**Question 1**

DNA has been extracted from an environment sample and paired-end reads were sequenced from the DNA afterwards. Below, we show read pairs aligned to a reference sequence called Ref_seq. Let's suppose that the length of each read is 100bp and that Ref_seq has a length of 400bp (the figure below is not to scale!).

![Reference sequence and reads alignment](alignment_diagram.png)

<table>
<thead>
<tr>
<th>Ref_seq</th>
<th>A_R1</th>
<th>B_R1</th>
<th>C_R1</th>
<th>A_R2</th>
<th>B_R2</th>
<th>C_R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start position on Ref_seq</td>
<td>51</td>
<td>201</td>
<td>201</td>
<td>150</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>End position on Ref_seq</td>
<td>150</td>
<td>300</td>
<td>300</td>
<td>201</td>
<td>150</td>
<td>300</td>
</tr>
</tbody>
</table>

*What is the insert size when calculated from the paired-end reads A_R1 and A_R2?*

**A:** It's 300 – 51 + 1 = 250 (distance between leftmost position of A_R1 and rightmost position of A_R2)

**Question 2**

In total 6 reads (3 pairs A, B and C), each with a length of 100bp have been mapped to the reference sequence.

*What is the average sequencing depth for Ref_seq?*

**A:** 6 * 100 / 400 = 6 / 4 = 1.5

**Question 3**

Suppose that the six reads together align to 350 nucleotide positions of Ref_seq at least once.

*What is the sequence coverage?*

**A:** 350 / 400 = 87.5 %

**Question 4**

Originally the DNA from the environment sample was fragmented with a fragment size as calculated in Question 1.

*How many fragments aligned to the reference sequence named Ref_seq?*

**A:** It's three fragments. Each fragment is sequenced from both ends, resulting in one read pair per fragment.
**Question 5**

In general, because we end up with a random sample of DNA fragments from our environmental sample, we should expect relatively many fragments from big genomes and few fragments from smaller genomes.

If we want to answer which reference sequences are most abundant in a sample, we need to normalize the corresponding read-counts (or fragment-counts) to the size of each reference sequence.

In addition, we also want to normalize to the total number of fragments (or reads) in a sample. This allows us to compare abundances between samples. In order to do this, we calculate FPKM values – FPKM is an abundance measure accounting for both the length of the reference sequence and the total number of reads in a sample:

\[
\text{FPKM} = \frac{\text{Fragments}}{\text{Size of reference sequence (in Kbp) \times \text{Total reads}}}
\]

Suppose we have obtained sequence data from two samples and we map the data to a reference database containing two sequences (Seq_1 and Seq_2). These are the resulting numbers of fragments we can map to each reference sequence:

<table>
<thead>
<tr>
<th></th>
<th>Fragments in sample 1</th>
<th>Fragments in sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq_1</td>
<td>320</td>
<td>350</td>
</tr>
<tr>
<td>Seq_2</td>
<td>1200</td>
<td>2000</td>
</tr>
</tbody>
</table>

The two reference sequences differ in size:

Seq_1: 800 bp
Seq_2: 1,100,000 bp

In addition, the samples yielded a different number of total reads:

Sample 1: 10,000,000 total reads
Sample 2: 20,000,000 total reads

**a)** Calculate the FPKM values for Seq_1 and Seq_2 in each sample.

<table>
<thead>
<tr>
<th></th>
<th>FPKM in sample 1</th>
<th>FPKM in sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq_1</td>
<td>320 / 0.8 / 10 = 40</td>
<td>350 / 0.8 / 20 = 21.875</td>
</tr>
<tr>
<td>Seq_2</td>
<td>1200 / 1100 / 10 = 0.109</td>
<td>2000 / 1100 / 20 = 0.09</td>
</tr>
</tbody>
</table>

**b)** The reference sequences Seq_1 and Seq_2 could represent two different organisms. Which organism is the most abundant in each sample?

A: Seq_1 is more abundant in both samples.

**Question 6**
Two forward reads are shown below in FASTQ format – the information for each read is contained in four lines:

**Header starting with “@” symbol**

**Nucleotide sequence**

“+” symbol

**Quality scores for each nucleotide position**

Here’s what actual sequence data may look like in a FASTQ file (the red highlights are added by us and would not usually be part of FASTQ):

```
@HWUSI-EAS664L:24:64FGCAAXX:4:1:2853:1232 1:N:0:CTTGTA
CCTCGGACGATTGCCGATTAATTCTGGGTACCACGATGCTTGTTTTCACCACAAAGAT
+HHHHHCHHHHHHHHHD#BDDABCCAGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHG#BFG
@HWUSI-EAS664L:24:64FGCAAXX:4:1:5315:1234 1:N:0:CTTGTA
CAGTGCCATCGTAATANTGAGTGCTGGCTCGAAGATGGAGAGCGTTAAGGCGATCCGAT
+IIIIIIIIIIIIIIIF#FFEFAFFEIIIIIIIFCGG?EEDGIEHIHHIIGHEEGGIEG
```

The characters in the quality score line represent phred-scores (Q) in ASCII-base 33.

This means that the quality represented by the letter ‘H’ can for instance be translated to the corresponding phred-score by looking up the integer value for the symbol ‘H’ and subtracting 33 from that number:

Integer value for ‘H’ is 72, this leads to the phred-score

\[ Q = 72 - 33 = 39 \]

The sequencing-error-rate, that is the probability that a particular nucleotide is wrongly sequenced, can then be calculated as:

Error rate \( p(E) = 10^{\frac{Q}{10}} \); in the example it’s \( p(E) = 10^{\frac{39}{10}} \approx 10^{-4} \)

\( a) \) *The integer value for the letter ‘?‘ is 63. What is the corresponding phred-score Q for this letter?*

A: \( Q = 63 - 33 = 30 \)

\( b) \) *What is the probability that a nucleotide has been wrongly determined if the quality letter is ‘?’*

A: \( p(E) = 10^{\frac{-30}{10}} = 10^{-3} = 0.001 \)