Martyna Wieczorek nr.albumu 203097

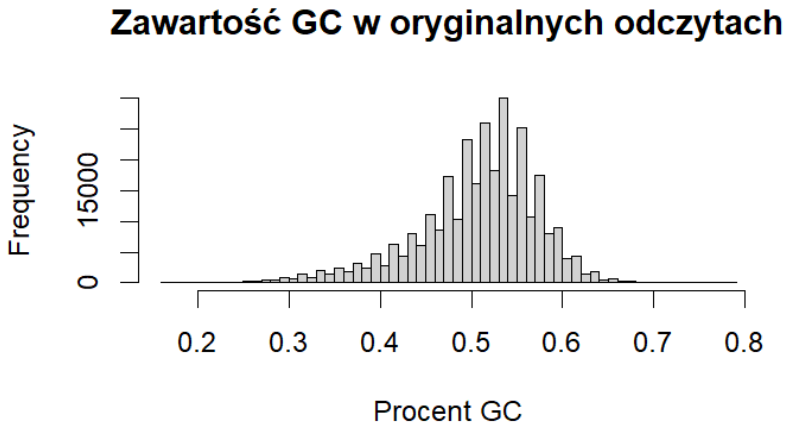
RAPORT – KOLOKWIUM 1 ABwBG

W poniższym pliku zamieszczone są interpretacje wykresów i cząstkowe wyniki dla poszczególnych analiz wykonanych za pomocą załączonego kodu w pliku Rmd.

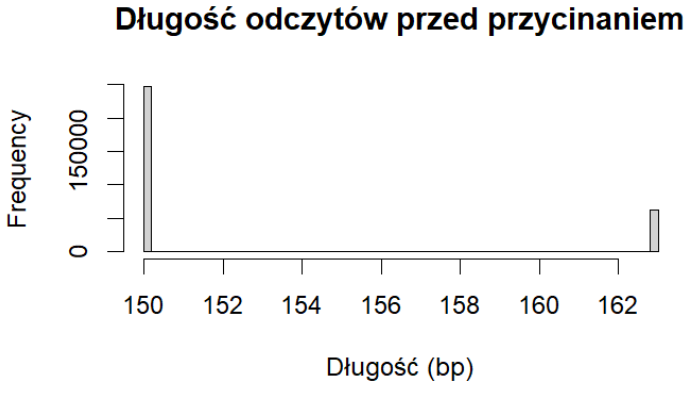
1.Podstawowe dane surowej sekwencji ecoli_raw1

- Liczba wszystkich odczytów: **309440**
- Długość sekwencji w nt: **47223547**
- Zawartość GC dla odczytów jest różnorodna
- Rozkład długości odczytów surowych – występują głównie odczyty o długości 150pz i nielicznie 162 pz

Wykres1. Histogram zawartości GC dla surowych odczytów



Wykres2. Rozkład długości odczytów surowych (bp).

