University of Pennsylvania
BIOL4536 Fall 2023
Professor: Gregory R. Grant
Exam#2 (Practice Exam)
- similar length to acutal exam -

November 1st, 2023 *25 Questions, 4 points each*

Question 1. Suppose an indel has negative score. True or False: an optimal local alignment can start with an indel.

ANSWER: False. Removing the indel at the end would raise the score.

Question 2. Suppose you have two sequences S_1 and S_2 and S_1 is twice as long as S_2. Then the optimal local aligment between them can have no indels.

ANSWER: True.

Question 3. True or False. It makes sense to report a multiple sequence alignment where one position in the alignment is an indel in all sequences.

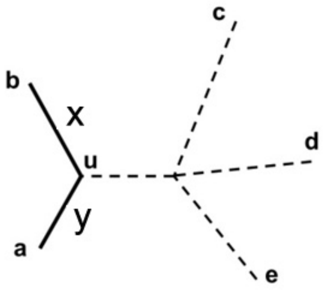
ANSWER: False.

Question 4. True or False. The columns of a Position Weight Matrix must add to one.

ANSWER: True.

Question 5. Consider the following distance matrix and unrooted tree. What's longer, $d(a, u)$ or $d(b, u)$? *Hint: You should do it without figuring out what X or y are exactly.*

	a	b	c	d	e
a	0				
b	5	0			
c	9	10	0		
d	9	10	8	0	
e	8	9	7	3	0



ANSWER: $d(b, u)$ is longer because $d(b, c) > d(a, c)$ (look them up in the table, the left-hand-side is 10 and the right-hand-side is 9.

Question 6. True or False. In a SAM file, whether the read is a PCR duplicate is encoded in the bitflag.

ANSWER: True.

Question 7. If we are clustering rows of the following block to construct a BLOSUM*N* matrix, what's the largest value of *N* that will result in just one cluster?

ANSWER: $N = 75$.

Is it possible to find an N that will give exactly two cluters?

ANSWER: No, $N = 75$ gives one and $N = 76$ gives four.

$$\begin{vmatrix} A & A & C & C \\ A & A & C & G \\ A & G & C & G \\ G & G & C & G \\ G & G & C & G \\ G & G & C & G \end{vmatrix}$$

Question 8. In a BLAST search, explain why the algorithm masks out polyA tails.

ANSWER: Because that's low complexity sequence and in fact all mRNA's have polyA tails.

Question 9. True or False. If the word size in constructing a BLAST index is twice as long, then the index will be twice as large.

ANSWER: False. There are a lot more than twice as many words of length $2N$ than length N.

Question 10. True or False, in the BLAST random walk, the max excursion is the maximum height obtained by the walk.

ANSWER: False.

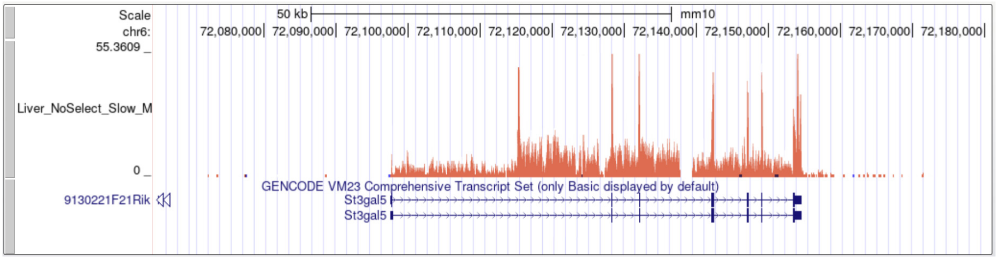
Question 11. True or False. In the alignment of an RNA-Seq read to the genome, one read can cross more than one exon/exon junction.

ANSWER: True.

Question 12. When doing RNA-Seq quantification from paired-end data, suppose the forward read's alignment all by itself is ambiguous, meaning there's more than one isoform it could have come from, and suppose the same is true for the reverse read. Is it possible that we could pin it down to one isoform by considering both forward and reverse read alignments together?

ANSWER: Yes, it is certainly possible. That's one of the advantages of paired-end RNA-Seq data.

