Julia Piskorz (202816)

Bioinżynieria zwierząt

**Raport**

**Analizy bioinformatyczne w genomice**

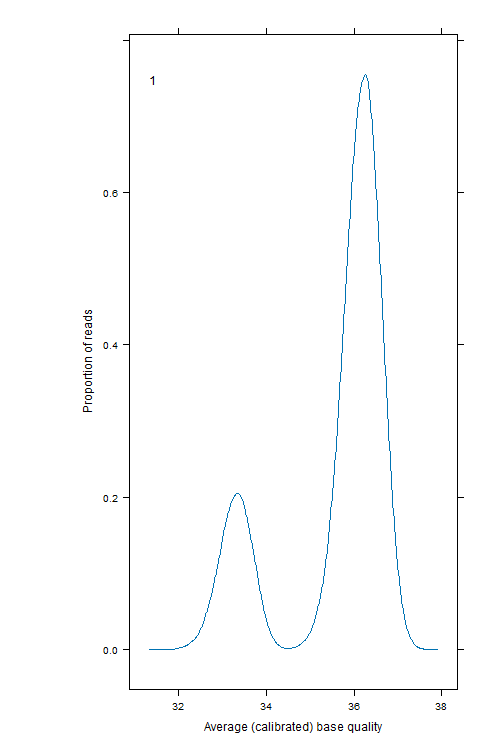
1. **Wprowadzenie**

Celem przeprowadzonej analizy było zmapowanie sekwencji *Escherichia coli* do genomu referencyjnego. W ramach pracy wykonano kontrole jakości danych oraz ich preprocesowanie w tym filtrowanie, przycinanie odczytów o niskiej jakości oraz usunięcie sekwencji adapterów. Podjęte kroki miały na celu maksymalne zwiększenie jakości danych aby uzyskać precyzyjne wyniki mapowania. Dokonano wizualizacji wyników mapowania, co obejmuje szczegółową interpretację uzyskanych danych.

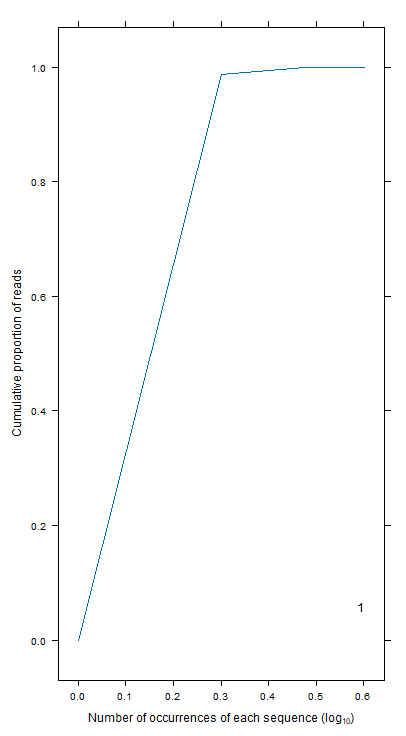
1. **Etapy analizy**
   1. **Wstępna analiza jakości surowych danych sekwencyjnych *E.coli***

W początkowym etapie zainstalowano niezbędne pakiety (Biostrings, ShortRead, Rqc, Rsubread) oraz zaimportowano plik zawierający odczyty sekwencyjne *E.coli*. Wygenerowano raport oceny jakości- „QA\_report-ecoli\_raw.html”, który umożliwił ocenę wielu kluczowych parametrów związanych z jakością danych. Po pierwsze jesteśmy w stanie określić, iż liczba odczytów wynosiła 309440. Dodatkowo w analizie wykresu ogólnej jakości odczyty (ryc.1) zaobserwowano dwa piki. Dominujący pik po prawej stronie wskazuje na regiony o dobrej jakości, natomiast drugi sugeruje obecność fragmentów o niskiej jakości lub możliwych zanieczyszczeń. W analizowanym raporcie obserwujemy kilka powtarzających się sekwencji (jedna sekwencja powtarza się 4-krotnie, pozostałe 3-krotnie oraz 2-krotnie). O występowaniu potencjalnych powtórzeń informuje nas również między innymi krzywa kumulacyjna pokrycia (ryc. 2), w idealnym przykładzie krzywa powinna przechodzić gwałtownie od niskiego do wysokiego pokrycia. W przypadku analizowanych danych, obserwujemy, że przejście jest mniej ostre co świadczy o obecności zanieczyszczeń lub powtórzeń. Z biegiem cykli jakości odczytów spada, co jest typowe dla sekwencjonowania (ryc. 3).

