University of Pennsylvania BIOL4536 Fall 2023 Professor: Gregory R. Grant Exam#1

October 2nd, 2023

Name:

25 Questions, 4 points each

Question 1. True or False. There are infinitely many *T*-distributions.

Question 2. Let H_0 be the hypothesis that a subject does not have disease X. Which of the following does not depend on the percent of the population being tested that are affected by disease X?

(A) Prob(not infected | test positive)

(B) Prob(test positive | not infected)

Question 3. A two-sample *T*-test (circle all that apply)

(A) Assumes individual observations are normally distributed.

(B) Can conclude the means of two groups of observations are different.

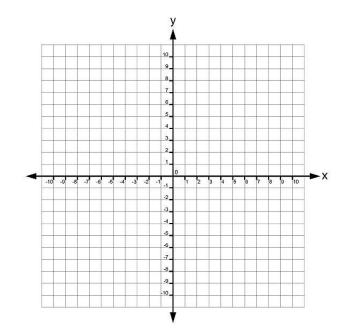
(C) Can conclude the means of two groups of observations are equal.

(D) Requires there to be the same number of replicates in each group.

Question 4. Consider the regression model

$$Y = X^2 + \epsilon$$

Draw the regression curve on the following coordinate system:



Question 5. Suppose we have the simple linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where ϵ is normally distributed and its distribution is independent of X (as usual).

Circle all of the following that are **necessarily** true. Could be none, one or more than one.

(A) The value of *X* determines the value of *Y*.

- (B) The value of *X* determines the probability distribution of *Y*.
- (C) The value of *Y* determines the value of *X*.
- (D) The slope of the regression line is positive.
- (E) The probability distribution of Y is different for different values of X.

Question 6. Estimating the regression curve of the following model means (circle one):

$$Y = \beta_0 + \beta_1 X + \epsilon$$

(A) Estimating *X* and *Y*.

(B) Estimating β_0 and β_1

(C) Estimating ϵ and σ

(D) All of the above

Question 7. Consider the genome browser track below. Circle all that are true.

(A) The gene Tmem231 is on the forward DNA strand.

(B) The codon for the 100th amino acid is CCG.

(C) There is one unique isoform of the gene Tmem231.

(D) The gene Tmem231 has no introns.

Question 8. According to the following genome browser gene annotation track.

Scale		100 kb		mm10	
chr8:	111,800,000	111,850,000	111,900,000	111,950,000	112,000,000
	Cidp1	GENCODE VM23 Comprehens	Tmem231	-←←← -←←← Gabarapl2 <mark>++→→→</mark>	Adatt

How many different genes are shown? Answer:

How many different single exon isoforms are shown? Answer:

How many isoforms with introns are on the reverse strand? Answer:

How many isoforms does the gene with the most isoforms have? Answer:

Question 9. Which of these genome browser Tools would you use to download all gene sequences on chromosome 12 from a particular annotation track? (circle the right one)

ñ	Genomes	Genome Browser	Tools	Mirrors	Downlo	ads My D	ata Vie	ew He
			Blat		t b	nome Brov	vser on N	louse (G
			In-Silic	o PCR	ti	>>> zoom i	in 1.5x 3x	< 10x bas
			Table B	rowser	tt	guided tutoria	al is availab	le for new
			LiftOve	r		•	Close Do	on't show ag
		multi-regio	Gene S	orter		133,303 bp.	anna chror	monomo ran
		Inditi-legio	Variant	Annotation Ir	ntegrator			
		chr8 (qE1) 8qA	Data In	tegrator		4 B1.1 qB	1.3 qB2 8qB	3.1 8qB3.3
		Scale		e Graphs) kb		
		chr8: 111,870,000	11 Gene Ir	nteractions		00 111,910,000 NCODE VM23 (
		Tmem170 matrix ←←	- Other T	ools		~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	H Tr	mem231
			Gm23853			Tmem2 Tmem2		····
	move start	Click on a feature tracks. Drag				. Press "?" for		nortcuts. Use
	collapse all				track se	earch hide all	add custon	m tracks
	-					Mapping	and Sequ	encing
	Base Positio		Patches		<u>aplotypes</u>			<u>Chromos</u>
	dense 🗸	hide	~	hide N		hide 🗸		hide 🗸

Question 10. This is the text box to enter a custom track in the genome browser. The info for a track has been filled in.

Paste URLs or data:	Or upload: Choose File No file chosen	Submit
chr12 12512209 1251230	8	Clear

How long will the custom track's span be exactly? Answer:

Question 11. Refer to the following browser gene annotation track.

Scale		
chr8:	111,870,000 111,880,000 111,890,000 111,900,000 111,910,000 111,920,000 111,930,000 111,940,000 111,950,000 111,960,000 111,950,000 110,0000 110,000 110,0000 110,0000 110,00000000	000
	GENCODE VM23	
Tmem170	Chst5 Chst5 Chst5	
	Tmem231 🔤 👫 😋 🗠 🗠	

(a) Which gene(s) have an intron in the 3' UTR of at least one of its isoforms, but do not have an intron in the 5' UTR of any isoform?

