Task 1

24

https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40

Task 2

1. Single-End (pojedyncze odczyty):

   Podczas sekwencjonowania "single-end" stosuje się tylko jeden zestaw reagentów do odczytu sekwencji. W tym procesie jest sekwencjonowany tylko jeden koniec (fragment) cząsteczki DNA lub RNA. Pliki w formacie "single-end" zawierają wyłącznie pojedyncze odczytane fragmenty sekwencji.

2. Double-End (parowane odczyty):

   Podczas sekwencjonowania "double-end" używa się dwóch zestawów reagentów: jeden do odczytu z jednego końca, a drugi do odczytu z drugiego końca cząsteczki DNA lub RNA. W tym procesie odczytywane są oba końce fragmentów, a nie tylko jeden z nich. Pliki w formacie "double-end" zawierają informacje z dwóch odczytów, które dotyczą tego samego fragmentu sekwencji, ale pochodzą z różnych końców.

Wykorzystanie technologii "double-end" umożliwia uzyskanie większej ilości informacji o sekwencji i zwiększa precyzję analizy danych sekwencyjnych. Pozwala to na bardziej dokładne złożenie sekwencji, identyfikację wariantów oraz przeprowadzanie analiz dotyczących struktury genomu lub transkryptomu. W przypadku dostarczonego pliku w formacie "single-end" mamy jednak do czynienia tylko z jednym odczytem sekwencji dla każdego fragmentu, co ogranicza dostępne informacje w porównaniu do pełnej technologii "double-end".

Task 3

HISEQ

Task 4

@HISEQ1:9:H8962ADXX:1:1101:1297:98785/1

CAAGAAATATGGGACTATGTGAAAAGACCAAATCTACTTCGGATTGGTGTACCTGAAAGTGATGGGGAGAATGG
AAACAAGTTGGAAAACACTCTGCAGGATATTATCCAGGAGAACTTCCCCAATCTAGCACGGCAGGCCAACGTTC

+

@@@DDDDADHFHHIB@;FF3<C@F<+AG>GHGEFEB>G9CF:FFFGD9BBFAGGGEA@)=@@FCC@EGEFBD@
DDECCCC@@A@>@ACCCBB@CC(5@>8<::@CC>AACCCBBBB@CACCC?C?C?34(:A@<5<.>0<?B?BBB<2

Task 5

1. Periodyczne zmiany jakości Dobre dane Illumina mają równomierne i stabilne wartości jakości na przestrzeni odczytów sekwencji. Złe dane Illumina mogą wykazywać periodyczne zmiany jakości, takie jak regularne spadki lub wzrosty jakości w określonych pozycjach sekwencji.

2. Niskie wartości jakości Dobre dane Illumina mają wysokie wartości jakości, zwykle reprezentowane przez symbole ASCII o wartościach wysokich lub bliskich maksymalnej wartości dla danej pozycji sekwencji. Złe dane Illumina mogą mieć niskie wartości jakości, co oznacza słabą jakość odczytu w określonych pozycjach sekwencji. Mogą występować niskie wartości jakości na początku lub końcu sekwencji.

3. Nadmiarowe sekwencje adaptora Dobre dane Illumina mają minimalne ilości sekwencji adaptora, które są używane podczas przygotowania biblioteki do sekwencjonowania. Złe dane Illumina mogą wykazywać nadmiarowe sekwencje adaptora, co sugeruje, że adaptory nie zostały prawidłowo usunięte, a sekwencje adaptora są obecne w odczytach sekwencji.

4. Niewłaściwa sekwencja zasadowa Dobre dane Illumina mają poprawne sekwencje zasadowe, zgodne z wybranym typem sekwencjonowania (np. sekwencjonowanie DNA lub RNA). Złe dane Illumina mogą zawierać niewłaściwe sekwencje zasadowe, co może wskazywać na problemy z przygotowaniem próbki lub sekwencjonowaniem.

