CENTRAL UNIVERSITY OF SOUTH BIHAR

Department of Bioinformatics

End-Term Open Book Examination Session: 2020-2022 Semester: 2nd (April-July 2021)

Programme: **M.Sc. Bioinformatics** Date: - - - - - -

Course Code: MSBIS2001E04 Course Title: Whole Exome Sequencing Data Analysis

Duration: **2 ½ hours** Course Credits: **4**

Maximum Marks: **50**

Instructions:

1. *Preferably write your answers on A4 size plain paper (non-ruled) sheets.*
2. *Write your required details on the first page in the same order as specified below:*

Name: …… **Programme**:…… **Semester**: ……

Course Title: …… **Course Code**: ……

Total No. of pages used: ……**Date**: …… **Signature**: ……..

3. *After completing the examination, write page number on the top right corner of each page in the format: 1/x, 2/x, ..., x/x where 'x' is the total number of pages used. If you have used total 6 pages then your page numbers will be 1/6, 2/6,.....6/6.*
4. *The students have to write the answers on both side of the sheet (A4 size plain paper non ruled sheet.*
5. *The questions asked here are basically designed to assess the interpretation and application of knowledge, comprehension skills, and critical thinking skills rather than only recalling capacity.*
6. *Total twelve short answer questions of **five points each** are given covering the entire course content.*
7. *Answer **any ten** questions in total in maximum **two and half hours**.*
8. *The maximum limit to answer a question is 200 -300 words.*
9. *At the start of the examination all the questions will be released through e-mail and/or WhatsApp.*
10. *The total time limit to attempt the question paper is **two and half hours**. Along with the two and half hours, extra 30 minutes will be given for IT related activities such as downloading questions, scanning of answer sheets and uploading/emailing them.*
11. *After completing the examination within the stipulated time (two and half hours, scan your answer sheets or click pictures and submit it electronically in **one single file** (preferably PDF) to the course instructor through e-mail (**vksingh@cub.ac.in**) strictly within stipulated time limit for submission (Three hours). Before submitting, rename your file and keep your name and enrolment number as file name.*
12. ***Please note** that do not use these extra 30 minutes for writing answers. Rather, finish writing as soon as possible within two and half hours and immediately submit your answers in the prescribed way given below. **Due to any reason, if a student is unable to submit the answer sheet file within the time limit, the university will not consider this examination and conduct another examination in the conventional mode whenever the conditions return to normal and circumstances permit or the university deems suitable. No other option or reason shall be entertained**.*
13. *In case you feel difficulty in submitting the answer sheet file through e-mail, then you are required to submit it to the concerned course instructor through WhatsApp **within the stipulated time only** and email it later on (within 48 hours) along with the screen shots of WhatsApp submission.*

……..

Question Paper (Answer any 10 questions)

1. The third line (line starting with + sign) in FASTQ file format is redundant and does not have any additional information presented in first line (line starting with @ sign). Why the third line should/shouldn't be removed from the FASTQ format? Justify your answer.
2. All SNP(s) can be considered as SNV(s) but all SNV(s) are NOT SNP(s). Why??
3. Calculate "Sanger format phred quality score" and "solexa/illumina 1.0 format quality score for base call error probabilities 0.75, 0.6, 0.3, 0.001, 0.0001. What you can interpret by comparing the two quality scores? [Calculate the two type of quality score for each of the given base call error probabilities and then make your interpretation]
4. Identification of low quality reads (possibly to remove them from further analysis) on the basis of average phred quality score is not recommended. Does the statement made here is valid statement ? Provide proper reasoning in support of your answer.
5. Recent genome-wide association studies (GWAS) on COVID patients has been performed either using SNP arrays or whole genome sequencing technology. What could be the possible reasons because of which Exome sequencing method was not used for these studies? In your opinion, for what purpose Exome sequencing experiments has most usability in present day context?
6. A hypothetical example of a read aligned on reference is presented below. Assume phred quality score of 30 for all the bases of the read and calculate phred based mapping quality score for this hypothetical alignment.

Reference:	**A**	**T**	**G**	**G**	**C**	**T**	**G**	**T**	**C**	**A**	**G**	**C**
Read					**C**	**T**	**A**	**T**				

7. Use Lander-Waterman sequencing statistics to calculate following:
   a) Percentage of genome not sequenced if sequencing was done at 2x fold coverage
   b) Percentage of genome sequence 2 times or less if sequencing was done at 5x fold coverage.
8. What are the CIGAR operations that consume query only? Why CIGAR operation "N" is not usually observed for Exome sequencing reads? Given reference sequence as:

   ATGCTCATGTCATCGTGATGCTGAT

   Find out the hypothetical read sequences that produces following CIGAR string when aligned to given reference.

   Read 1 5M2D3M1I6M
   Read 2 4=2X3I4=1X2D6=
9. Why rare genetic diseases are most likely to be associated with rare genetic variants?
10. Discuss the minimum essential steps and tools with their order in which they should be used in a pipeline to analyses Exome sequencing data.
11. A particular genomic location was observed to exist in balletic form (allele A and B) in population of 2250 individuals. The minor allele (allele B) frequency (MAF) was find out to be 0.2. The total number of individuals that have a specific disease in this cohort of 2250 individuals was found out to be 532, 246 and 56 for the genotype AA, AB and BB respectively. Using the data presented find out if the occurrence of genotype BB is significantly associated with the disease?
12. In relation to Exome sequencing experiments discuss the ethical issues related to
    a) Data collection protocols
    b) Confidentiality of the generated reports
    c) About the ownership and access to the generated data.