz.

Przed odfiltrowaniem danych niedotyczących euglenin obiekt *phyloseq* zawierał 3 166 OTU oraz 7 poziomów taksonomicznych; próbka z najmniejszą liczbą odczytów (67_RybM) miała ich 22 003, a próbka z największą liczbą (348.Cho) miała ich 71 507. Po odfiltrowaniu liczba OTU spadła do 2 776, a minimalna i maksymalna liczba odczytów w próbkach wynosiła odpowiednio 18 046 (67_RybM) i 56 098 (202_IK1) (Rys. 3b). Jak łatwo zauważać, próbka 348.Cho nie jest już najliczniejszą próbka po odfiltrowaniu OTU nienależących do euglenin – jest to spodziewane biorąc pod uwagę, że jest to próbka w której ASV przypisane do tej grupy miały najmniejszy procentowy udział (Rys. 2b).

Aby uniknąć wrażenia fałszywej dokładności analiz, ASV zostały połączone w gatunki do których zostały przypisane. Wszystkie kolejne kroki zostały przeprowadzone na OTU dopowiadających gatunkom, połączonym z przypisanych

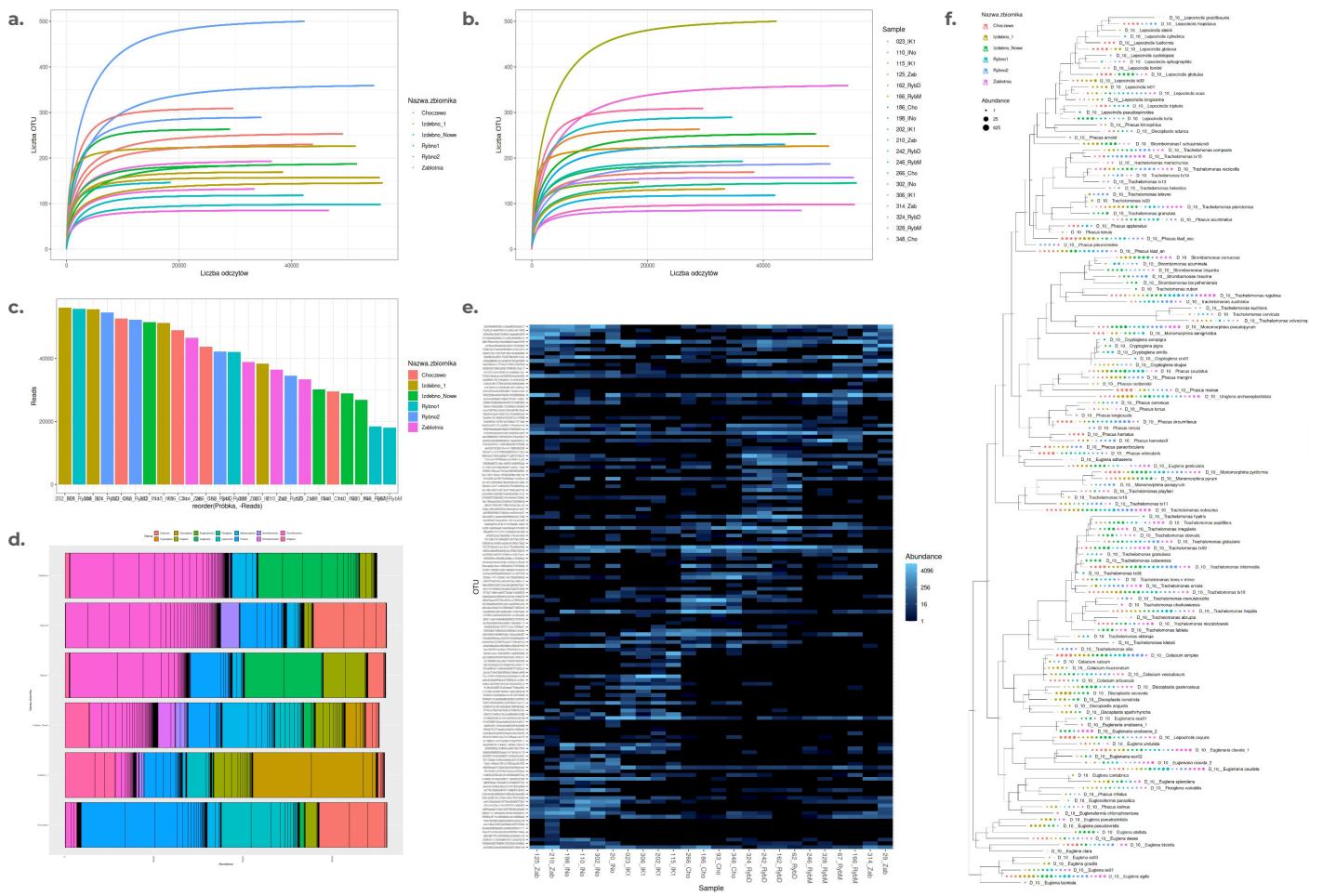
¹Żeby dokładnie to ocenić należałoby sprawdzić istotność statystyczną różnic między grupami, ale zabrakło na to czasu