Task 6

230282

Task 7

Liczba sekwencji zakwalifikowanych jako złej jakości wynosi 0. Oznacza to, że żadna sekwencja nie została uznana za niskiej jakości na podstawie metryk jakości sekwencji.

Task 8

148

Task 9

Wykres "Per base sequence quality" w raporcie FastQC reprezentuje jakość sekwencji w zależności od pozycji w odczycie. Oś pozioma przedstawia pozycje w sekwencji, a oś pionowa reprezentuje wartości jakości sekwencji. Na wykresie można zauważyć trzy obszary oznaczone kolorem: zielonym, żółtym i czerwonym. Zielony obszar reprezentuje wysoką jakość sekwencji, żółty obszar oznacza jakość umiarkowaną, a czerwony obszar oznacza niską jakość sekwencji.

Task 10

Na podstawie tego wykresu można stwierdzić, że analizowane dane są dobrej jakości. Wszystkie punkty na wykresie znajdują się w zielonym obszarze, co sugeruje wysoką jakość sekwencji.

Task 11

Jeśli error rate wynosi 0,2%, oznacza to, że 0,2% wszystkich odczytów w zestawie danych jest błędnych. Mówiąc inaczej, 0,2 na 100 odczytów może zawierać błędy. Jeśli error rate wynosi 1%, oznacza to, że 1% wszystkich odczytów w zestawie danych jest błędnych. W tym przypadku 1 na 100 odczytów może zawierać błędy. Error rate w sekwencjonowaniu informuje o dokładności procesu sekwencjonowania. Im niższy współczynnik błędu, tym bardziej pewne i wiarygodne są uzyskane sekwencje. Wysoki współczynnik błędu może wynikać z różnych czynników, takich jak błędy w procesie sekwencjonowania, szumy sygnału, czy też problemy z jakością próbki DNA lub RNA.

Task 12

Ilość AT (adenina i tymina) oraz CG (cytozyna i guanina) w sekwencji powinna się zgadzać ze względu na zasady parowania zasad w DNA. Adenina (A) łączy się z tyminą (T) za pomocą dwóch wiązań wodorowych, podczas gdy cytozyna (C) łączy się z guaniną (G) za pomocą trzech wiązań wodorowych. Ta zasada parowania jest znana jako reguła Chargaffa. Jeśli sekwencja DNA jest dwuniciowa i dobrze sparowana, to ilość adeniny (A) będzie równa ilości tyminy (T), a ilość cytozyny (C) będzie równa ilości guaniny (G). Dlatego ilość AT oraz CG powinna się zgadzać, a na wykresie "Per base sequence content" linie odpowiadające tym zasadom powinny być zbliżone.

Task 13

Symbol "N" w sekwencji DNA oznacza nieoznaczoną lub nieznaczącą zasadę azotową. Jest to specjalny symbol używany w sekwencjonowaniu DNA, gdy nie można jednoznacznie określić konkretnej zasady azotowej w danym miejscu sekwencji.

Task 14

Narzędzie "Map with BWA-MEM" służy do mapowania sekwencji odczytów na referencyjny genom za pomocą algorytmu BWA-MEM.

Task 15

Optymalna długość odczytów dla użycia narzędzia "Map with BWA-MEM" może zależeć od wielu czynników, takich jak organizm, rodzaj sekwencji, dostępność referencyjnego genomu itp. Jednak ogólnie rzecz biorąc, BWA-MEM dobrze radzi sobie z odczytami o długościach z zakresu 70-100 par zasad (bp).

Task 16

230552 odczyty przeszły kontrolę jakości (quality control).

Task 17

Z linijki nr 7 odczytujemy, że 99.99% sekwencji zostało zmapowanych (mapped) do genomu.

Task 18

Mapowanie sekwencji do genomu oznacza proces dopasowania krótkich sekwencji DNA (nazywanych odczytami) do referencyjnego genomu, czyli kompletnego zestawu genetycznego danego organizmu.