Answer:

(b) Which gene(s) have an intron in the 5' UTR of at least one of its isoforms, but do not have an intron in the 3' UTR of any isoform?

Answer:

(c) Which gene(s) have no intron in either UTR of any of its isoforms? Answer:

(d) Which gene(s) have an intron in both UTRs of all of its isoforms? Answer:

Question 12. You're looking at all of human chromosome M.

	mu		> >> >>> Zoor		(GRCh38/hg38) ase zoom out 1.5x 3x	go examples		
chrM				М				
Scale chrM:	1,000 2,000	3,000 4,000	5,000 6,000	7,000 8,000	9,000 10,000 11,0 VCODE Version 43 (Ensemb		14,000 15,000	16,000
MT-TF MT-RNR1	MT-TV MT-RNR2	MT-TL1 MT-TI MT-ND1 >>>>>>> MT-TQ MT-TM	MT-TW MT-TA MT-TN MT-TC	MT-TS1 MT-TD MT-TD MT-CO2 MT-TK	MT-CO3 MT-TG MT-TG MT-TG MT-TG MT-TG MT-TG MT-TG MT-TG MT-TR MT-TR MT-TR MT-TR MT-TR MT-TR MT-TR MT-TR MT-TR M	MT-TH MT-TS2 MT-TL2 MT-TL2 MT-ND5 >>>>>>>>	MT-ND6 KKKK MT-TE M MT-CYB SSS	MT-TT I MT-TP I
		MT-ND2	MT-TY MT-CO1	MT-ATP8 MT-ATP6		››››››››››››››››››››››››››››››››››››››		

According to the annotation shown:

(a) How many protein coding genes are there on chrM? Answer:

(b) True of False. All protein coding genes on chrM are on the forward strand.

Question 13. There is only one directory in a UNIX file system that is not contained in another (parent) directory.

(a) What is the name of that directory? Answer:

(b) Give a unix command that will take you to that directory. Answer:

(c) If you are in that directory and you execute pwd then what will be returned? Answer:

Question 14.

(a) How do you make a file hidden in UNIX? Answer:

(b) What option of the 1s command makes it display the hidden files? Answer:

Question 15. Suppose file.fa is a fastA file of sequences. Construct a UNIX command to return just the reads in file.fa that contain five (capital) G's in a row but not six, and also contain five (capital) T's in a row but not six.

Question 16. Connect the UNIX control-key-sequence on the left to its action on the right.

ctrl-c	search backwards through history
ctrl-z	cut text from cursor to end of line
ctrl-r	kill current job
ctrl-a	suspend current job
ctrl-e	move cursor to end of line
ctrl-k	paste text from the clipboard to the screen
ctrl-y	move cursor to beginning of line

Question 17. What will the following UNIX command achieve? cat *fa | grep A | grep C | grep G | grep T | wc -1

Question 18. The following grep command will match which of the following in the file sequence.fa? You can assume they will not match any of the ID lines, just the sequence. Circle all that apply

ggrant@workstation:~\$ grep -v AAAA sequence.fa | grep AAA*

(A) AAA
(B) AAAA
(C) AAAAA
(D) AAGG
(E) AAGGAA

Question 19. Construct a UNIX command to count the lines from the file file.txt that consist of exactly five characters (not including the newline).

Question 20. Write down the optimal alignment corresponding to the trace-back in the filled-in Needleman-Wunch table and give its score.

		Α	С	т	G	Α	т	т	С	Α
		-2	-4	-6	-8	-10	-12	-14	-16	-18
Α	-2	2	0	-2	-4	-6	-8	-10	-12	-14
С	-4	0	4	ą	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
С	-8	-4	0	-1	گ	1	-1	-3	0	-2
Α	-10	-6	-2	-3	0	Æ	2	0	-2	2
т	-12	-8	-4	0	-2	2	6	-4	2	0
С	-14	-10	-6	-2	-4	0	4	2	6	4
Α	-16	-12	-8	-4	-5	-2	2	1	4	8

Answer:

Question 21. True or False. $(n + 2)^2 = O((n + 1)^2 + (n - 1)^2).$

Question 22. True or False. If two sequences are the same length, then the optimal global alignment has no indels.

Question 23. Fill in the next two squares in the Needleman-Wunch table. Matches score +1, mismatches score -1 and indels score -1. Draw in also the appropriate arrows.

		Α	С	Α	Α
	0	-1	-2	-3	-4
Α	-1	1	0	-1	-2
С	-2	0	2	1	0
т	-3	-1	1	1	0
G	-4	-2	0	0	
Α	-5	-3	-1		

Question 24. Consider the following two alignments. If there were a scoring scheme such that the second one is higher than the first, then what does that tell you about the score for a mismatch?

	A C T C
	A C - C
And	
	C A G –
	C A C T
Answer:	

Question 25. Assume there are 20 amino acids. A "peptide" is a short sequence of amino acids of length N where $2 \le N \le 50$. What's the smallest value of N so that there are more different peptide sequences of length N than there are DNA sequences of length 5?

Answer: