Mapowanie do genomu referencyjnego (alignment)

ABwG, ćwiczenia 7 | 19 listopada 2024

Co to jest genom referencyjny?

Standardowa sekwencja DNA danego gatunku, która służy jako punkt odniesienia w badaniach genomowych. Umożliwia:

- mapowanie odczytów sekwencyjnych,
- identyfikację wariantów genetycznych takich jak SNP, indels,
- porównania międzygatunkowe i badania ewolucyjne.

Jak znaleźć odpowiedni genom referencyjny

- Konsultacja z literaturą przedmiotu (Google Scholar)
- Bazy danych genomowych (NCBI, Ensembl, UCSC Genome Browser)
- Kryteria wyboru
  - aktualność – wersje są aktualizowane
  - adnotacja – dostępność informacji o genach i elementach regulacyjnych
  - zgodność z próbkami badawczymi

Gatunek a specyfika genomu referencyjnego

- Gatunki modelowe (*E.coli*, *Drosophila melanogaster*, *Mus musculus*) mają dobrze zbadane i adnotowane genomy
- U gatunków bardzo zmiennych genetycznie (np. człowieka) wymagane jest uwzględnienie wielu polimorfizmów
- Genomy prokariotów są zazwyczaj mniejsze i mają mniej skomplikowaną strukturę niż genomy eukariotów.

Cele mapowania sekwencji

- Lokalizacja odczytów: określienie położenia sekwencji w genomie referencyjnym
- W przypadku danych RNA-seq analiza ekspresji genów
- Wykrywanie wariantów genetycznych: identyfikacja mutacji i polimorfizmów

Metody i algorytmy

- Algorytmy dopasowania dokładnego: wymagają idealnego dopasowania sekwencji
- Algorytmy dopasowania niedokładnego: pozwalają na pewną liczbę niedopasowań
- Metody indeksowania genomu: umożliwiają szybkie wyszukiwanie sekwencji

BAM/SAM: format plików ze zmapowanymi odczytami

- SAM (Sequence Alignment/Map): tekstowy format zapisu dopasowanych odczytów.
- BAM (Binary Alignment/Map): skompresowana, binarna wersja pliku SAM.

Struktura:

- Nagłówek: informacje o sekwencji referencyjnej i parametrach mapowania
- Dane odczytów: Informacje o każdym odczycie, jego pozycji i jakości dopasowania.

Pokrycie genomu (coverage)

- Oznacza ile razy dany fragment genomu został odczytany podczas sekwencjonowania.
- Średnie pokrycie: suma długości wszystkich odczytów podzielona przez długość genomu referencyjnego.

Wpływ na analizę danych genomowych:

- Wysokie pokrycie: zwiększa pewność w wykrywaniu wariantów.
- Niskie pokrycie: może prowadzić do pominięcia rzadkich mutacji lub błędów w analizie.

Mapowanie genomu z pakietami Bioconductor

- Rsubread
- Rsamtools
- GenomicAlignments
- ShortRead
- Biostrings