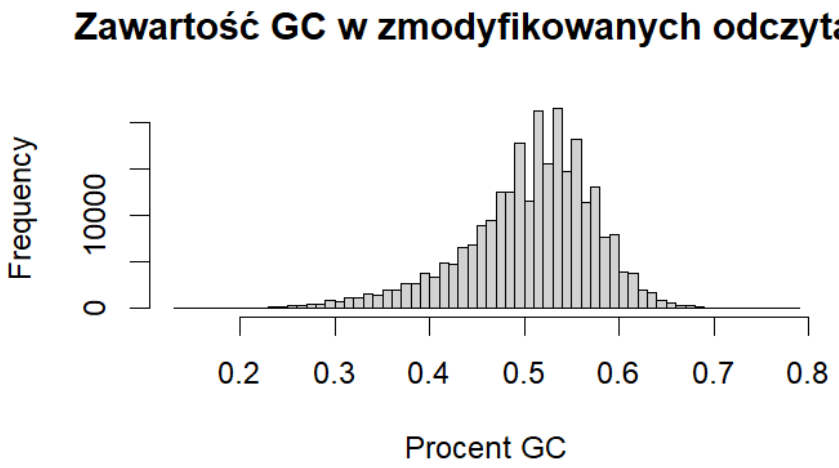


2. Dane sekwencji po trymowaniu i filtrowaniu

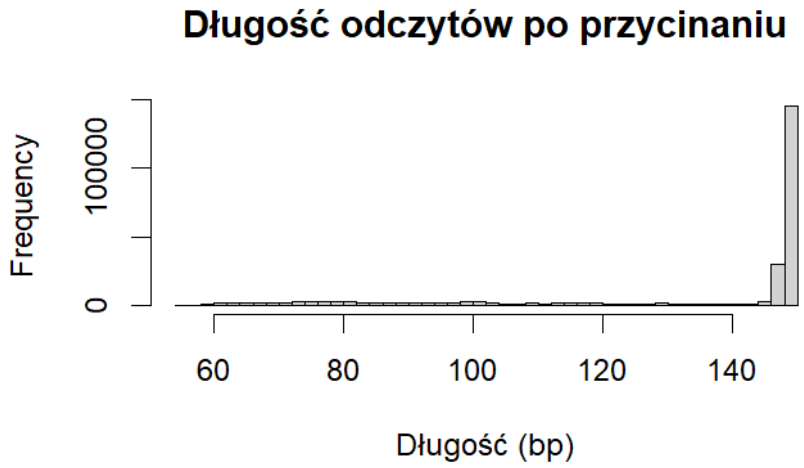
- TRIMMING
    - Przycinanie baz: przycinanie odczytów o niskiej jakości na końcach odczytów - zadane parametry funkcji
    - a ---> Phred+33+ kodowanie - wybierny próg Phred30
    - k = 3 tzn.funkcja będzie szukać co najmniej trzech kolejnych baz o jakości poniżej progu Phred30 aby rozpocząć przycinanie
    - halfwidth ustawiam na szerokość 5 baz - funkcja obliczać będzie średnią wartość jakości w oknie o szerokości 5 baz
    - [1] 309440 – liczba odczytów przed przycinaniem
    - [2] 298001 – liczba odczytów po przycinaniu
    - [3] 193342 – liczba zmodyfikowanych odczytów
    - [4] 96.30332 - % sekwencji, która nie uległa modyfikacji
    - [5] 3.696678 - % sekwencji zmodyfikowanej

- FILTERING
    - Przycinanie całych odczytów o niskiej jakości, przy minmalnej długości odczytu = 60pz
    - [1] 298001- liczba odczytów po przycinaniu baz
    - [2] 266966 – liczba odczytów po filtracji
    - [3] 89.58561 -% odczytów pozostałych po filtracji
    - [4] 10.41439 - podczas filtracji odrzucono 10,41 % odczytów
    - Filtracja i trymowanie spowodowało, że rozkład zawartości GC w odczytach jest bardziej jednorodny – jednocześnie ich udział w sekwencji zwiększył się
    - Rozkład długości odczytów po przycinaniu i filtracji – uległ zmianie, zostały przycięte bazy i odczyty o niższej jakości co spowodowało wzrost frekwencji krótszych odczytów od 60 do 150 pz, przy czym odczyty w zakresie 60-140 mają jednorodną frekwencję, natomiast nadal dominują odczyty o długości w okolicach 150 pz.

Wykres.3 Histogram zawartości GC w zmodyfikowanych odczytach



Wykres4. Rozkład długości odczytów po przycinaniu i filtracji.



3. Wykrywanie i usuwanie sekwencji adapterów: na podstawie analizy można stwierdzić, że sekwencje adapterów "AGATCGGAAGAGC" nie występowały w ogóle w sekwencji, prawdopodobnie zostały skutecznie usunięte, jeśli występowały, na etapie przycinania i filtracji.

- [1] 266966 – liczba odczytów po filtracji
- [2] 266966 – liczba odczytów po usuwaniu adapterów
- [3] 0 - liczba zmodyfikowanych odczytów równa 0 sugeruje, że adapterów nie było.
- Całkowita długość sekwencji w nukleotydach po usuwaniu sekwencji adapterów – 35300224
- Całkowita długość sekwencji surowej po modyfikacjach(przycinanie, filtracja i usuwania adapterów) uległa skróceniu o 11923323 nukleotydy

