University of Pennsylvania

BIOL4536 Fall 2023

Professor: Gregory R. Grant

Exam#1

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

ANSWER: True, one for each degrees of freedom $1, 2, 3, \ldots$.

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 $\mid$ test positive)

(B) Prob(test positive $\mid$ not infected) ⟵ **THIS ONE**

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

(A) Assumes individual observations are normally distributed. ⟵ **THIS ONE**

(B) Can conclude the means of two groups of observations are different. ⟵ **THIS ONE**

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

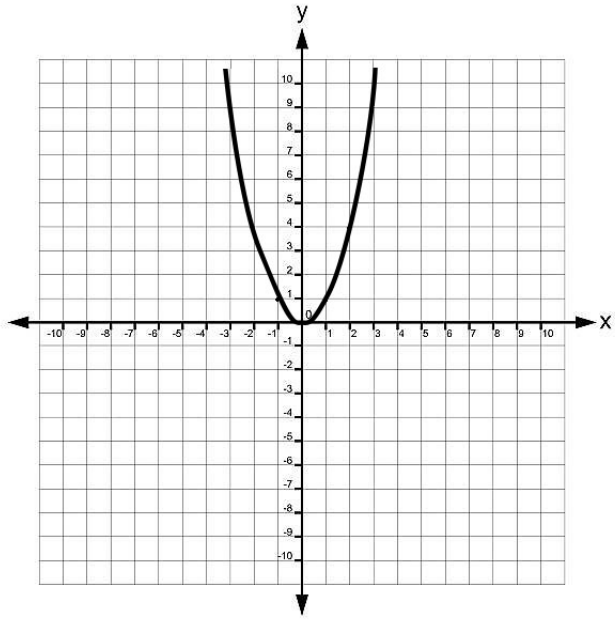
(D) Requires there to be the same number of replicates in each group.

Question 4. Consider the regression model

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

Draw the regression curve on the following coordinate system:

ANSWER: It's the parabola $Y = X^2$



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.

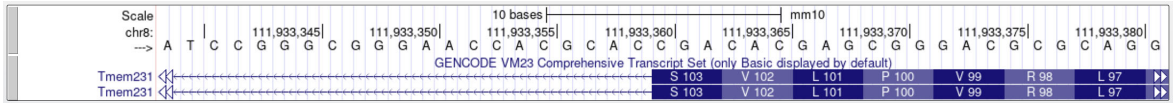
(A) The value of X determines the value of Y.
(B) The value of X determines the probability distribution of Y. ⟵ **THIS ONE**
(C) The value of Y determines the value of X.
(D) The slope of the regression line is positive.
(E) The probability distribution of Y is different for different values of X.

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

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

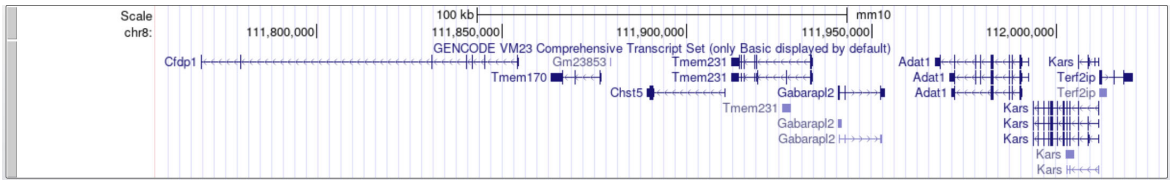
(A) Estimating X and Y.
(B) Estimating β_0 and β_1 ⟵ **THIS ONE**
(C) Estimating ϵ and σ
(D) All of the above

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. ⟵ **THIS ONE**
(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.