Task 19

Większość odczytów mapowanych jest do chromosomu "chr10".

Task 20

NC_000010

Task 21

CYP2C18, CYP2C19

Task 22

Genotyp osoby może być GG, GA lub AA.

Task 23

Jeśli osoba ma genotyp GG, może to oznaczać, że osoba ta jest mniej podatna na działanie klopidogrelu. W przypadku genotypu GA, reakcja na klopidogrel może być zmienne, a skuteczność leczenia może być mniejsza niż u osób z genotypem GG. Natomiast w przypadku genotypu AA, osoba ta może być bardziej podatna na działanie klopidogrelu, co może wiązać się z większym ryzykiem działań niepożądanych leku.

Task 24

Narzędzie "bcftools mpileup" służy do generowania plików VCF lub BCF poprzez przeprowadzenie wielopozycyjnego wywołania wariantów dla zbioru odczytów sekwencji DNA przyporządkowanych do referencyjnego genomu.

Task 25

chr10 510710 . T 0 . DP=1;I16=0,1,0,0,29,841,0,0,2,4,0,0,2,4,0,0;QS=1,0;FS=0;MQ0F=0 PL 0,3,4

Task 26

- bcftools mpileupon data - ~600,000 lines
- bcftools call - 308 lines

Task 27

Prawdopodobną przyczyną różnicy w wynikach analizy może być zastosowanie różnych genomów referencyjnych. Parametr "Select reference genome" podczas generowania narzędzia "bcftools mpileup" wskazuje, że użyto genomu referencyjnego dla człowieka (Homo sapiens) o oznaczeniu hg19. Skoro użyto różnych genomów referencyjnych, może to prowadzić do różnic w wynikach analizy, w tym różnicy w liczbie linii w pliku wyjściowym narzędzia "bcftools mpileup".

Task 28

0/1:255,0,255 C/G

0/1:255,0,120 atttttttttttttt/aTTTTtttttttttttttt

0/1:255,0,255 G/A

Task 29

Parametr "sequence depth" odnosi się do liczby odczytów obejmujących daną pozycję w genomie. W kontekście sekwencjonowania DNA lub RNA, głębokość sekwencjonowania odnosi się do liczby razy, jakie dana sekwencja została przeczytana podczas procesu sekwencjonowania. Wyższa wartość parametru DP wskazuje na większą ilość odczytów dla danej pozycji, co może wskazywać na większą pewność i precyzję w wykrywaniu wariantów genetycznych lub mutacji na tej pozycji.

Task 30

Narzędzie o nazwie "bcftools counts" służy do analizy wariantów genetycznych obecnych w plikach VCF (Variant Call Format).

Task 31

#samples	SNPs	INDELs	MNPs	others	sites
1	262	46	0	0	308

Task 32

- SNP (Single Nucleotide Polymorphism) to pojedyncza zmiana nukleotydu w sekwencji DNA między osobnikami tego samego gatunku.
- INDELs (Insertions and Deletions) to mutacje polegające na wstawieniu lub usunięciu nukleotydów w sekwencji DNA.
- MNP (Multiple Nucleotide Polymorphism) to polimorfizm genetyczny, w którym występuje jednoczesna zmiana więcej niż jednego nukleotydu w sekwencji DNA.

Task 33

Narzędzie "VCFfilter" umożliwia użytkownikowi zastosowanie różnych filtrów i kryteriów w celu wyodrębnienia interesujących wariantów z pliku VCF. Można go użyć do odfiltrowania wariantów o określonym poziomie jakości, pokryciu sekwencji, allelicznym stosunku i innych parametrach

Task 34

"QUAL > 200" oznacza, że jakość przekracza wartość 200, co sugeruje wysoką pewność co do poprawności tego zdarzenia lub sekwencji.

Task 35

Zostało wyświetliono 249 lines z 308 lines, to 80,8%

Task 36

0/1:255,0,255 GA