4. Mapowanie odczytów do genomu referencyjnego

- Total_reads 266966
  Mapped_reads 266918

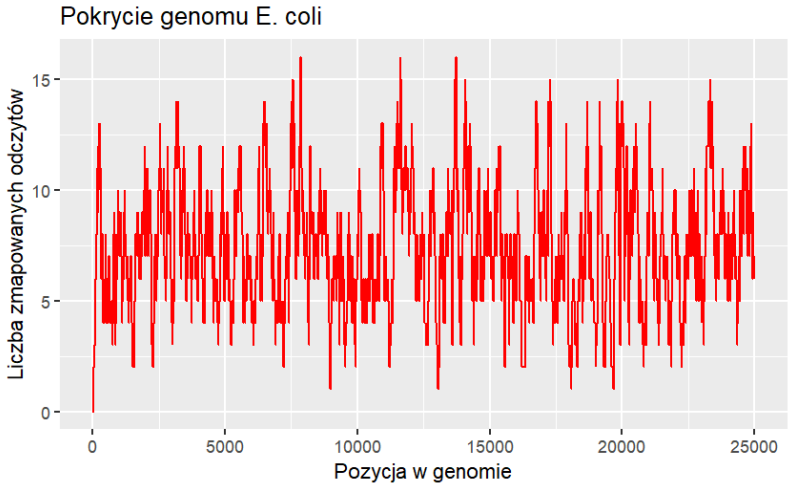
Uniquely_mapped_reads 262013
Multi_mapping_reads 4905
Unmapped_reads 48
Indels 129

- **Procent odczytów zmapowanych: 99,98**
- **Procent odczytów niezmapowanych: 0,018**

Przyczyny niezmapowania odczytów: insercje i delecje obecne w sekwencji (określone jako Indels) lub wystąpienie SNP (polimorfizmów pojedynczych nukleotydów)

Średnie pokrycie genomu: 7.602825 --> pokrycie genomu poniżej 10 mówi nam, że jest to bardzo niskie pokrycie, które może prowadzić do pominięcia rzadkich mutacji i błędów w analizie

Wykres5. Wykres pokrycia genomu

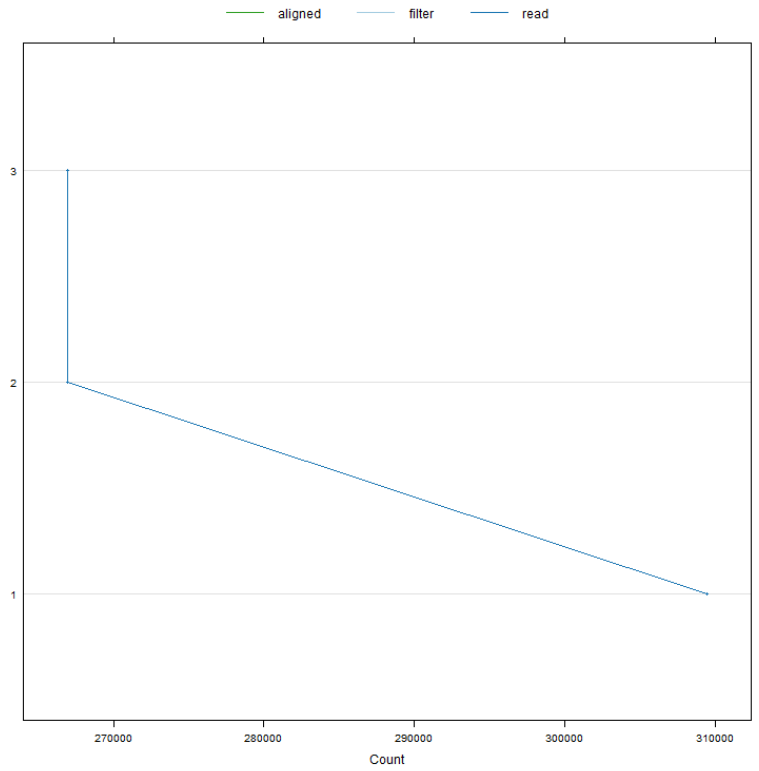


Na podstawie wykresu można zauważyć ze liczbą odczytów o najwyższym pokryciu charakteryzowały się odczyty w rejonie od pozycji ok 7500,12500pz, 1400, 17500, 20000,20350.