do nich ASV. Następnie przeprowadzono analizę wysycenia próbek i stworzono krzywe wysycenia (Rys. 3a), aby ocenić strategię próbkowania w dół. Na podstawie krzywych wysycenia zdecydowano aby w czasie próbkowania w dół sprowadzić liczbę odczytów w próbkach do liczby odczytów w najmniej licznej próbce (18046 odczytów – 67_RybM, Rys. 3b) – w ten sposób nie musimy pozbywać się żadnej próbki z analizy, nie tracąc przy tym znacząco informacji o różnorodności biologicznej z bardziej licznych próbek.

Po próbkowaniu w dół, otrzymano próbki o równej liczności odczytów, co znacząco ułatwia przeprowadzanie i interpretację analiz. Przed przystąpieniem do analiz różnorodności biologicznej przeprowadzono eksploracyjną analizę danych poprzez wizualizację składu taksonomicznego. Zgodnie z tym co zaobserwowano przed wyrównaniem wielkości próbek (Rys. 2c) organizmy z rodzaju *Trachelomonas* były najliczniejszą grupą w wielu zbiornikach; wraz z rodzajami *Colacium*, *Euglenaria* i *Phacus* stanowiły większość organizmów (Rys. 3d).

Eksploracyjna analiza danych wskazywała również na większe podobieństwo między próbками z tego samego zbiornika, w stosunku do podobieństwa między próbками z różnych zbiorników, bez wyraźnej segregacji po porach roku, czy regionie Polski (Rys. 3e). Nie zaobserwowano również klarownej zależności między pokrewieństwem filogentycznym poszczególnych gatunków, a ich występowaniem w konkretnych zbiornikach, czy regionach lub w czasie konkretnych pór roku (Rys. 3f).

Aby zweryfikować te wstępne obserwacje przeprowadzono ilościowe analizy różnorodności biologicznej.



Rysunek 3 | Zbiorniki różnią się pod względem ilości odczytów, obecnych OTU i ilością OTU przypisanych do różnych rodzajów Euglenida
a. b. Krzywe wysycenia dla wszystkich próbek, pokolorowane po zbiorniku (**a**) lub po próbce (**b**). **c.** Liczba odczytów w poszczególnych próbkach po odfiltrowaniu odczytów nienależących do grupy Euglenida. **d.** Wykres słupkowy obrazujący skład gatunkowy euglenin w różnych zbiornikach, pokolorowany po rodzajach. **e.** Heatmapa pokazująca ilość poszczególnych OTU (gatunków) w próbkach. **f.** Drzewo filogenetyczne gatunków ze wszystkich próbek, pokazujące liczebność poszczególnych OTU (gatunków) w próbkach.

Różnorodność biologiczna

Jako pierwszy krok w ocenie różnorodności biologicznej i różnic w różnorodności między próbками, obliczono alfa-różnorodność – tj. wewnętrzną różnorodność próbek – wykorzystując trzy miary różnorodności: obserwowe bogactwo (ilość OTU w próbce), wskaźnik Shannona (entropia Shannona próbki) oraz wskaźnik Simpsona (Rys. uzup. 4a). Tylko jeden z użytych wskaźników (Shannon: p-value = 0.0122) wskazywał na istnienie istotnych statystycznie różnic wewnętrznej różnorodności biologicznej między zbiornikami (Observed: p-value = 0.231, Simpson: p-value = 0.062). Żaden ze wskaźników nie wskazywał na istotne statystycznie różnice w wewnętrznej różnorodności biologicznej pomiędzy próbками z różnych części Polski, czy z różnych pór roku.

Aby potwierdzić hipotezy postawione po eksploracyjnej analizie danych, o większym podobieństwie między próbками z tego samego zbiornika, analiza alfa-różnorodności jest niewystarczająca. W tym celu przeprowadzono analizę beta-różnorodności, która pozwala na kwantyfikację różnic w różnorodności pomiędzy próbками. Jako wskaźników beta-różnorodności użyto odległości Jaccarda, która jest miarą podobieństwa pomiędzy skończonymi zbiorami próbek, definiowaną jako wielość przecięcia tych zbiorów

podzieloną przez wielkość sumy tych zbiorów:

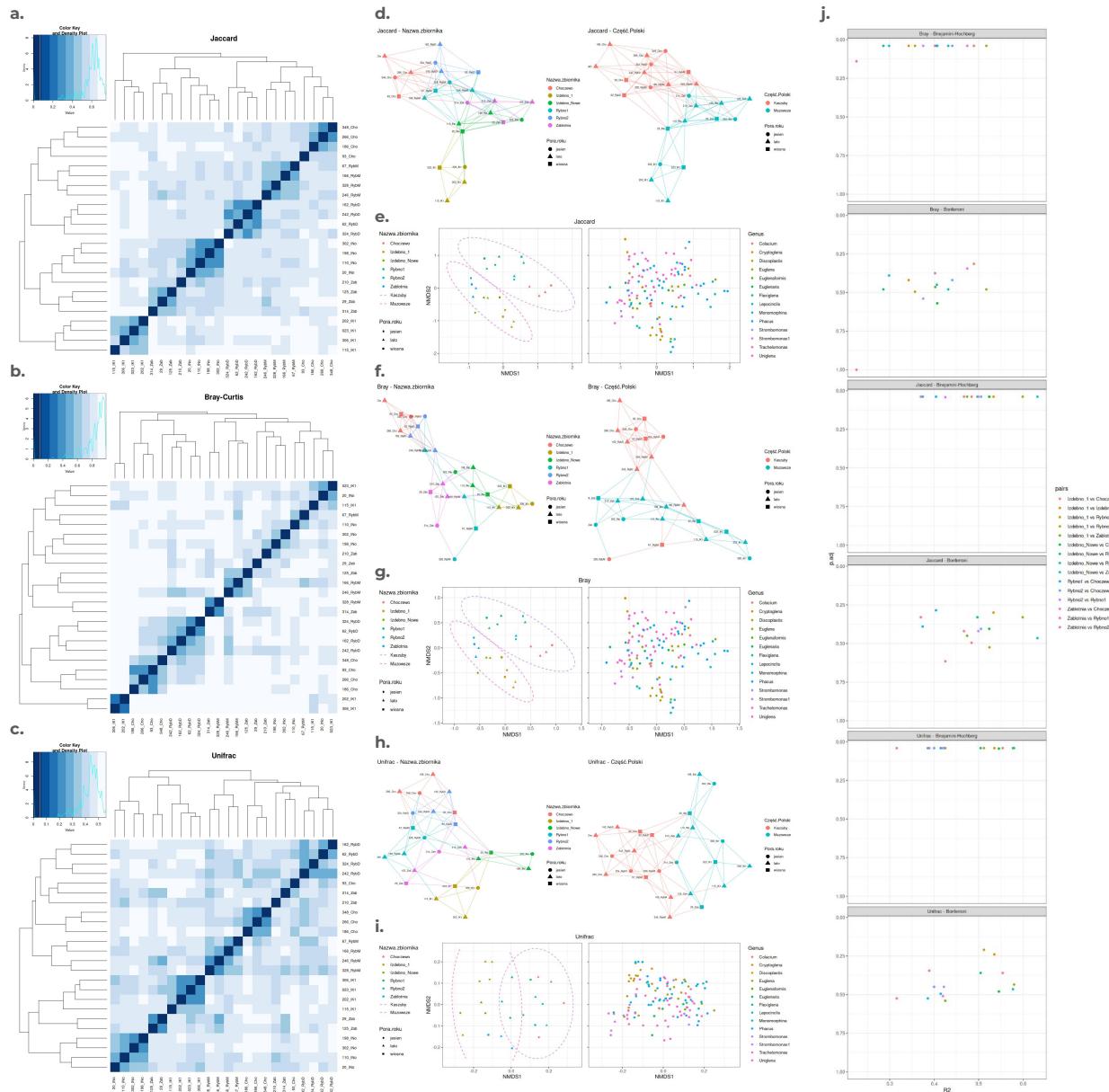
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \quad (1)$$

