Q1) A)

Velvet:

1) K = 21
    - Longest contig size is: 808
    - N75 is = 41
    - N80 is = 41

2) K = 55
    - Longest contig size is: 265397
    - N75 is = 60972
    - N80 is = 52331

SOAP:

1) K = 13
    - Longest contig size is: 55
    - N75 is = 14
    - N80 is = 14

2) K = 21
    - Longest contig size is: 845
    - N75 is = 26
    - N80 is = 24

3) K = 55
    - Longest contig size is: 6742
    - N75 is = 111
    - N80 is = 103

SPAdes:

2) K = 21
    - Longest contig size is: 131828
    - N75 is = 14323
    - N80 is = 12823

3) K = 55
    - Longest contig size is: 224454
    - N75 is = 56527
    - N80 is = 45407

Q1) B

Algorithm Description:

Consider we have contigs C1, C2, C3... Cn with lengths L1, L2, L3...Ln.

Then my algorithm to find NXX will work as follows,

-> First sort lengths L1, L2, L3...Ln in descending order(nlog(n)) and find total length = L1 + L2+...+Ln

-> then find fraction = total length * (XX/100)

-> then go on subtracting sorted lengths of contig from fraction until you end up (fraction <=0)

-> Last length subtracted from fraction is your NXX statistics.

Time and Space Complexity:

Time Complexity - O(n log(n)) -- (Considering sorting complexity)

Space complexity - O(1)

==========

Q 2.

Idea of Hammer

-> Methods proposed before Hammer assume that the sequenced reads are having uniform coverage but that is not always the case specially with single-cell sequencing. On the other hand Hammer method provides a solution regardless of such assumption. Hammer is based on a combination of the Hamming graph and a simple probabilistic model. Hammer forms a hamming graph where vertices represent k-mers and this vertices are connected with edges if two vertices are having hamming distance within T (some constant). Once graph is build, it is traversed to find connected components also called clusters using algorithm similar to Reptile. Each of the cluster formed is processed separately to output error corrected sequencing datasets.

For example, Consider if we have following Kmers,

GAAATACTGACTCA 1 (multiplicitie)
GACATACTGAGTCA 1 (multiplicitie)
GACATAGTGACTCA 1 (multiplicitie)

Consensus output will be,
GACATACTGACTCA

Idea of BayesHammer

In Hammer, a central k-mer in each cluster of the Hamming graph is considered as error free while other k-mers in the cluster are erroneous. In cases, when genome has repeated regions with similar but not identical genomic sequences which would be bundled together by Hammer. For the clusters with large enough diameter or cluster with several k-mers of large

multiplicities, it is more sensible to consider more than one central k-mer with large multiplicities. This is a basic idea of BayesHammer method. First step in BayesHammer method is same as Hammer in which cluster are created in hamming graph. In the next step, read quality value and Bayesian penalties is considered to sub-cluster existing clusters. Sub-clustering helps to capture the complex structure of repeats in the genome by separating even very similar k-mers that come from different instances of a repeat, which is not possible with Hammer.