STUDENT'S NAME: LUTAAYA FESTUS DAVID

REGISTRATION NUMBER: 2022/HD07/2046U

STUDENT NUMBER: 2200702046

COURSE UNIT: BASH

INSTRUCTOR: Mr. LUJUMBA IBRA

MANIPULATING VCF FILES

1. Variant Call Format file is a text-based file that stores information about genetic variations in a population. It is a commonly used file format in the field of genetics to share and store data about variations in the genome. This file contains specific details such as the location of the variation, the reference allele, and the alternate allele, as well as information about the accuracy of the variant calls.

2. The header section of a VCF file contains metadata about the file, such as:The version of the file format, reference genome used, information about the samples in the file, information about the tools and parameters used to generate the variants.

3. bcftools query -l sample.vcf | wc -l

They are 6

4. bcftools view -H sample.vcf | wc -l

Answer: 398246

5.bcftools query -f '%CHROM\t%POS[\t%QD;%MQ]\n' sample.vcf > vcffile1.txt

6. awk '$1=="2" || $1=="4" || $1=="MT"' sample.vcf > vcffile2.vcf

7. awk '$1 != "20" || ($1 == "chr20" && ($2 < 1 || $2 > 30000000)) \

{print $1, $2, $4, $5}' sample.vcf > vcffile3.vcf

8. bcftools query -f '%CHROM\t%POS\t%REF\t%ALT\n' -s SRR13107019 sample.vcf > vcffile4.txt

9. bcftools filter -i 'INFO/QD>7' sample.vcf > vcffile5.vcf

10. bcftools view -h sample.vcf | grep -o -w 'contig=[^;]*' | sort | uniq | wc -l

Answer: 2211

11. The two columns depict the format of the genotype data for each sample. The eighth column refers to the identifier for the sample and the ninth column refers to the genotype data for that sample.

12. bcftools query -f '%DP\n' -s SRR13107018 sample.vcf > vcffile6.vcf

13. bcftools query -f '%CHROM\t%POS\t%AF\n' sample.vcf > vcffile7.vcf

MANIPULATING SAM FILES

1. The structure of a SAM file is a text-based format that holds information about how short DNA sequences align to a reference genome. It includes details about the position of the read, the quality of the alignment.

2. The header section of a SAM file contains metadata about the file, such as the version of the file format, the reference genome used, and information about the samples in the file, as well as information about the sequencing run like the instrument and run parameters, and any program specific options used. It is also used to store information about the reference sequences used in the alignment, read groups, and program options. This information is stored in a series of lines starting with "@" symbol.be included.

3. grep '^@RG' -c sample.sam | cut -f2

Answer: 249

4. grep -v '^@' sample.sam | wc -l

Answer: 36142

5. samtools flagstat sample.sam > samfile1.txt

6. head -n1 sample.sam | tr '\t' '\n' | wc -l

Answer: 4

7. grep '^@SQ.*NT_' sample.sam

8. grep '^@RG.*LB:Solexa' sample.sam

9. awk '$1 !~ /^@/ && $2 == "99" || $2 == "83"' sample.sam > samfile2.sam

10. awk '$1 !~ /^@/ && ($3 == "1" || $3 == "3")' sample.sam | samtools view -Sb - > samfile3.bam

11. samtools view -f 4 sample.sam > samfile4.sam

12. grep -c "^4\t" sample.sam

Answer: 0

13. The read name and read flag are included in the SAM file, with the second column indicating the name of the read and the sixth column indicating an integer value that provides information about the alignment of the read, including if it is mapped or unmapped.

14. awk '{ for (i=11; i<=NF; i++) print $i }' sample.sam > optional_fields.txt