miary różnicy Braya-Curtisa, definiowanej jako:

$$BC_{jk} = 1 - \frac{2 \sum_{i=1}^p \min(N_{ij} N_{ik})}{\sum_{i=1}^p (N_{ij} + N_{ik})} \quad (2)$$

gdzie N_{ij} jest liczbą odczytów dla gatunku i w próbce j , N_{ik} jest liczbą odczytów dla gatunku i w próbce k , a p jest całkowitą liczbą gatunków w obydwu próbках, oraz odległość unifrac, której definicja jest bardziej skomplikowana, ale co istotne, bierze pod uwagę dystanse filogenetyczne pomiędzy gatunkami obecnymi w próbках (na podstawie drzewa filogenetycznego). Odległość Jaccarda i unifrac nie biorą pod uwagę ilościowych danych na temat liczby odczytów dla poszczególnych OTU w porównywanych próbках, podczas gdy miara Braya-Curtisa bierze pod uwagę liczbę odczytów.

Analiza beta-różnorodności przy pomocy odległości Jaccarda i Unifrac, popiera hipotezę o niższej różno-



Rysunek 4 | Różnorodność biologiczna pomiędzy zbiornikami jest znacznie większa niż wewnętrz zbiorników, oraz różnorodność między rejonami Polski jest znacznie większa niż wewnętrz jednego rejonu a. b. c. Heatmapa obrazująca beta-różnorodność między próbami mierzona przy pomocy odległości Jaccarda (a) Braya-Curtisa (b) i unifrac (c). d. f. h. Sieć obrazująca podobieństwo między próbami przy pomocy odległości Jaccarda (d, min.dist = 0.6), Braya-Curtisa (f, min.dist = 0.8) i unifrac (h, min.dist = 0.4), pokolorowana po zbiorniku (lewo) i rejonie Polski (prawo) e. g. i. Ordynacja NMDS pokazująca redukcję wymiarowości przestrzeni odległości Jaccarda (e), Braya-Curtisa (g) i unifrac (i), dla próbek (lewo) i gatunków (prawo). Elipy obrazują rejon Polski, z którego pochodzą próbki. j. Wykresy obrazujące istotność statystyczną różnic pomiędzy poszczególnymi parami zbiorników, oraz ułamek zmienności tłumaczyony przez zmienność między daną parą. Statystyki wyliczone przy pomocy korekty Bonferroniego i przy pomocy procedury Benjamini-Hochberga.