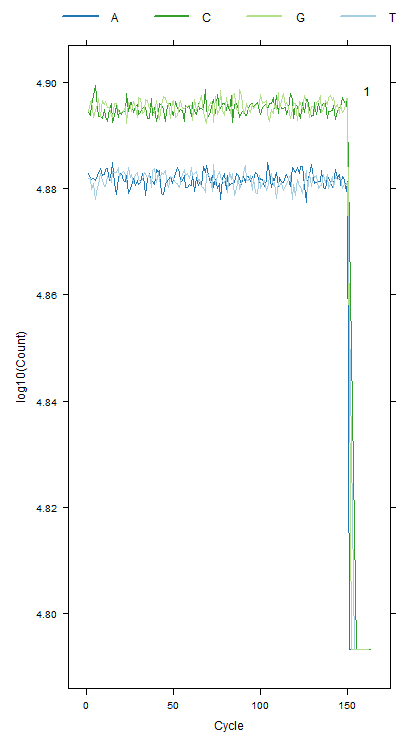
W celu poprawy jakości danych, a także eliminacji potencjalnych błędów, wymagane jest przeprowadzenie preprocessowania odczytów.



**Rycina 1:** Ogólna jakość odczytu



**Rycina 2:** Krzywa kumulacyjna pokrycia odczytów



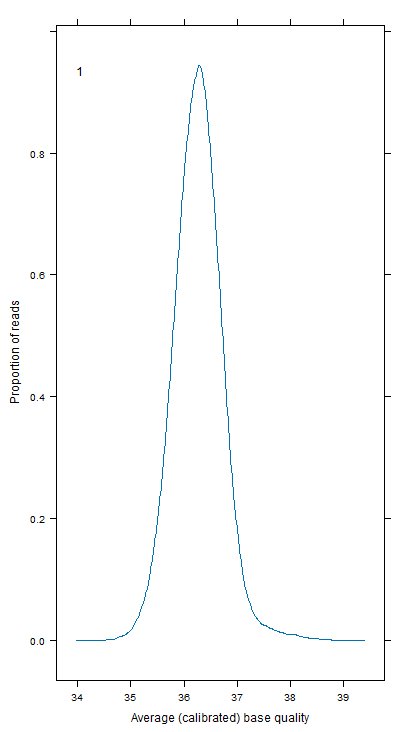
**Rycina 3:** Liczba odczytów zasad nukleotydowych (A, C, G, T) w kolejnych syklach sekwencjonowania.

* 1. **Preprocesing- filtrowanie i przycinanie danych sekwencyjnych**

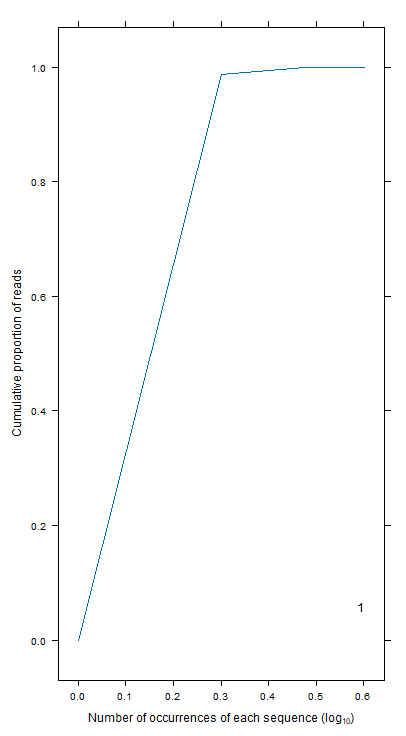
Odczyty przycięto, wykorzystując funkcję trimTailw, która usuwa niskiej jakości bazy na końcach sekwencji. Po wykonaniu preprocesingu wygenerowano nowy raport- „QA\_report\_e.coli\_processed.html”. Wykazał on poprawę jakości danych sekwencyjnych. Dzięki zastosowaniu filtrowania i przycinania pozbyliśmy się zanieczyszczeń powodując, że powstał tylko 1 dominujący pik (Ryc.4) a nie dwa tak jak w danych przed zmianami. W tabeli 1 obserwujemy, że spadła liczba odczytów z 299809 przed filtracją do 291316 po filtracji. Oznacza to, że odrzucono 2.83% odczytów podczas filtracji.