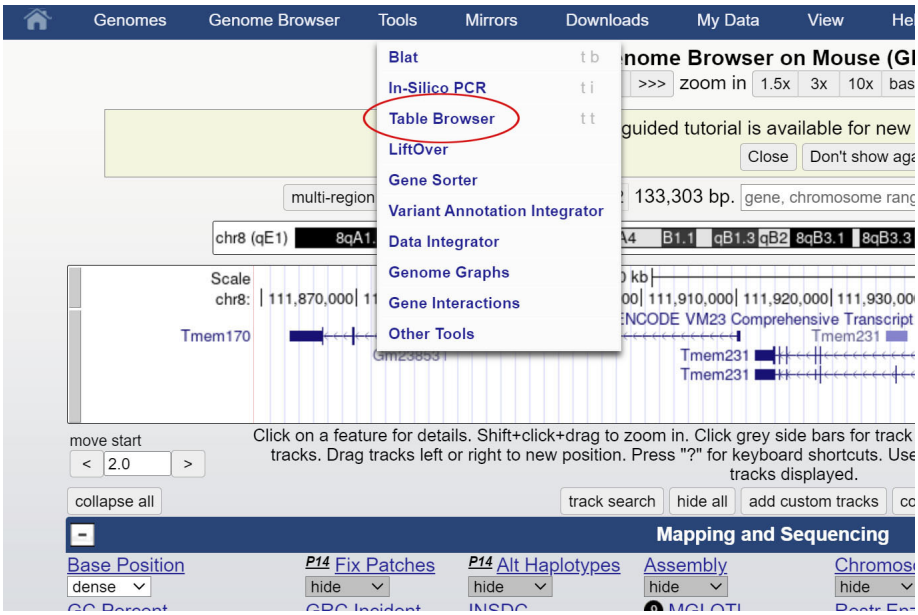
How many different genes are shown? **Answer:** Nine: Cfdp1, Gm23853, Tmem170, Chst5, Tmem231, Gabarapl2, Adat1, Kars and Terf2ip

How many different single exon isoforms are shown? **Answer:** Five: Gm23853, Tmem231, Gabarapl2 (first isoform), Terf2ip (2nd isoform) and Kars (4th isoform)

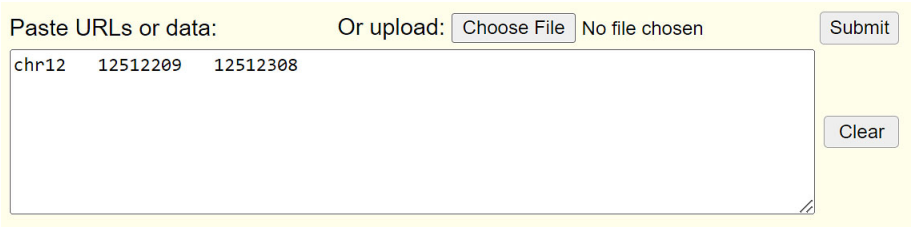
How many isoforms with introns are on the reverse strand? **Answer:** 13

How many isoforms does the gene with the most isoforms have? **Answer:** Kars has six.

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)

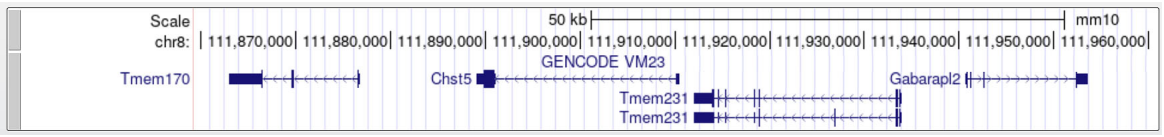


Question 10. This is the text box to enter a custom track in the genome browser. The info for a track has been filled in.



How long will the custom track's span be exactly? **ANSWER:** Since the left endpoint is not included, you can simply just subtract 12512209 – 12512308 = 99.

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



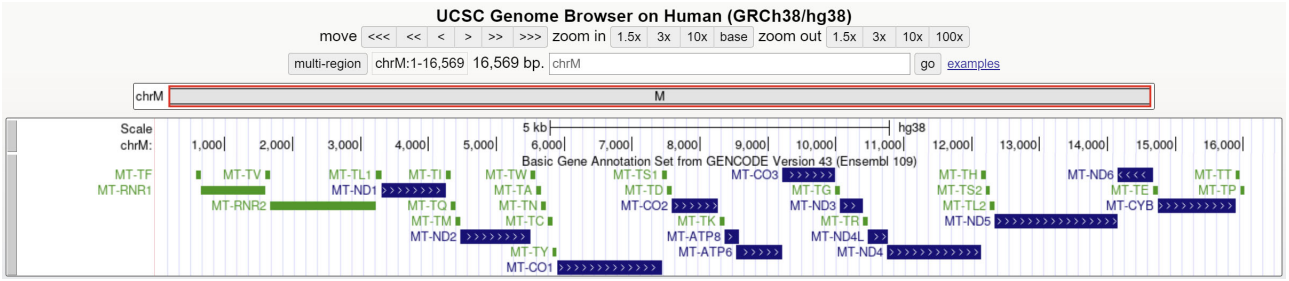
(a) Which gene(s) have an intron in the 3' UTR of at least one of its isoforms, but do not have an intron in the 5' UTR of any isoform?
ANSWER: Tmem231 has several introns in its 3' UTR and no introns in the 5' UTR of either isoform.

