1. Genomics

(1) What is the meaning of “coverage” in sequencing? The number of reads aligned to the position (2.5pt)

(2) Name one long-read sequencing technique PacBio/ONT (misspelling allowed) (2.5pt)

(3) Match the sequencing methods with their applications: (10pt, -2pt for each mistake, no pt if all wrong)

Methylated sites C
Long-range interactions D
Protein binding A
Chromatin accessibility B

A: ChIP-seq
B: ATAC-seq
C: Bisulfite-seq
D: Hi-C

2. Proteomics

(1) What is a major advantage of cryo-em compared to X-ray crystallography?

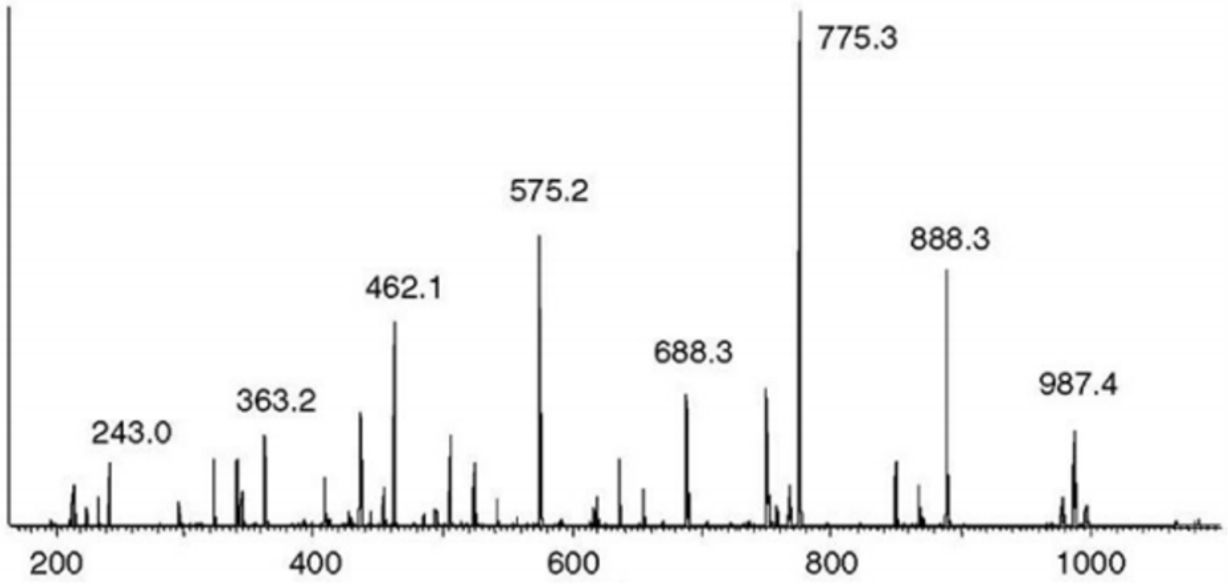
Does not need crystallization (other answers accepted) (5pt)

Material requirement

Suitable for large proteins/complexes, and membrane proteins/complexes

Capture multiple conformations

(2) Below is a mass spectrum result. What are the x-axis and y-axis?



X = m/z (mass-to-charge ratio)

Y = Relative abundance or Intensity

(2.5pt x2)

(half score if order is wrong)

3. Personal genomics

True or False: (5pt x2)

(1) Many psychiatric diseases, such as schizophrenia are highly heritable. True

(2) Typically, an individual could have ~30,000 single nucleotide polymorphism (SNP) with respect to the human reference genome. False

4. Sequence comparison

(1) Compute the position probability profile matrix for the following nucleotide sequences (you just need to consider the simplest case): (5pt, -1pt for one mistake)

ACTGG
ACGGC
ATGCG
ATCGC

A	1				
T		.5	.25		
G			.5	.75	.5
C		.5	.25	.25	.5

(2) What is the probability of observing sequence ACGCC, given the profile matrix in (1)?

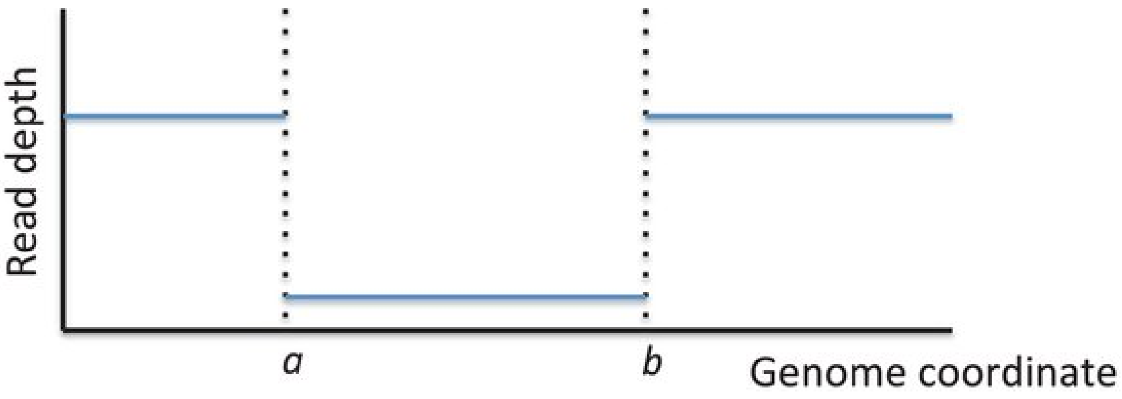
Note: Please write out the expression. You do not have to calculate the exact number. (5pt)