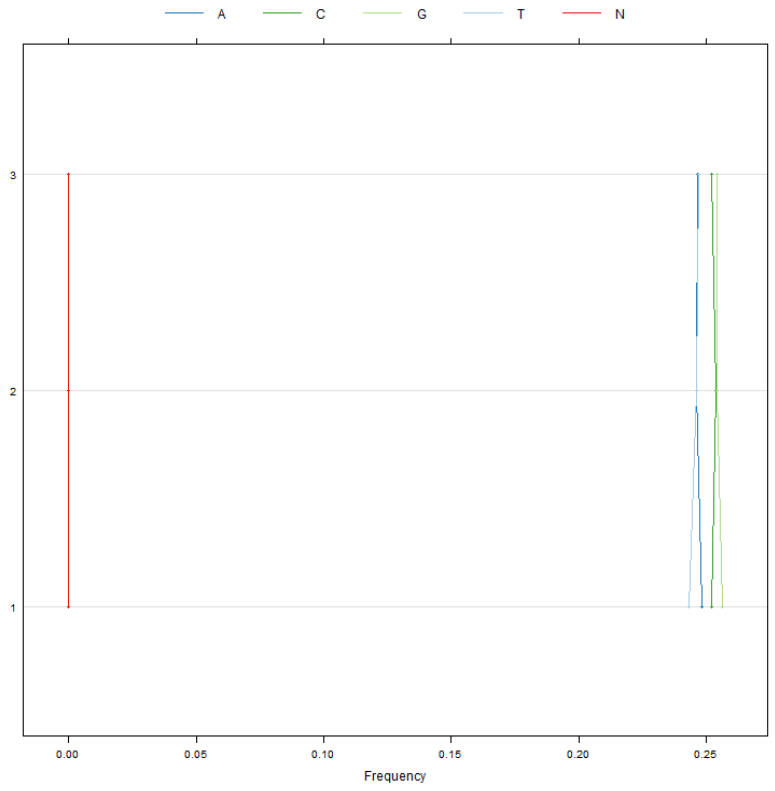
Podsumowanie poprawy jakości odczytów – na podstawie raportu QCmulti dla plików

- **[1]ecoli_raw1** (surowa sekwencja)
- **[2]ecoli_raw2** (sekwencja po trymowaniu i filtrowaniu)
- **[3]ecoli_raw3** (sekwencja po przycięciu adaterów)

1.PlotReadCount – obrazu zmianę długości wszystkich odczytów w trakcie procesowania sekwencji, na wykresie widać, że liczba odczytów wszystkich po trymowani i filtracji nie zmieniła się na skutek usuwania sekwencji adapterów.

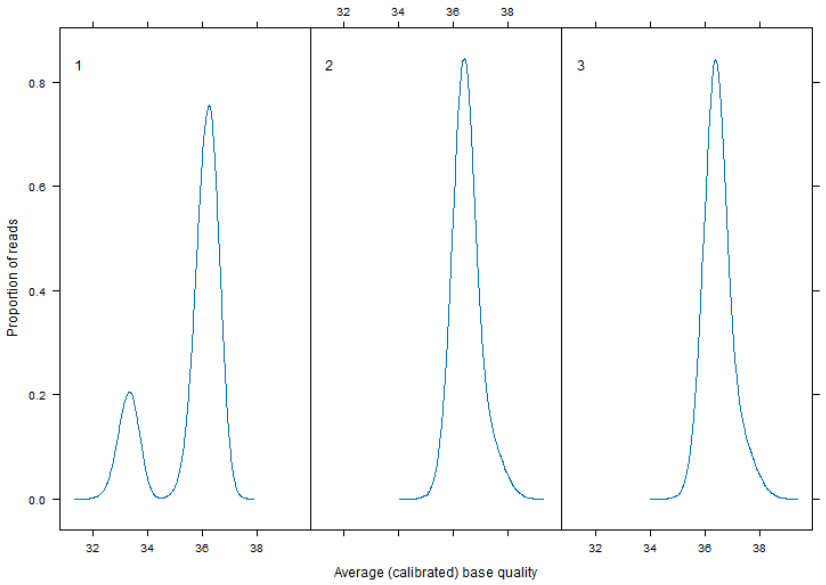


2.Base call frequency over all reads – frekwencja nukleotydów na skutek modyfikacji sekwencji – przycinanie, filtracja i usuwanie sekwencji adapterów różniła się w stosunku do sekwencji surowej, jednak widać, że dla wszystkich trzech sekwencji frekwencja każdej zasady oscylowała w okolicach 0,25%, natomiast N – liczba niedopasowanych nukleotydów nie zmieniała się i wynosiła 0.



3. **Overall read quality –jakość baz w odczytach**

- Dla danych surowych [1] ogólna jakość odczytów jest zmienna, na wykresie widać dwa piki – sugerujące duży udział odczytów o niższej jakości baz w skali Phred, jakość jest niższa na końcach odczytów
- Obróbka sekwencji – przycinanie baz o niższej jakości, filtracja oraz usuwanie sekwencji adapterów (sekwencje [2] i [3]) wskazuje na poprawę jakości, szczególnie w regionach, które wcześniej wykazywały spadki.