(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: Chst5 has an intron in its 5' UTR ut not in its 3' UTR

(c) Which gene(s) have no intron in either UTR of any of its isoforms? **ANSWER:** Two of them: Tmem170, Gabarapl2

(d) Which gene(s) have an intron in both UTRs of all of its isoforms? **ANSWER:** None.

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



According to the annotation shown:

(a) How many protein coding genes are there on chrM? **ANSWER:** The protein coding ones are the thicker ones (which also happen to be colored blue). There are 13 of them.

(b) True of False. All protein coding genes on chrM are on the forward strand.
ANSWER: False, there is one that goes the other way, MT-ND6.

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:** root

(b) Give a unix command that will take you to that directory. **ANSWER:** `cd /`

(c) If you are in that directory and you execute `pwd` then what will be returned? **ANSWER:** `/`

Question 14.

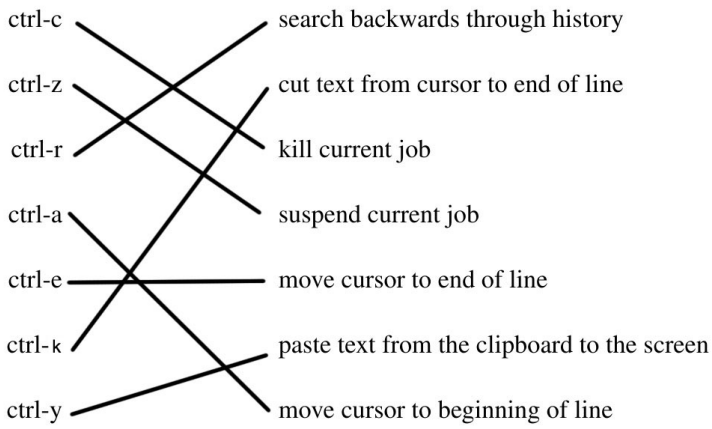
(a) How do you make a file hidden in UNIX? **ANSWER:** You start its name with a dot.

(b) What option of the `ls` command makes it display the hidden files? **ANSWER:** `-a`

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.

ANSWER: `grep GGGGG file.fa | grep -v GGGGGG | grep TTTTT | grep -v TTTTTT`

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



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

ANSWER: This will count the number of rows *in all files in the current directory that end in* `.fa` which have all four letters A, C, G and T.

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 ⟵ **THIS ONE**
(B) AAAA
(C) AAAAA
(D) AAGG ⟵ **THIS ONE**
(E) AAGGAA ⟵ **THIS ONE**

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).

ANSWER: `grep ^.....$ file.txt`

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

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-4	0	4	2	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

ANSWER: Its score is 8 and the alignment is:

$$\begin{array}{cccccccccc} A & C & T & G & - & A & T & T & C & A \\ A & C & - & G & C & A & T & - & C & A \end{array}$$

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

ANSWER: True, because

$$\lim_{n\to\infty} \frac{(n+2)^2}{(n+1)^2 + (n-1)^2)} = \lim_{n\to\infty} \frac{n^2+4n+4}{2n^2+2} = \frac{1}{2}$$

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

ANSWER: False. If a mismatch has a more negative score than an indel, then this can happen.

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.

		A	C	A	A
	0	-1	-2	-3	-4
A	-1	1	0	-1	-2
C	-2	0	2	1	0
T	-3	-1	1	1	0
G	-4	-2	0	0	0
A	-5	-3	-1	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?

$$\begin{array}{cccc} A & C & T & C \\ A & C & - & C \end{array}$$

And

$$\begin{array}{cccc} C & A & G & - \\ C & A & C & T \end{array}$$

ANSWER: Both alignments have two `C`'s aligned to each other and two `A`'s aligned to each other. And both also have a `T` aligned to an indel. So the only difference to the score is the first has another `C` aligned with a `C` (a match) while the second has a `C` aligned to a `G` (a mismatch). If the second alignment scores higher than the first, then that means the score for a mismatch must be higher than the score for a match (which would be weird).

Question 25. Assume there are 20 amino acids. A "peptide" is a short sequence of amino acids of length N where $2 \leq N \leq 50$. 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: There are $4^5 = 2^{10} = 1024$ DNA sequences of length 5. So we need $20^N > 1024$. Since $20^2 = 400$ and $20^3 = 8000$ the answer is $N = 3$.