1 Data formats for NGS data - Answers

1. There are 10 sequences in this file. To count all the header lines, we can use `grep -c ">" data/example.fasta`

2. There are 8 reads in this file. We can use `grep` to search for /1 or /2:

```
grep -c "/1" data/example.fastq
```

Alternatively, we can use `wc -l` to count the lines in the file and then divide this by 4.

3. RG = Read Group

4. Illumina. See the __PL__field.

5. SC. See the **CN** field.

6. ERR003612. See the **ID** field.

7. 2kbp. See the **PI** field.

8. The quality is 48. We can use `grep` to find the id, followed by `awk` to print the fifth column:

```
grep "ERR003762.5016205" data/example.sam | awk '{print $5}'
```

9. The CIGAR is 37M. We can use `grep` and `awk` to find it:

```
grep ERR003814.6979522 data/example.sam | awk '{print $6}'
```

10. 213. The ninth column holds the insert size, so we can use awk to get this:

```
grep ERR003814.1408899 data/example.sam | awk '{print $9}'
```

11. The CIGAR in **Q9** was 37M, meaning all 37 bases in the read are either matches or mismatches to the reference.

12. CIGAR: 4M 4I 8M. The first four bases in the read are the same as in the reference, so we can represent these as 4M in the CIGAR string. Next comes 4 insertions, represented by 4I, followed by 8 alignment matches, represented by 8M.

13. NCBI build v37

14. There are 15 lanes in the file. We can count the @RG lines manually, or use standard UNIX commands such as:

```
samtools view -H data/NA20538.bam | grep ^@RG | wc -l
```

or

```
samtools view -H data/NA20538.bam | awk '{if($1=="@RG")n++}END{print n}'
```

15. Looking at the @PG records ID tags, we see that three programs were used: GATK IndelRealigner, GATK TableRecalibration and bwa.

16. The @PG records contain a the tag **VN**. From this we see that bwa version 0.5.5 was used.

17. The first collumn holds the name of the read: ERR003814.1408899

18. Chromosome 1, position 19999970. Column three contains the name of the reference sequence- and the fourth column holds the leftmost position of the clipped alignment.

19. 320 reads are mapped to this region. We have already sorted and indexed the BAM file, so now we can search for the reagion using **samtools view**. Then we can pipe the output to **wc** to count the number of reads in this region:

```
samtools view data/NA20538_sorted.bam 1:20025000-20030000 | wc -l
```

20. The reference version is 37. In the same way that we can use `-h` in samtools to include the header in the output, we can also use this with bcftools:

```
bcftools view -h data/1kg.bcf | grep "##reference"
```

21. There are 50 samples in the file. The `-l` option will list all samples in the file:

```
bcftools query -l data/1kg.bcf | wc -l
```

22. The genotype is A/T. With `-f` we specify the format of the output, `-r` is used to specify the region we are looking for, and with `-s` we select the sample.

```
bcftools query -f'%POS [ %TGT]\n' -r 20:24019472 -s HG00107 data/1kg.bcf
```

23. There are 4778 positions with more than 10 alternate alleles. We can use `-i` to specify that we are looking for instances where the value of the INFO:AC tag (Allele Count) is greater than 10:

```
bcftools query -f'%POS\n' -i 'AC[0]>10' data/1kg.bcf | wc -l
```

24. There are 451 such positions. The first command picks out sample HG00107. We can then pipe the output to the second command to filter by depth and non-reference genotype. Then use `wc` to count the lines:

```
bcftools view -s HG00107 data/1kg.bcf | bcftools query -i'FMT/DP>10 & FMT/GT!="0/0"' -
f'%POS[ %GT %DP]\n' | wc -l
```

25. 26. The first base is at position 9923 and the last is at 9948.

26. G. To reduce file size, only the first base is provided in the REF field.

27. 10. See the MinDP tag in the INFO field.

2 QC assessment of NGS data

1. The peak is at 140 bp, and the read length is 100 bp. This means that the forward and reverse reads overlap with 60 bp.

2. There are 400252 reads in total.

Look inside the file and locate the field "raw total sequences". To extract the information quickly from multiple files, commands similar to the following can be used:

```
grep ^SN lane*.sorted.bam.bchk | awk -F'\t' '$2=="raw total sequences:"'
```

3. 76% of the reads were mapped. Divide "reads mapped" (303036) by "raw total sequences" (400252).

4. 2235 pairs mapped to a different chromosome. Look for "pairs on different chromosomes"

5. The mean insert size is 275.9 and the standard deviation is 47.7. Look for "insert size mean" and "insert size standard deviation".

6. 282478 reads were properly paired. Look for "reads properly paired".

7. 23,803 (7.9%) of the reads have zero mapping quality. Look for "zero MQ" in the "Reads" section.

8. The forward reads. Look at the "Quality per cycle" graphs.

3 File conversion - Answers

1. The CRAM file is ~18 MB. We can check this using:

```
ls -lh data/yeast.cram
```

2. Yes, the BAM file is ~16 MB bigger than the CRAM file. We can check this using:

```
ls -lh data/yeast*
```