rodności (większym podobieństwie) pomiędzy próbami z tego samego zbiornika, niż pomiędzy próbami z różnych zbiorników. Najbardziej klarownie zależność ta jest widoczna w wypadku odległości Jaccarda. Hierarchiczna analiza skupień odległości Jaccarda pomiędzy próbami, grupuje niemal wszystkie próbki (za wyjątkiem 201_Zab, która jest zgrupowana z próbami z Izdebna Nowego) na podstawie zbiornika, z którego pochodzą (Rys. 4a). Ta sama zależność dla odległości Jaccarda jest również wyraźnie widoczna przy wizualizacji odległości jako sieci ($\text{max.dist} = 0.6$) i jako ordynacji. Sieci i ordynacja pokazują również grupowanie próbek na podstawie rejonu Polski z którego pochodzą, ale nie na podstawie pory roku, w czasie której zostały pobrane (Rys. 4d, e).

Miara Braya-Curtisa nie wskazuje wyraźnie na większe podobieństwo wewnętrz zbiorników, ale zdaje się popierać obserwacje o większym podobieństwie próbek z

tego samego rejonu Polski. Hierarchiczna analiza skupień wskazuje na podobieństwo próbek z tych samych rejonów Polski, za wyjątkiem 67_RybM, która grupuje się z próbami z Mazowsza i 125_Zab, która grupuje się z próbami z Kaszub (Rys. 4b). Jest to również widoczne na sieci ($\text{max.dist} = 0.8$), ale mniej wyraźnie niż dla odległości Jaccarda (Rys. 4f)².

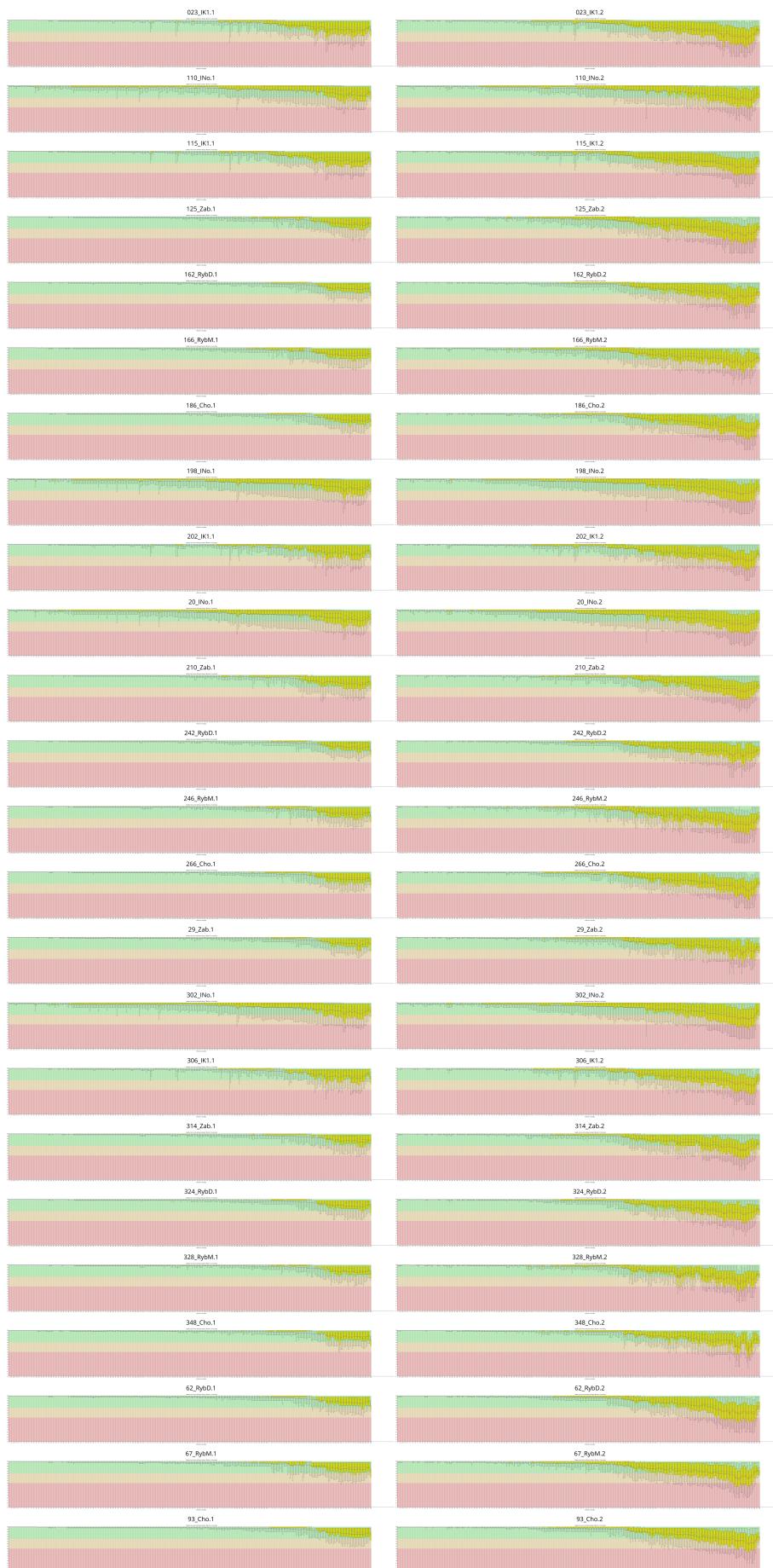
Powyzsze obserwacje o większych różnicach pomiędzy zbiornikami i rejonami Polski są również wyraźne dla analiz przy użyciu odległości unifrac, co wskazuje że różnice te wynikają nie tylko ze składu próbek, ale również z odległości filogenetycznych między OTU w próbach (Rys. 4c, h, i).

