Bioinformatics - Lab 7-8

Batuhan Seyhan

25.04.2023

TASK 1

The size of this chromosome is approximately 135 million base pairs. Referenced database is International Genome Sample Resource : https://www.internationalgenome.org/data-portal/sample/NA12878

TASK 2

The primary difference between single-end and paired-end sequencing is the amount of information obtained from each DNA fragment. Single-end sequencing provides sequence data from only one end, while paired-end sequencing provides sequence data from both ends. Paired-end sequencing allows for better accuracy in aligning the reads to the reference genome, detecting structural variations, and improving assembly and mapping of the sequenced fragments.

TASK 3

HISEQ1

TASK 4

```
@@@DDDDADHFHHIB@;FF3<C@F<+AG>GHGEFEB>G9CF:FFFGD9BBFAGGGEA@)=@@FCC@EGEFBD@DDECCCC@@A@>@ACCC
@HISEQ1:9:H8962ADXX:1:1101:1521:25554/1
GTTTGTTGTTTTCATTTTCCCATACTCCTAGAGTACTTGCCAAGGTAGCTCTGGTGAGTGGCTGGAAAGGGGTGTTGGGAGCAAACTGAG
+
```

TASK 5

Good Quality (Good Illumina Data):

Base Call Accuracy: Good quality data typically exhibit high base call accuracy, meaning the probability of correctly calling the nucleotide at each position is high. Per Base Sequence Quality: The quality scores associated with each base in the sequencing read are high and consistent. Good quality data show high-quality scores across the entire length of the read. Sequence Length Distribution: The majority of sequences in the dataset have uniform lengths, indicating successful sequencing and minimal variability. Adapter Content: The presence of adapter sequences can indicate incomplete library preparation or contamination. In good quality data, adapter content

is low or absent. Overrepresented Sequences: Good quality data usually show a low presence of overrepresented sequences, which could indicate contamination or biased amplification. GC Content Distribution: The distribution of GC content in the data is generally even, indicating a lack of bias or specific amplification preferences.

Bad Quality (Bad Illumina Data):

Low Base Call Accuracy: Low-quality data may exhibit a high number of sequencing errors, resulting in a decreased accuracy of base calls. Per Base Sequence Quality: The quality scores associated with the bases may drop below a certain threshold at certain positions in the read, indicating regions of lower confidence or potentially problematic sequencing. Sequence Length Distribution: Poor-quality data may have a wide range of sequence lengths, suggesting issues with library preparation or variability in the data. Adapter Content: High levels of adapter content suggest incomplete removal of adapter sequences during library preparation, potentially impacting downstream analysis. Overrepresented Sequences: The presence of numerous overrepresented sequences may indicate contamination, adapter artifacts, or other biases that could affect the accuracy of downstream analysis. GC Content Distribution: Significant deviations in GC content distribution may suggest biases in the sequencing process or amplification issues.

TASK 6

230282 sequences

TASK 7

There is no sequence flagged as poor quality

TASK 8

Length of sequences: 148

TASK 9

The Per Base Sequence Quality graph in FastQC provides information about the quality scores associated with each base position in the sequencing reads. The quality scores represent the confidence in the accuracy of the base call at a given position.

In the graph, the x-axis represents the position in the read, starting from the 5' end (beginning) of the sequence to the 3' end. The y-axis represents the quality scores, typically given in Phred scores. Higher quality scores indicate higher confidence in the base call accuracy.

The Per Base Sequence Quality graph gives you an overview of the quality distribution across the entire length of the reads. By analyzing this graph, you can assess potential issues with sequencing quality, such as poor signal, errors, or biases.

TASK 10

Based on the graph the analyzed data has an average of very good quality across almost all of the bases.

TASK 11

The error rate in sequencing refers to the rate at which sequencing errors occur during the process. These errors can arise from a variety of sources, such as limitations in the sequencing technology, sample preparation, or other experimental factors.

Error Rate of 0.2%: A low error rate of 0.2% means that, on average, only 0.2% of the bases in the sequencing reads are incorrectly called or have sequencing errors. This indicates high sequencing accuracy and reliability. In such a scenario, the vast majority of the sequencing data is likely to be of high quality, with minimal errors affecting downstream analysis.

Error Rate of 1%: A higher error rate of 1% suggests that, on average, 1% of the bases in the sequencing reads contain errors. While this error rate is still relatively low, it indicates a slightly lower quality and accuracy compared to the 0.2% error rate. In this situation, it is important to carefully evaluate the specific error types and their impact on downstream analysis to ensure accurate interpretation of the data.

