1. Which of the following strings cannot denote a DNA sequence:

   GTTACGCGAT

   GGGGGGGGG

   GTTGAGCCGGCGT

   MASLLRG

2. How many lines does it take to specify:

i) one fasta sequence? and ii) one fastq sequence?

Select the best answer.

   Fasta – 3 lines; fastq – 3 lines

   Fasta – 4 lines; fastq – 4 lines

   Fasta – a fasta header followed by any number of sequence lines; fastq – 4 lines

   Fasta – 1 line; fastq – 4 lines

3. Which of the following is incorrect:

   The symbols H and S in the CIGAR field of a SAM record represent 'hard' and 'soft' clipping, respectively.

   SAMtools can be used to convert data from SAM to BED format and vice-versa.

   The CIGAR alignment field can contain insertions, deletions, matches or substitutions, and hard and soft clipping.

   Introns are represented with Ns in a CIGAR string.

4. Which of the following is incorrect:

   The BED format can be used to represent gene features.

   The GTF format can be used to represent gene features.

   SAMtools flagstats reports the total number of mapped reads.

   The SAM format is used to represent alignments.

5. Which of the following is NOT an alignment operation:

   Hard clipping

   Soft clipping

   Cut and paste

   Padding

6. What is the minimum number of columns that are sufficient to specify a BED format?

   6

   5

   7

   3

7. Which of the following represents the most accurate conversion into BED of the GTF record:

```
chr1    CLASS   exon    516     811     100 +   .   gene_id “genA”; transcript_id “genA.1”;
chr1    CLASS   exon    1001    1115    100 +   .   gene_id “genA”; transcript_id “genA.1”;
chr1    CLASS   exon    3010    3312    100 +   .   gene_id “genA”; transcript_id “genA.1”
```

```
chr1    515 3312    genA.1  100 +   515 3312    0   3   296,115,303 0,485,2494
```

```
chr1 515 3312 genA.1 + 515 3312 0 3 296,115,303 516,1001,3010
```

```
chr1 515 811 genA.Exon1 100 + 515 811 255,0,0
chr7 3009 3312 genA.Exon3 100 + 3009 3312 255,0,0
```

```
chr1 515 811 genA.Exon1 100 + 515 811 255,0,0
chr7 1000 1115 genA.Exon2 100 + 1000 1115 255,0,0
chr7 3009 3312 genA.Exon3 100 + 3009 3312 255,0,0
```

8. Determine the number of genes, transcripts, exons per transcript, gene orientation (strand), and the length of 5' most exon(s) from the GTF snippet below. Select the correct answer.

```
chr1    HAVANA  gene        3205901 3671498 .   -   .   gene_id "MUSG51951.5";
chr1    HAVANA  transcript  3205901 3216344 .   -   .   gene_id "MUSG51951.5"; transcript_id "MUST162897.1";
chr1    HAVANA  exon        3213609 3216344 .   -   .   gene_id "MUSG51951.5"; transcript_id "MUST162897.1";
chr1    HAVANA  exon        3205901 3207317 .   -   .   gene_id "MUSG51951.5"; transcript_id "MUST162897.1
chr1    HAVANA  transcript  3206523 3215632 .   -   .   gene_id "MUSG51951.5"; transcript_id "MUST159265.1";
chr1    HAVANA  exon        3213439 3215632 .   -   .   gene_id "MUSG51951.5"; transcript_id "MUST159265.1";
chr1    HAVANA  exon        3206523 3207317 .   -   .   gene_id "MUSG51951.5"; transcript_id "MUST159265.1";
```

Genes: 1; Transcripts: 2; Exons: 2,2; Strand: -; Length of 5' exon(s): 2736, 2194.

MUST162897.1: length = 3216344 - 3213609 + 1 = 2736

MUST159265.1: length = 3215632 - 3213439 + 1 = 2194

Genes: 1; Transcripts: 2; Exons: 1,3; Strand: -; Length of 5' exon(s): 2736,1417.

Genes: 2; Transcripts: 1; Exons: 4; Strand: -; Length of 5' exon(s): 2736, 2194.