Analiza podobieństw ANOSIM, także wskazuje na

²Ordynacja dla miary Braya-Curtisa jest z nieznanego mi powodu taka sama jak dla odległości Jaccarda. Niestety nie miałem czasu dokładnie zbadać dlaczego.

istnienie znacząco większych różnic między zbiornikami niż wewnętrz, oraz między rejonami Polski niż wewnętrz, ale nie pomiędzy porami roku (Rys. uzup. 4d). Należy przy tym pamiętać, że zbiornik pochodzenia próbki i rejon pochodzenia próbki są zmiennymi wzajemnie splątanymi, tj. zbiornik pochodzenia próbki jednoznacznie dyktuje rejon pochodzenia. Sparowana analiza wariancji (pairwise PERMANOVA) również potwierdza istnienie znaczących różnic między zbiornikami (chyba, że użyta zostanie procedura korekcji Bonferroniego dla wielokrotnych testów) (Rys. 4j) oraz między rejonami Polski (adjusted p-value ≤ 0.001 dla wszystkich miar; Bonferroni i Benjamini-Hochberg).

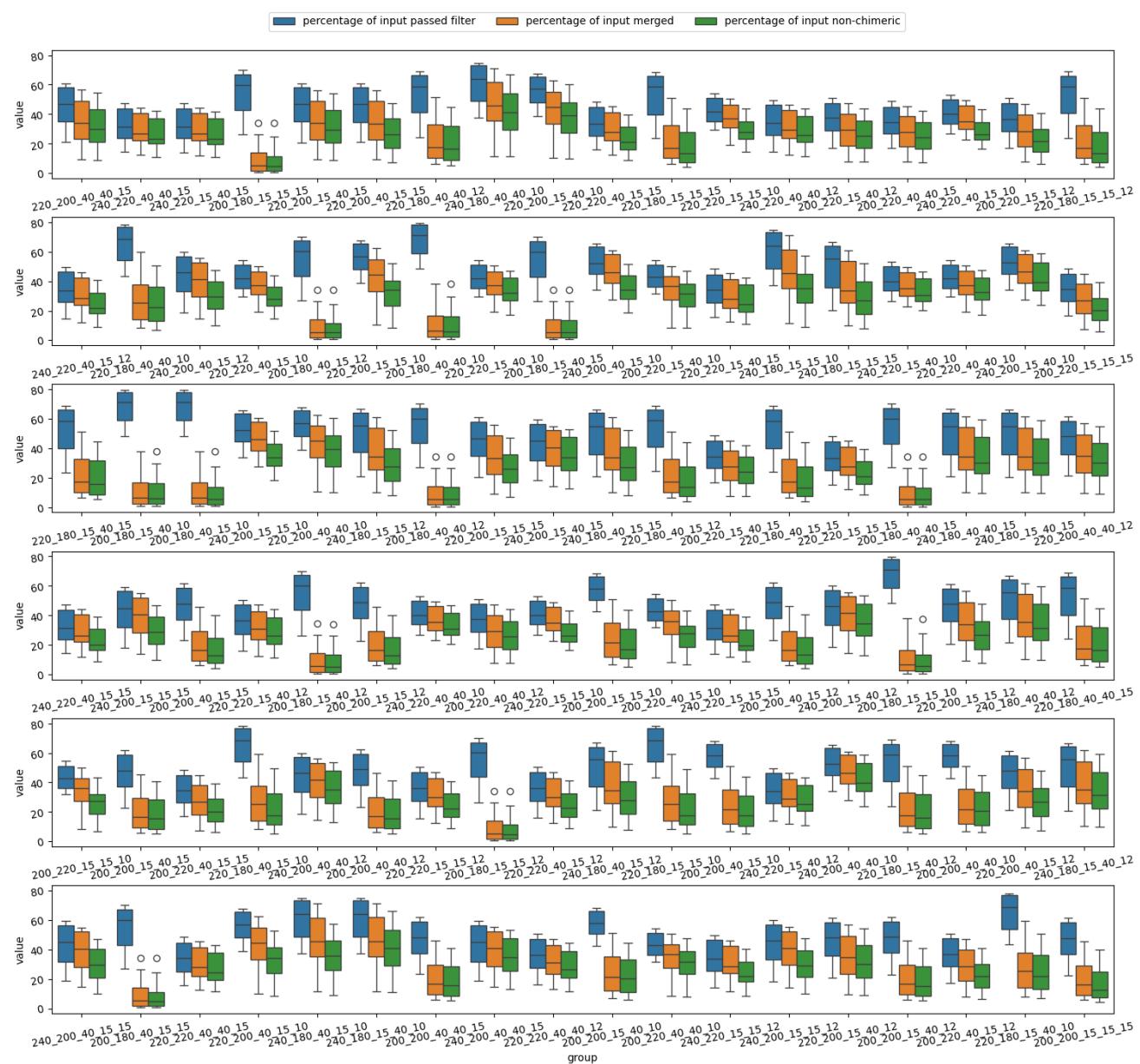
Podsumowanie



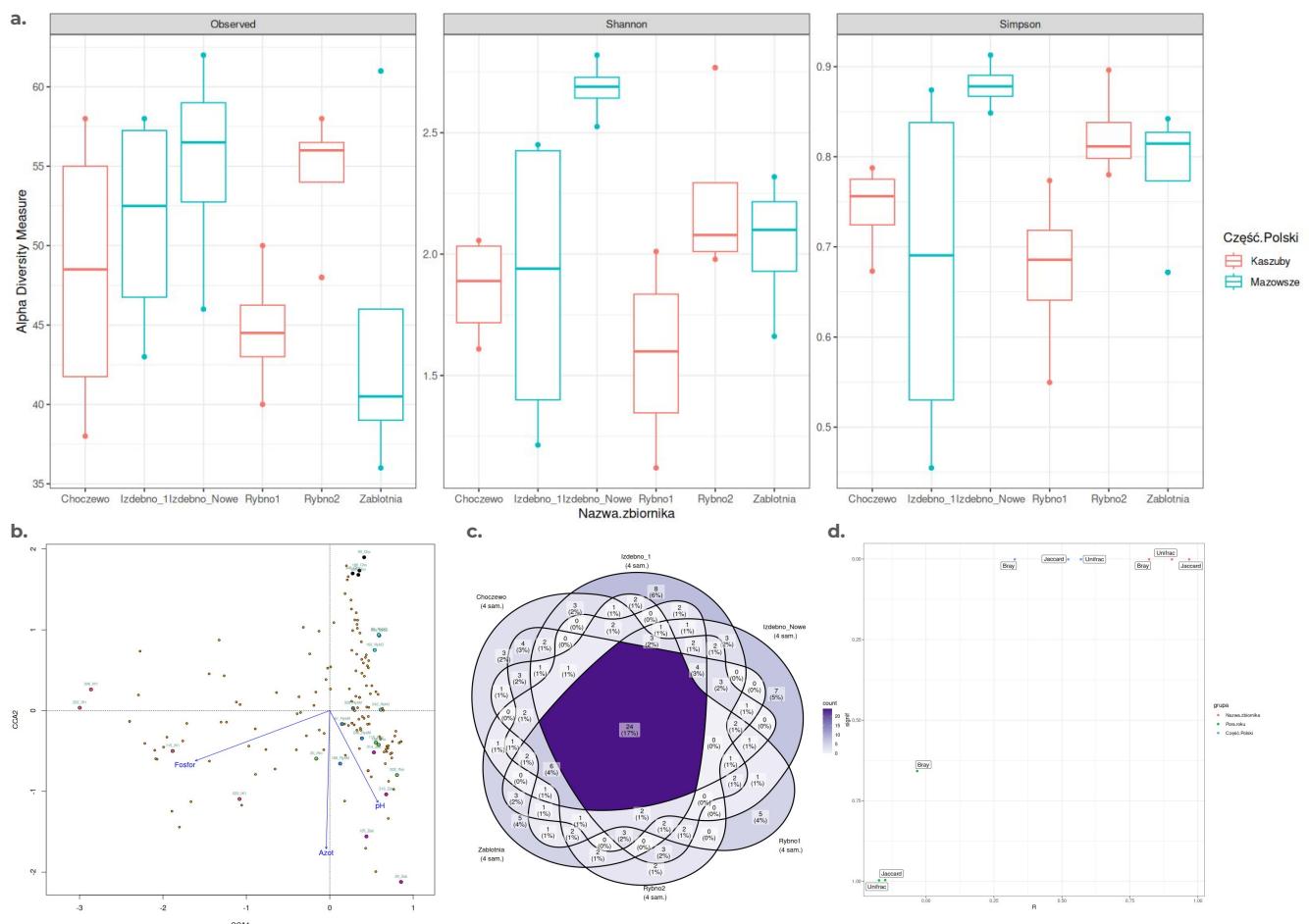
Rysunek uzupełniający 1 | Jakość odczytów na poszczególnych pozycjach dla wszystkich analizowanych próbek. W lewej kolumnie pokazane są wykresy jakości dla odczytów *forawrd*, a w prawej odpowiadające wykresy jakości odczytów *reverse* dla tej samej próbki.



Rysunek uzupełniający 2 | Częstość poszczególnych nukleotydów na kolejnych pozycjach dla wszystkich próbek. Duża powtarzalność tych samych sekwencji na początkach odczytów powoduje, że początki sekwencji niosą niską ilość informacji użytecznych w dalszych analizach.



Rysunek uzupełniający 3 | Statystyki filtrowania dla różnych kombinacji parametrów. Wykresy pudełkowe przedstawiające ilości odczytów jakie zostały w próbkach po filtrowaniu, łączeniu i usuwaniu chimerycznych odczytów, dla różnych użytych parametrów filtrowania. Parametry zakodowane są następująco: 220_200_15_15_10 oznacza przycięcie forward do długości 220, reverse do długości 200, usunięcie 15 pierwszych zasad z odczytów forward i 15 z odczytów reverse oraz odrzucenie odczytów o jakości niższej niż 10.



Rysunek uzupełniający 4 | .