TASK 1

Ma rozmiar: 249280735 bp Link: Homo sapiens isolate NA12878 chromosome 1, whole genome shotgun sequen - Nucleotide - NCBI (nih.gov)

TASK 2

Sekwencjonowanie "double-end" polega na sekwencjonowaniu obydwu końców DNA, a "single-end" na tylko jednym końcu.

"Single-end" jest prostsze i tańsze, ale mniej dokładne niż "double-end", które zaweira informacji o długości i orientacji fragmentów DNA względem siebie.

TASK 3

HISEQ

TASK 4

1. @HISEQ1:9:H8962ADXX:1:1101:1297:98785/1
2. CAAGAAATATGGGACTATGTGAAAAGACCAAATCTACTTCGGATTGGTGTACCTGAAAGTGATGGGGAGAATGGAAACAAGTTGGAAAACACTCTGCAGGATATTATCCAGGAGAACTTCCCCAATCTAGCACGGCAGGCCAACGTTC
3. +
4. @@@DDDDADHFHHIB@;FF3GHGEFEB>G9CF:FFFGD9BBFAGGGEA@)=@@FCC@EGEFBD@DDECCCC@@A@>@ACCCBB@CC(5@>8<::@CC>AACCCBBBB@CACCC?

   Dodatkowo w podglądzie był jeszcze ciąg znaków: C?C?34(:A@<5<.>0.

TASK 5

Dane dobrej jakości (Good Illumina Data) mają wysoką jakość sekwencjonowania i niski poziom błędów.

Dane niskiej jakości (Bad Illumina Data) mają niską jakość sekwencjonowania i wysoki poziom błędów.

Większość sekwencerów generuje raport QC, jednak jest on zazwyczaj skupiony na identyfikowaniu problemów wygenerowanych przez sam sekwencer.

TASK 6

Total Sequences: 230 282

TASK 7

Nie

TASK 8

Sequence length: 148

TASK 9

Są to dane Illumna - graficzną reprezentację statystyk jakości na przestrzeni wszystkich baz.

TASK 10

Bazując na dokumentacji - **tak**, dane są dobrej jakości.

TASK 11

Error rate wynosi 0.2% w sytuacji, gdy obserwowana średnia jakość jest poniżej 27.

Error rate wynosi 1%, gdy błąd jest zgłaszany oraz jeśli najczęściej obserwowana średnia jakość jest poniżej 20.

W sekwencjonowaniu error rate oznacza procent baz wywołanych niepoprawnie w dowolnym cyklu. Można zaobserwować jego wzrost wraz z długością odczytu.

TASK 12

Ilość AT oraz CG powinna się zgadzać, ponieważ DNA składa się z par zasad azotowych:

- adeniny (A) łączącej się z tyminą (T)
- cytozyny © łączącej się z guaniną (G).

W każdej parze zasad azotowych ilość A powinna być równa ilości T, a ilość C równa ilości G. Dlatego też, jeśli ilość AT oraz CG nie zgadza się, może to wskazywać na obecność błędów w sekwencjonowaniu.

TASK 13

N w sekwencji DNA pochodzącej z sekwencjonowania oznacza, że sekwencer nie był w stanie dokonać poprawnego wywołania "basee call" zasady azotowej w danym miejscu i zastąpił ją literą N.

Wartość Per base N content pokazuje procentowy udział N w każdym cyklu sekwencjonowania

TASK 14

To narzędzie służy do mapowania średnich i długich odczytów względem genomu referencyjnego.

TASK 15

Odczyty o długości większej niż 100 bp

TASK 16

230552 odczytów

TASK 17

99.99%

TASK 18

Mapowanie sekwencji DNA do genomu referencyjnego polega na przyrównywaniu odczytów sekwencjonowanych fragmentów DNA do sekwencji wzorcowej, czyli takiej która jest reprezentatywna dla danego gatunku. Mapowanie pozwala naokreślenie umiejscowienia w genomie, z którego pochodzi dany fragment DNA.

TASK 19

chr10

TASK 20

99871

TASK 21

CYP2C18 i CYP2C19

TASK 22

Oznacza to, że analizowana osoba ma genotyp GA, czyli posiada jedną kopię allelu z "G" jedną z "A" dla tej pozycji.

TASK 23

Osoby z tym genotypem moga mieć zmiejszoną skuteczność leku Clopidogrel i zwiększone ryzyko wystąpienia powikłań sercowo-naczyniowych.

TASK 24

Służy do wykrywania wariantów genetycznych w próbkach sekwencjonowanych.

Generuje plik mpileup, który ma informacje o pokryciu sekwencji i zidentyfikowanych wariantach genetycznych.

TASK 25

##contig=<ID=chr10,length=135534747>

TASK 26

~600,000 linii

TASK 27

Różnica ta wynika z faktu, że narzędzie bcftools mpileup służy do szacowania prawdopodobieństwa genotypów w każdej pozycji genomu na podstawie danych sekwencjonowania. Natomiast narzędzie bcftools call identyfikuje zarówno warianty, jak i genotypy, czyli dokonuje właściwego wywołania.

bcftools mpileup zawiera informacje o pokryciu sekwencji i prawdopodobieństwie genotypów w każdej pozycji genomu. Natomiast plik bcftools call zawiera informacje o zidentyfikowanych wariantach genetycznych i przypisanych im genotypach

TASK 28

idąc od 1 jako POS 96400844

1. AA, ponieważ jest 1/1 i alt to A
2. CG, ponieważ jest 0/1 i ref: C, alt: G
3. TT, ponieważ jest 0/1 i z `attttttttttttt` mamy T, oraz z `aTTTTttttttttttttt`, tak samo T dominuje

TASK 29

Jest to głębokość czytania sekwencji. Czyli np jak 23 "szare paski" - odczyty nachodzą na siebie, będzie to głębokość DP = 23. W rekordzie 1 DP = 196, czyli 196 odczytów sekwencji (sequence reads) nachodzi na siebie w tym miejscu.

TASK 30

bcftools counts — to narzędzie do zliczania liczby próbek, SNP-ów, INDEL-ów, MNPs i całkowitej liczby miejsc w pliku VCF.

TASK 31

Po kolei: SNPs, INDELs, MNPs i inne.

TASK 32

SNP to Single Nucleotide Polymorphism, INDEL to Insertion/Deletion, a MNP to Multiple Nucleotide Polymorphism.

TASK 33

VCFfilter to narzędzie do filtrowania wariantów w pliku VCF, pozwala na wybranie różnych atrybutów takich jak jakość wariantu, głębokość pokrycia sekwencji czy liczba próbek z danym wariantem.

TASK 34

QUAL > 200 oznacza, że zostaną wybrane tylko te warianty, których jakość (QUAL) jest większa niż 200.

TASK 35

$$\frac{249}{308} \approx 0.80 \text{ zmienności}$$

Czyli daje nam to 80% spełniających zadane przeze mnie warunki.

TASK 36

Mamy Ref: G i Alt: A, 0/1, czyli **GA**.