Genes: 2; Transcripts: 2; Exons: 3,1; Strand: -; Length of 5' exon(s): 2736, 795.

9. Which of the following is FALSE for the following read alignments:

```
R1  83  chr12   9232390 255 50M =   9232180 0
ATGGCAGAGCCTAATATGTCTCCTAGAGAATGGGAGAGATGGGAAGTCAT  HGHHHHHHHHHHHHHHHHHHHHHHHHHHHI
GIIIIHHHHHHHHHHHGHHFH  NM:i:0  NH:i:1  HI:i:0
R2  97  chr12   9232391 255 28M278N22M  =   9242529
0   TGGCAGAGCCTAATATGTCTCCCAAAACTGAGACAGAAGCTCGGGCAGAT  D>DDDHHHHHHHHHHIHIHHHHHHIH
HHHIGFFGGGHHHHHHHHHHFB.F   NM:i:4  NH:i:3  HI:i:0  XS:A:+  NS:i:2
R3  77  *   0   0   0   *   *   0   0   CTGATATGAGGAAAGAGGATTGCTTAAGCCCAGGAGGTAGA
GGCTGTACC  @@@FFDFFHFFHHJJJJIJEGFGIGHHIHIIIIGCDE?D?FGGCBHDGGG
```

R1: FLAG 83 (0x53) 0x1: paired, 0x2: properly paired, 0x10: reverse strand, 0x40: first in pair

R2: FLAG 97 (0x61) 0x1: paired, 0x20: reverse strand, 0x40: first in pair; mate is supposed to be at 9242529 → has a mapped mate based on this but 0x8 (mate unmapped) is not set

R3: FLAG 77 (0x4D) 0x1: paired, 0x4 read unmapped; 0x40: first in pair

R2's mate is unmapped.

R3 is unmapped.

R1 maps uniquely to the genome.

The R1 alignment is the primary mapping (hit index 0) for that read.

R1 has an exact match to the genome.

R1, R2 and R3 all have length 50.

R2 maps in 3 places within the genome.

R2 has an exact match to the genome.

10. For the alignment below, which statements are FALSE? The binary encoding for 97 is 97 = 0000 0110 0001. Select all answers that apply.

```
R2  97  chr12   9232391 255 28M278N22M  =   9242529
0   TGGCAGAGCCTAATATGTCTCCCAAAACTGAGACAGAAGCTCGGGCAGAT  D>DDDHHHHHHHHHHIHIHHH
HHIHHHHIGFFGGGHHHHHHHHHHFB.F   NM:i:4  XS:A:+  NS:i:2
```

FLAG 97 (0000 0110 0001) 0x1: paired, 0x20 read reverse strand; 0x40: first in pair; no 0x4 or 0x8 set → both read and mate are mapped

The read matches to the genome with 4 differences. NM:i:4

The length of the read is 50 bp. CIGAR: 28M278N22M 28M + 22M = 50 bases

This is the first read in the pair.

The read sequence is reverse complemented in the alignment.

The two mates are identical in sequence.

Both the read and its mate are mapped.

The alignment represents a potential PCR or optical duplicate. No 0x400 (duplicate) flag

The sequence of the read's mate is reverse complemented in its alignment.

The alignment passes quality checks.

Both the read and its mate are mapped.

The read matches to the genome with 4 differences.

11. Files ‘A.bed’ and ‘B.bed’ contain the following sets of intervals:

```
File A                      File B
chr1    100 400             chr1    300 500
chr1    1000    1400                chr1    900 1600
chr1    2000    2400                chr12   2000    2200
```

What would be the answers for the following sequence of commands:

```
bedtools intersect –wao –a A.bed –b B.bed | sort –u | wc -l
bedtools intersect –wo –a A.bed –b B.bed  | cut –f1-3 | sort –u | wc -l
bedtools intersect –wo –a A.bed -b B.bed  | cut -f4-6 | sort –u | wc -l
```

3, 2, 2

3, 6, 3

3, 6, 6

3, 2, 6