TASK 12

In an ideal scenario, the amount of A and T nucleotides (which form a complementary base pair) and the amount of C and G nucleotides (which also form a complementary base pair) should match. This is because in a DNA molecule, A pairs with T, and C pairs with G, in a complementary manner. This complementary base pairing is a fundamental principle in DNA structure and replication.

If the amount of A and T nucleotides or the amount of C and G nucleotides is not balanced, it may indicate potential issues or biases in the sequencing data.

TASK 13

In DNA sequencing, the letter "N" represents an ambiguous or unknown nucleotide base at a specific position in the DNA sequence.

When the sequencing technology cannot confidently identify the exact nucleotide at a specific position, it is denoted by the letter "N". The letter "N" is used as a placeholder to indicate that the nucleotide base at that position is unknown or cannot be determined with certainty.

TASK 14

This tool is a widely used algorithm for mapping next-generation sequencing (NGS) reads to a reference genome. It is particularly well-suited for aligning longer reads, such as those generated by Illumina sequencing technology.

TASK 15

Medium and long reads: larger than 100 bp

TASK 16

230552 reads passed quality control.

TASK 17

99.99%

TASK 18

Mapping sequences to a genome refers to the process of aligning or matching DNA or RNA sequencing reads to a known reference genome. The reference genome represents a complete or representative sequence of a specific organism's genome, often obtained through extensive sequencing and assembly efforts.

TASK 19

Chromosome: SN:chr10
Reads: 135534747

TASK 20

Accession number: NG_008384.3

TASK 21

CYP2C18 and CYP2C19

TASK 22

The genotype of an individual refers to the combination of alleles they possess at a specific genomic locus. In this case, the two possible alleles at that position are "G" and "A". If an individual is homozygous for the "G" allele, all the reads would be expected to align to the "G" position. Similarly, if an individual is homozygous for the "A" allele, all the reads would align to the "A" position.

TASK 23

Molecular consequence: NM_000769.4:c.681G>A - synonymous variant Functional consequence: cryptic splice acceptor activation [PubMedVariation Ontology: 0375]; protein loss of function [Variation Ontology: 0043]

TASK 24

It is used for analyzing genomic variation from next-generation sequencing data, specifically from aligned sequencing reads in BAM format.

TASK 25

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

TASK 26

bcftools mpileup: around 600000 lines bcftools call: 308 lines

TASK 27

The difference in the number of generated lines between bcftools mpileup and bcftools call is primarily due to the fact that bcftools mpileup provides coverage and per-base information for all positions in the reference genome, whereas bcftools call focuses on reporting variant calls and includes only the positions with identified variants.

TASK 28

##contig=<ID=chr6_apd_hap1,length=4622290> ##contig=<ID=chr6_cox_hap2,length=4795371> ##contig=<ID=chr6_dbb_hap3,length=4610396>

TASK 29

DP stands for "Read Depth". It represents the total number of reads (sequencing reads) that were aligned to the genomic position considered in the variant calling process.

TASK 30

It is a tool which is commonly used in bioinformatics for analyzing genetic variation from next-generation sequencing data. This tool provides additional functionality to perform counting and summary operations on variant call format (VCF) files.

TASK 31

Total 308 including SNPs, INDELs and others.

TASK 32

SNP: Single Nucleotide Polymorphism - a genetic variation where a single nucleotide at a specific position in the genome differs among individuals. INDELs: Insertion and Deletion - genetic variations caused by the insertion or deletion of nucleotides in the genome, resulting in a change in the length of the DNA sequence. MNPs: Multiple Nucleotide Polymorphisms - genetic variations where multiple adjacent nucleotides at a specific position in the genome differ among individuals.

TASK 33

It is tool used for filtering and manipulating Variant Call Format (VCF) files, which are commonly used for storing genomic variants identified from DNA sequencing data.

TASK 34

By specifying QUAL>200, we are setting a threshold for the variant quality score. This means that we only want to retain variants in the VCF file that have a QUAL value greater than 200. Variants with a lower quality score will be filtered out or excluded from the output.

TASK 35

After filtering we obtain 249 lines out of 308 lines which were retrieved by the bcftools call previously. Therefore the percentage of volatility is approximately 19.1%

TASK 36

Genotype: C