Question 13. There are many observations one could make about this RNA-Seq coverage plot. But one definitely stands out, what is it? (*You can make more than one observation if you want, as long as the one we're looking for is among them you'll get full credit.*)



ANSWER: There's clearly an unannotated isoform, the first (leftmost) spike of signal is clearly an exon, but there's no exon annotated there in either isoform. So there's an unknown isoform expresssed in this sample. Another observation is that this gene has particularly high intron signal and in fact differential intron signal (it's lower in the leftmost intron). A third observation is that there's a curious dropout of signal, perhaps due to low complexity sequence or a region that just doesn't like to sequence for some reason.

Question 14. True or False. In order to do phasing, you need a read long enough to cover two SNP's at a time.

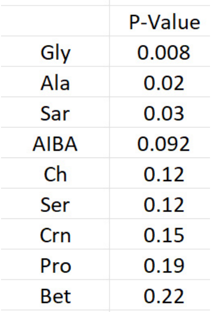
ANSWER: False, you can use population information.

Question 15. H3K4me3 is a(n):
(A) Transcription factor binding site
(B) Binding motif ID.
(C) SNP location
(D) Histone modification ⟵ **THIS ONE**
(E) Open chromatin region

Question 16. Suppose a *q*-value cutoff of 0.25 leads us to expect 70 false positives. How many genes have *q*-value cutoff ≤ 0.25?

ANSWER: 280

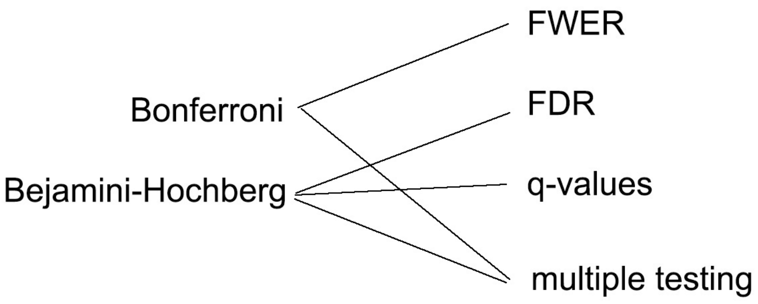
Question 17. Suppose you have done ten tests with the following *p*-values. You want to use the Bonferroni correction to control the FWER at the 0.05 level. How many significant genes do you get?



	P-Value
Gly	0.008
Ala	0.02
Sar	0.03
AIBA	0.092
Ch	0.12
Ser	0.12
Crn	0.15
Pro	0.19
Bet	0.22

ANSWER: None. You'd have to have one with *p*-value ≤ 0.005

Question 18. Connet the things on the left to (all of) the appropriate things on the right.



Question 19. True or False. We use the FWER for a small number of tests and the FDR for a large number.

ANSWER: True.

Question 20. Consider three consecutive SNPs. There are eight possible haplotypes. Is it possible that a population has five haplotypes and a subpopulation of it has six?

ANSWER: No it's not possible, a subpopulation cannot have a haplotype that's not in the greater population because it's a subset of individuals.

Question 21. What's the long term behavior of the Markov Chain given by this matrix?

$$\begin{array}{c|ccc|} & A & B & C \\ \hline A & 0 & 0 & 1 \\ B & 1/3 & 1/3 & 1/3 \\ C & 1 & 0 & 0 \\ \hline \end{array}$$

ANSWER: It will eventually just bounce back and forth between A and C forever, and never return to B.

Question 22. True of False: BLAST needs to build a different index for each substitution matrix. True of False: BLAST needs a different index for blastp and blastx.

ANSWER: First one True, second one False.

Question 23. Suppose there are two experimental conditions C_1 and C_2 with no differentially expressed genes between them. And suppose there are two other experimental conditions C_3 and C_4 where every gene is differentially expressed between them. For which analysis do we have a bigger multiple testing problme? C_1 vs. C_2 or C_3 vs. C_4?

ANSWER: We have a bigger problem with C_1 vs. C_2. There's in fact no problem with C_3 vs. C_4 since there cannot be any false-positives.

Question 24. Consider a SNP location in the genome that has the two variants A/G. Suppose we do DNA-seq and get 40 reads covering this SNP location. Suppose all reads indicate an A except one G. What could explain this besides heterozygosity?

ANSWER: A sequencing error.

Question 25. Consider the following CIGAR string: 24M2D30M2042N40M7I2S. What is the read length?

ANSWER: Everything adds to the read length excetp N's and D's. So 24 + 30 + 40 + 7 + 2 = 103.