**Tabela 1:** Dane liczbowe przedstawiające zmiany wynikające z przycięcia i filtracji sekwencji E.coli.

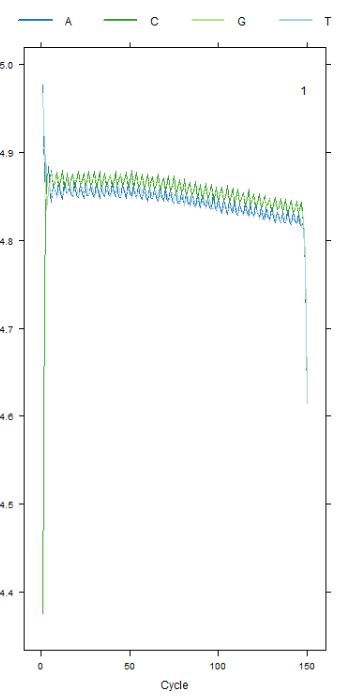
|  |  |
| --- | --- |
| Liczba odczytów w niezmienionej sekwencji | 309440 |
| Ilość przyciętych sekwencji | 95137 |
| Liczba odczytów przed filtracją | 299809 |
| Liczba odczytów po filtracji | 291316 |
| Procent odczytów odrzuconych podczas filtracji | 2.83 % |



**Rycina 4.** Ogólna jakość odczytu



**Rycina 5:** Krzywa kumulacyjna pokrycia odczytów



**Rycina 6:** Liczba odczytów zasad nukleotydowych (A, C, G, T) w kolejnych syklach sekwencjonowania.

* 1. **Wykrywanie i usuwanie sekwencji adapterów**

Zidentyfikowano oraz wyszukano sekwencje adapterów (AGATCGGAAGAGC) w odczytach. Długość przed przycięciem i po przycięciu adapterów nie zmieniła się (wynosi 291316) co oznacza że w sekwencji nie znajdowały się adaptery

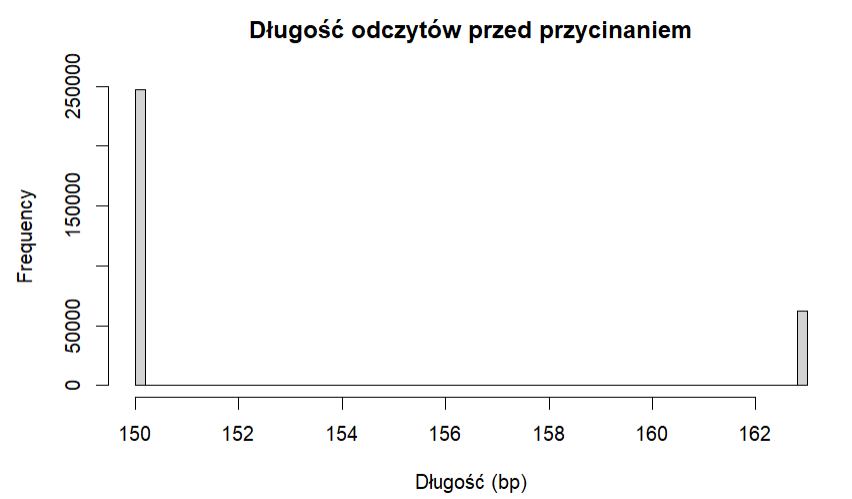
Ilość zmodyfikowanych odczytów: 113288

* 1. **Analiza danych zmodyfikowanej sekwencji**

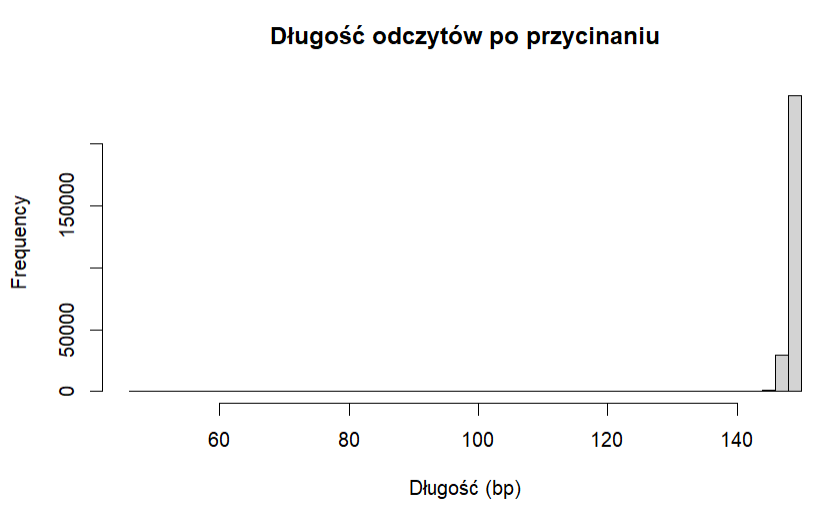
Wygenerowano ostateczny raport w celu analizy danych zmodyfikowanych po przycięciu, filtracji i usunięciu adapterów- „ecoli\_final\_raport.html”. Obserwowana jest poprawa jakości po wykonaniu obróbki danych sekwencyjnych (Ryc. 4). Nie obserwuje się adapterów.