1x0.5x0.5x0.25x0.5=0.03125

5. Variant identification

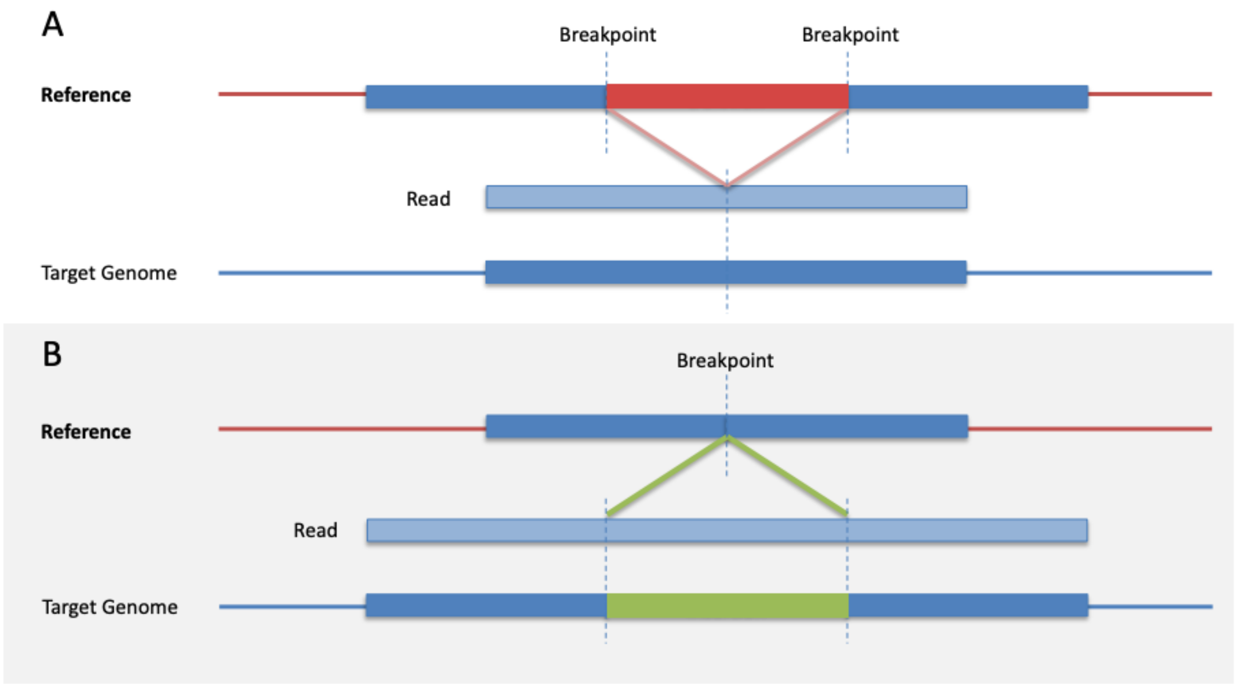
(1) The following read depth coverage may indicate a _______ (a type of structural variant) between a and b

Deletion (5pt)



(2) Choose the split-read diagram for the structural variant in (1)

A (5pt, only has to match part 1)



6. a. A relational database has the following columns for a student table:

Student ID (primary key)	Student Name	Student's Department	Advisor ID	Advisor Name

Explain why this table is not in third normal form

Transitive dependency exists: advisor name depends on advisor ID, not student ID (5pt)

b. Explain and draw columns for a new table (or tables) for how you would normalize this database further (to third normal form): (5pt, -2pt for one mistake)

Remove transitive dependencies: Split table so that all non-primary key columns are dependent only on the primary key of its table (will accept answers above 3NF if they make sense)

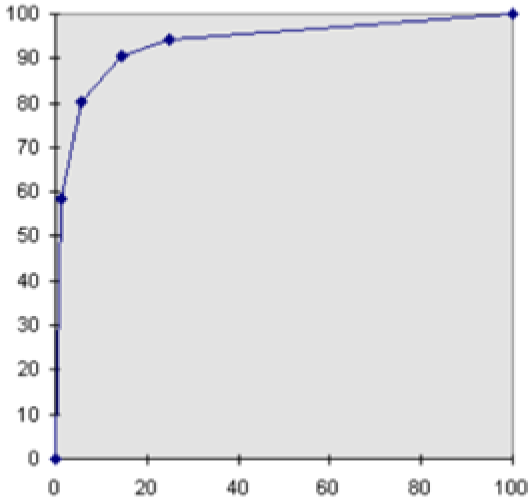
Student ID (primary key)	Student Name	Student's Department	Advisor ID