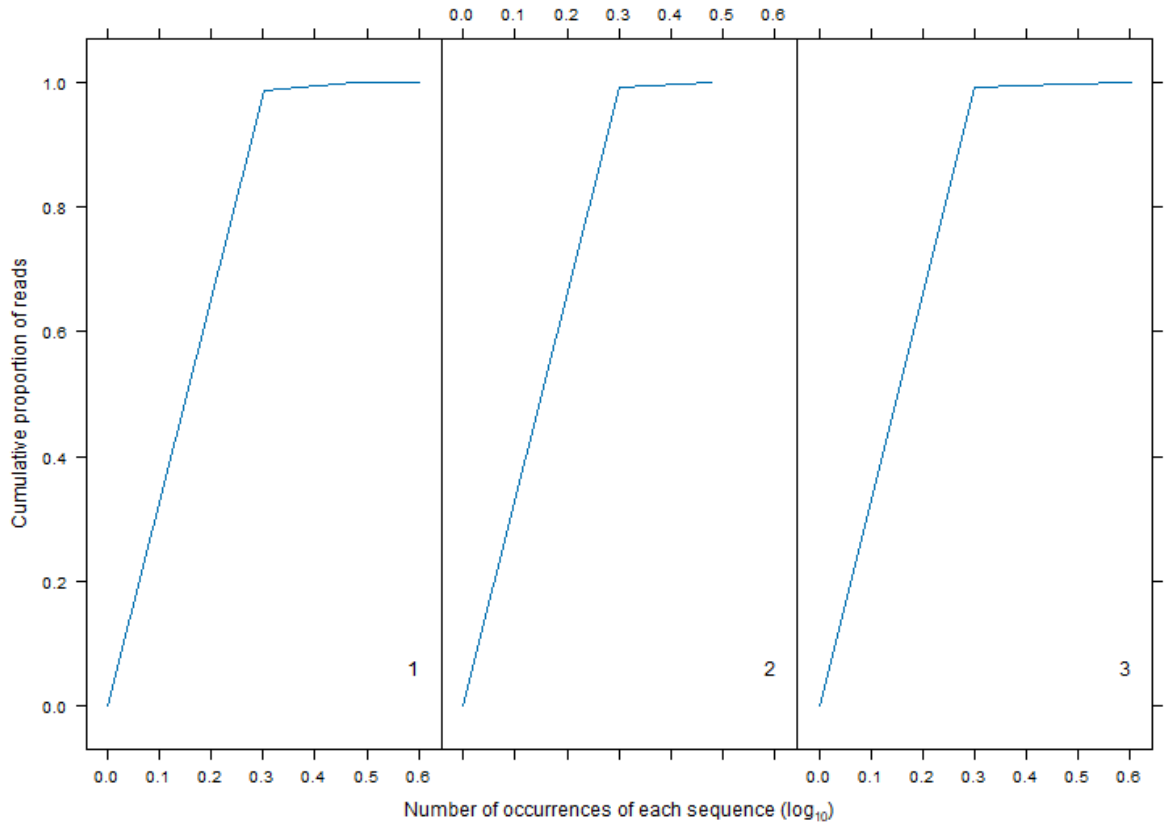


4. plotReadOccurrences

Zarówno w sekwencji [1] ecoli_raw1, [2] ecoli_raw2 i [3] ecoli_raw3:

Krzywa przechodzi gwałtownie od niskiej do wysokiej skumulowanej proporcji odczytów. Sugeruje to, że większość odczytów występuje zbliżoną liczbę razy, co jest charakterystyczne dla danych o równomiernym rozkładzie pokrycia. Wydaje się, że początkowa sekwencja nie miała dużego z problemu z sekwencjami często powtarzającymi się (mogących wynikać z obecności artefaktów, takich jak adaptery, ogonki poli-A, czy powtarzalne regiony w DNA), natomiast modyfikacje sekwencji również znacznie tego nie skorygowały.