Odczyty przed przycięciem znajdują się w przedziale 150-162pz (wykres 1), wysoki pik w okolicach 150pz oznacza że większość odczytów ma tę długość natomiast w niewielkiej ilości występują też odczyty w okolicach 162pz, może to wskazywać na obecność artefaktów i zanieczyszczeń. Po dokonaniu przycinania i filtracji obserwujemy redukcje długości, jest skondensowana w zakresie ok 140pz i poniżej (wykres 2). Taki rozkład wskazuje, że proces przycinania skutecznie usunął niskiej jakości regiony na końcach odczytów, zdecydowany spadek liczby odczytów o długości powyżej 150 bp wskazuje, że potencjalnie problematyczne fragmenty zostały wyeliminowane. Długość po przycinaniu wykazuje lepszą jednorodność co może pozytywnie wpłynąć na dokładność mapowania do genomu referencyjnego.

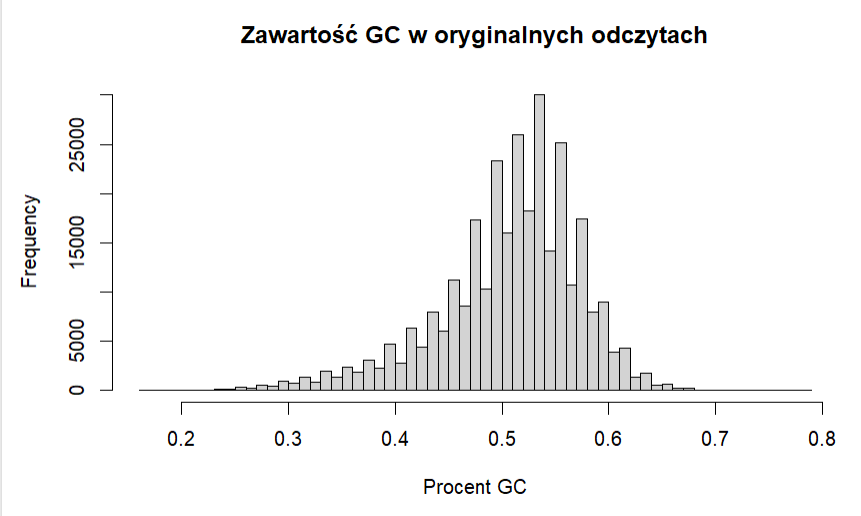
Wykresy przedstawiające zwartość zasad GC w odczytach przed i po przetworzeniu (wykres 3 i 4) charakteryzują się rozkładem zbliżonym do normalnego. Natomiast po filtracji skrajne wartości (poniżej 30% lub powyżej 70%) są mniej widoczne, co sugeruje, że usunięto potencjalne błędy lub odczyty niskiej jakości. Rozkład jest bardziej symetryczny i skoncentrowany w węższym zakresie (przede wszystkim między 45% a 55% GC).



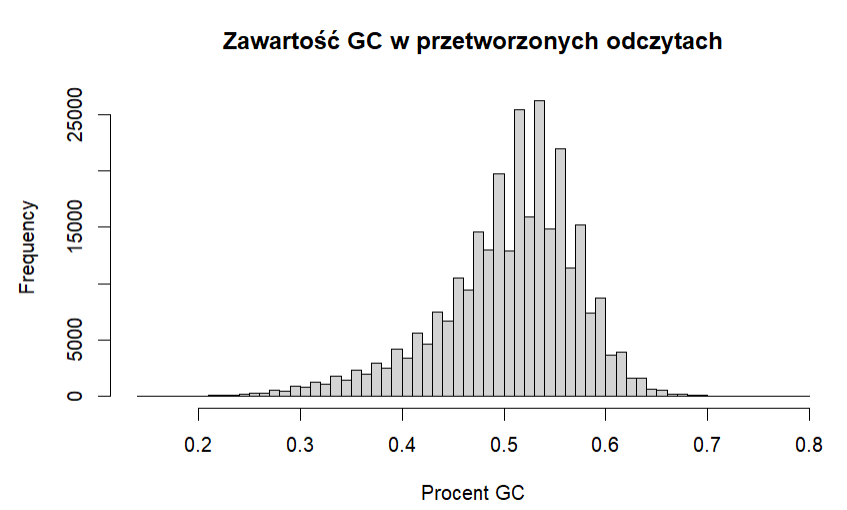
**Wykres 1:** Częstotliwość występowania długość odczytów przed przycięciem danych sekwencyjnych. Długość podana w parach zasad (bp).