Advisor ID (primary key)	Advisor Name

7. eQTLs are genomic loci that explain variation in ____________________ (5pt)

mRNA expression levels

8. In the following ROC (Receiver Operating Curve), the area under the curve indicates that this classifier performs better than random guessing. Label the axes and also define each axis in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN)

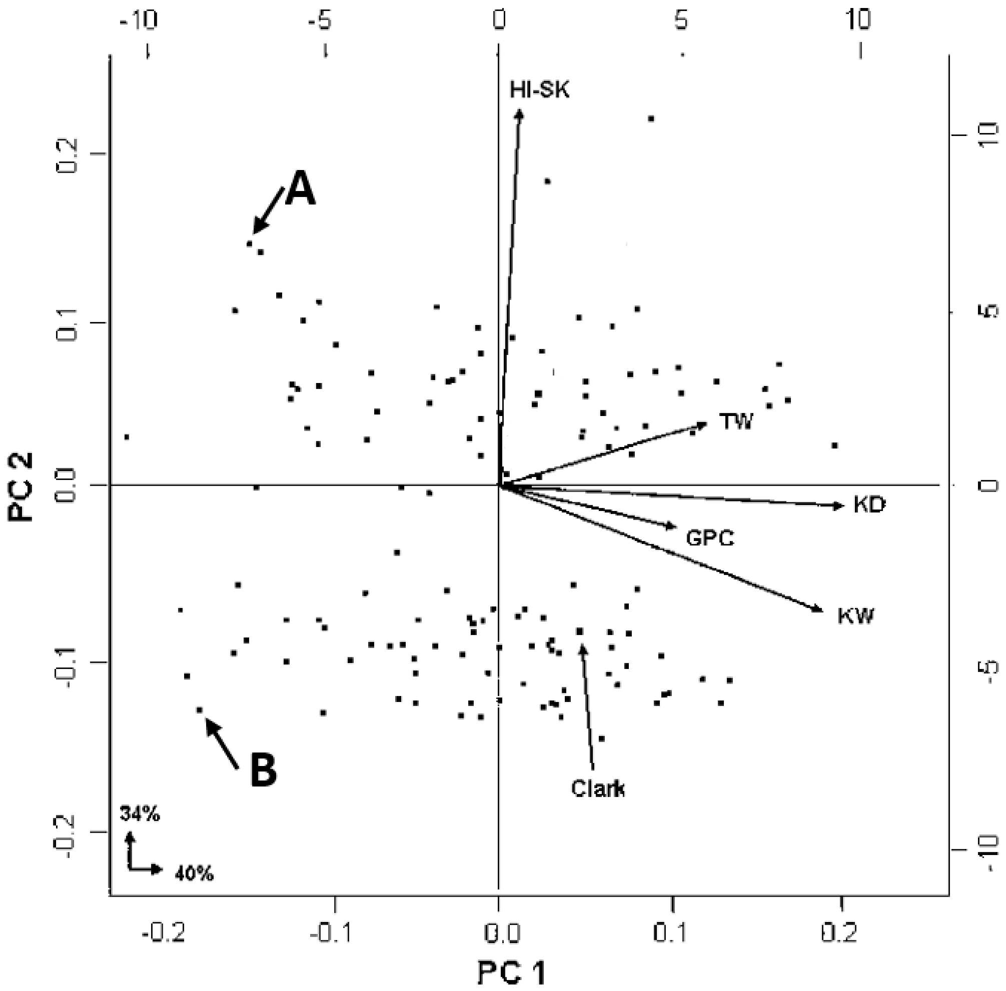


y-axis: TPR = TP/(TP+FN)

x-axis: FPR = FP/(TN+FP) or 1-TN/(TN+FP)

(5pt x2, half points if label wrong but eq are correct)

9.



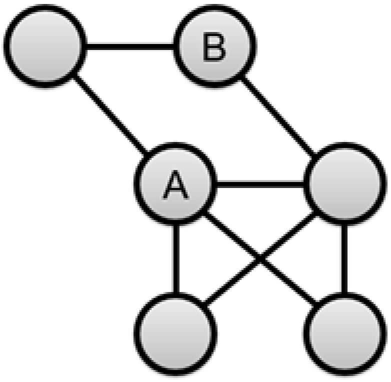
For the above PCA biplot, assume that the first 2 components account for 95% of the variance in the original dataset. Answer True or False: (2.5pt x4)

KD and GPC are strongly correlated True

KD and HI-SK are strongly correlated False

KD strongly projects onto principal component 2 False

Points A and B differ mostly in HI-SK True



10. For the above network, what are the following values: (2.5pt x4)

The degree of node A?

4

The clustering coefficient of node A?

2/6 or 1/3

The clustering coefficient of node B?

0

The shortest path length from A to B?

2