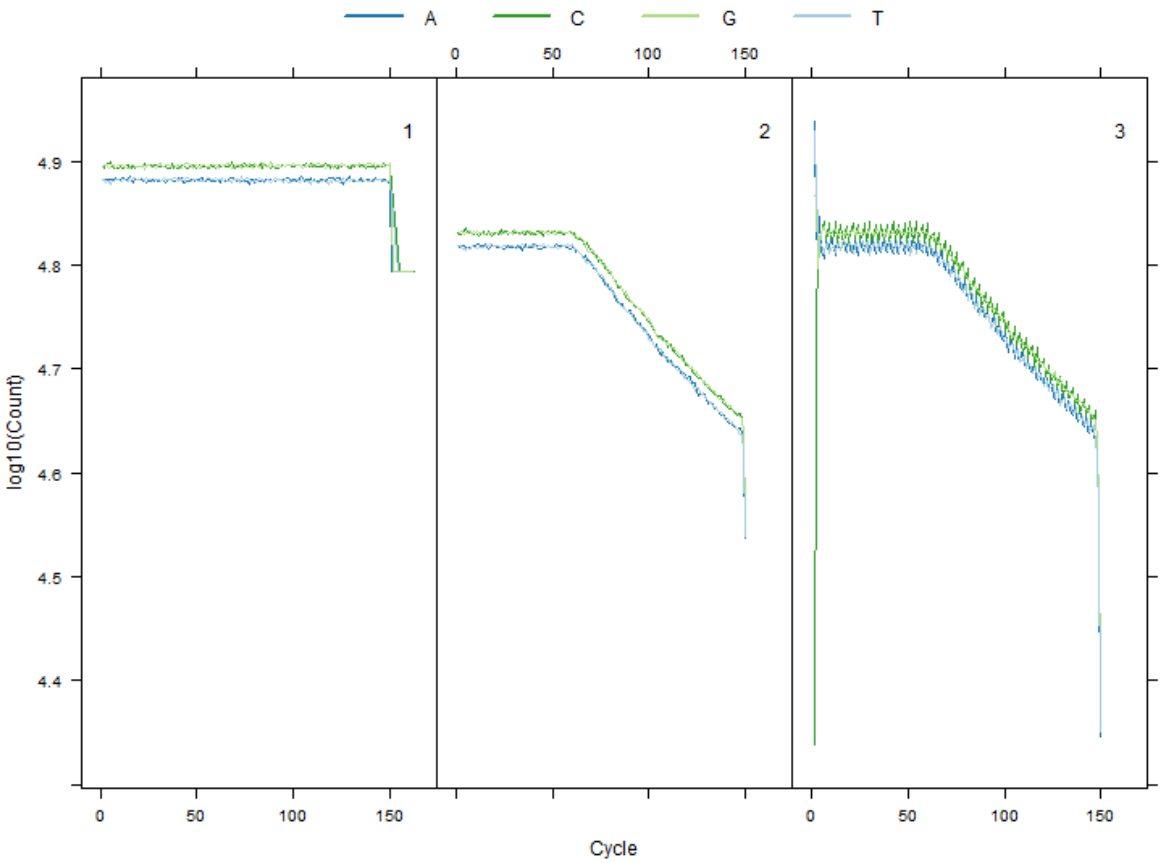
Dla wykresu [2] po trymowaniu i filtracji obserwujemy lekki spadek częstości występowania sekwencji nadreprezentowanych.



5. Per-cycle base call

Analiza jakości baz na cykl pokazuje, jak często poszczególne nukleotydy (A, T, C, G) są wykrywane w każdym cyklu sekwencjonowania. Wyniki te powinny być w przybliżeniu równomierne dla wszystkich cykli, jeśli sekwencjonowanie działa prawidłowo i próbka jest dobrze przygotowana.

Na poniższych wykresach widać, że obróbka danych spowodowała spadek w odczycie wraz z wrzostem liczby cykli, ze względu na interpretację histgramu zawartości GC po modyfikacji, gdzie obserwujemy wzrost udziału zasad GC można stwierdzić, że regiony o wysokiej zawartości GC mogą sprawiać trudności w dokładnym odczycie i doprowadziły do spadku dokładności identyfikacji zasad od 50 cyklu.



Podsumowanie:

Biorąc pod uwagę całą analizę można stwierdzić, że średnie pokrycie genomu po mapowaniu było bardzo niskie, może być to związane z nieodpowiednim doborem parametrów do funkcji wykorzystanej przy trymowaniu sekwencji co wpłynęło na jakość odczytów, która co prawda wzrosła ze względu na **usunięcie sekwencji o niskiej jakości** W przypadku *raw2* i *raw3*, trymowanie i filtrowanie usunęło słabej jakości odczyty, być może zadane parametry funkcji trimmed_reads spowodowały w rezultacie pozostajawienie odczytów z częściowo poprawioną jakością, lub była zbyt mało rygorystyczna i nie przycinała odpowiednio odczytów o niskiej jakości.