**Wykres 2:** Częstotliwość występowania długość odczytów po przycięciu danych sekwencyjnych. Długość podana w parach zasad (bp).



**Wykres 3:** Częstotliwość występowanie zasad GC w oryginalnych odczytach



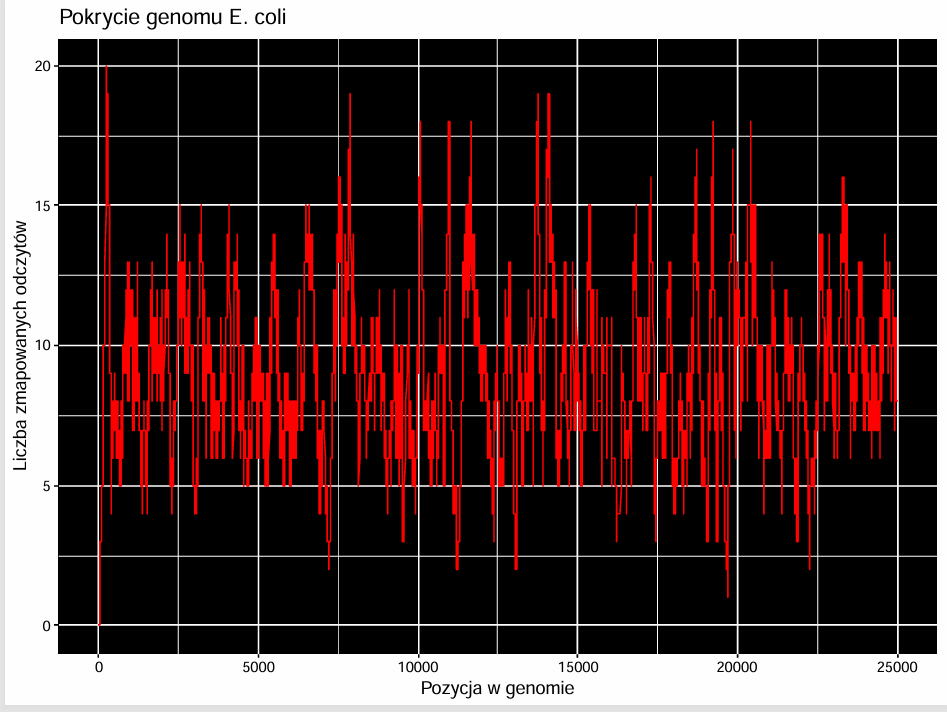
**Wykres 4:** Częstotliwość występowania zasad GC w przetworzonych odczytach

1. **Mapowanie**

Oczyszczone odczyty zostały zmapowane do genomu referencyjnego *E.coli.* Po wykonaniu mapowania mogliśmy dane z tabeli 2 posłużyły nam do obliczenia procentu zmapowanych odczytów który wynosił 99.99% oraz średnią wartość pokrycia która wynosiła 9.15. Liczba unikalnych zmapowanych odczytów wskazuje na wysoką specyficzność mapowania. Uzyskanie 84 niezmapowanych odczytów wskazuje na praktycznie kompletną zgodność danych sekwencyjnych z genomem referencyjnym. Dokonano wizualizacji pokrycia genomu *E.coli* przedstawiony na rycinie 6, określa on liczbę zmapowanych odczytów w danej pozycji w genomie.

**Tabela 2:** Dane liczbowe dotyczące mapowania sekwencjo do genomu referencyjnego

|  |  |
| --- | --- |
| Całkowita liczba odczytów | 291316 |
| Mapowane odczyty | 291232 |
| Unikalne mapowane odczyty | 286087 |
| Liczba odczytów mapujących w wielu miejscach | 5145 |
| Niezmapowane odczyty | 84 |
| Indels | 170 |



**Rycina 7:** Wizualizacja pokrycia danych sekwencyjnych do genomu referencyjnego *E.coli.*

1. **Podsumowanie**

Przeprowadzona analiza doprowadziła do uzyskania wysokiej jakości danych, które zostały zmapowane do genomu referencyjnego *E.coli,* uzyskując wysoki procent zmapowanych odczytów co świadczy o wysokiej specyficzności mapowania.