University of Pennsylvania

BIOL4536 Fall 2023

Professor: Gregory R. Grant

Exam#1 (MAKEUP)

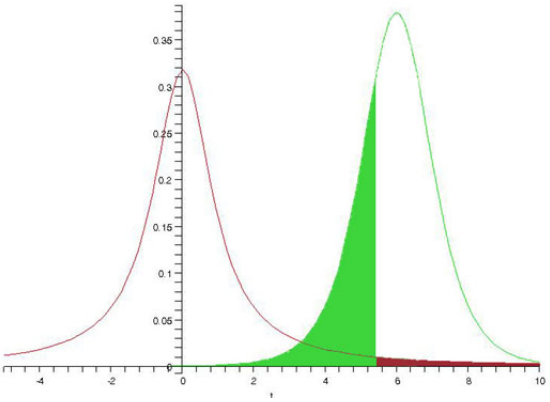
Question 1. True or False. The p-value for a hypothesis test is 0.16, therefore we conclude the means are equal.

ANSWER: False. You make no conclusions from a non-significant p-value.

Question 2. In the AIDS testing example in class, the probability of a subject being negative even though they tested positive was high nearly 80%. What course of action is taken to deal with this problem? A satisfactory answer can be given in less than 10 words, try not to use more than 30. An overly-wordy answer will be penalized.

ANSWER: In a word, "validation". Those who test positive the first time are tested a second time.

Question 3. This graph shows the null and true (unknown) distributions for a T-test run with a significance threshold of 0.01.



If the number of replicates is increased: (circle the one correct answer)

(A) The red shaded area will decrease.
(B) The red shaded area will increase.
(C) The green shaded area will decrease. ⟵ **THIS ONE**
(D) The green shaded area will increase.

Question 4. True or False. To reasonably estimate the regression curve, you need a set of data with at least two different values of X.

ANSWER: True.

Question 5. Consider the regression model

$$Y = 42 + 17X + \epsilon$$

Circle the deterministic part of the right hand side of the equation.

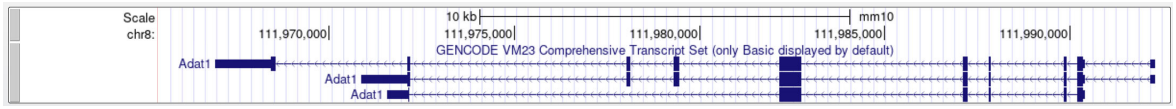
Question 6. Consider the simple linear regression model

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

Suppose there is a significant relationship between X and Y. Then which of the following is true?

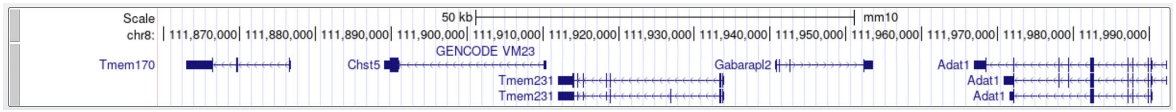
(A) $\beta_0 > 0$
(B) $\beta_0 \neq 0$
(C) $\beta_1 > 0$
(D) $\beta_1 \neq 0$ ⟵ **THIS ONE**
(E) $E[\epsilon] \neq 0$

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



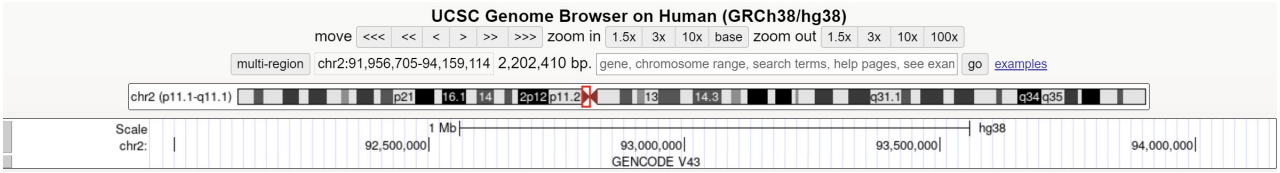
(A) The gene Adat1 is on the reverse strand. ⟵ **THIS ONE**
(B) All isoforms of Adat1 have an intron in their 5' UTR.
(C) No isoforms of Adat1 have an intron in their 3' UTR. ⟵ **THIS ONE**
(D) It's possible that Adat1 has more than three isoforms. ⟵ **THIS ONE**
(E) There is an annotated isoform of Adat1 with the same number of introns as another annotated isoform has exons. ⟵ **THIS ONE**

Question 8. True or False. The following browser gene annotation track shows a gene where its entire coding sequence is contained in one exon.



ANSWER: True, Chst5 is like that.

Question 9. The following genome browser screen shot shows a region of chromosome 2 that's 2,202,410 bases long but has no annotated genes. Examine the screen shot carefully and give the best explanation of why there are no genes here.

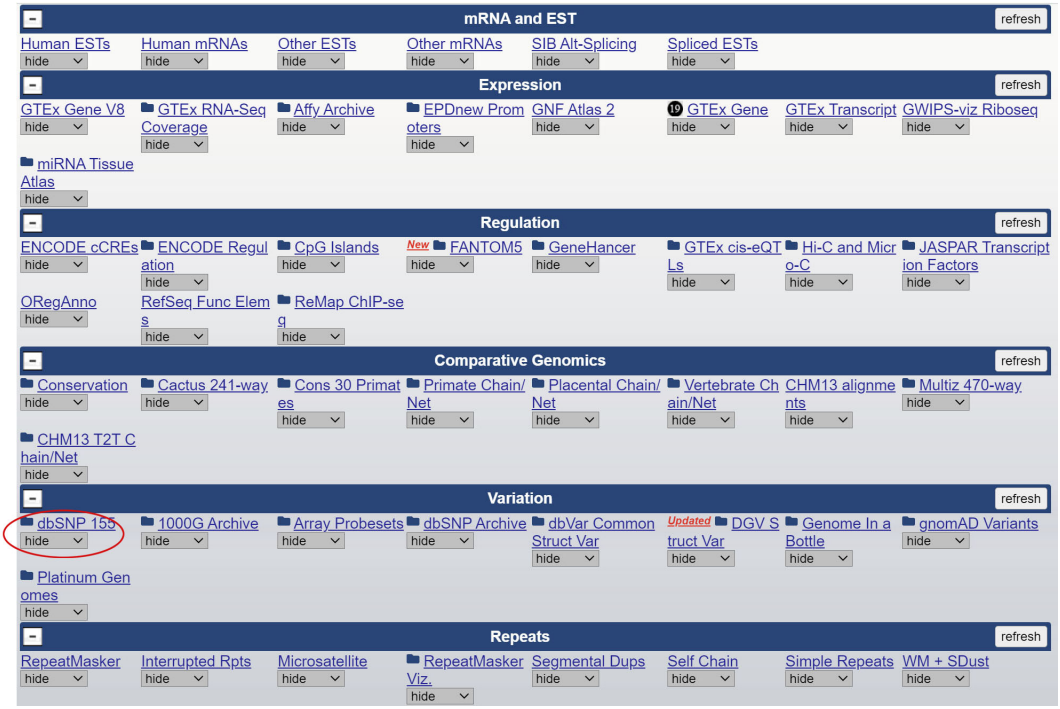


ANSWER: It's because the region shown is in the centromere.

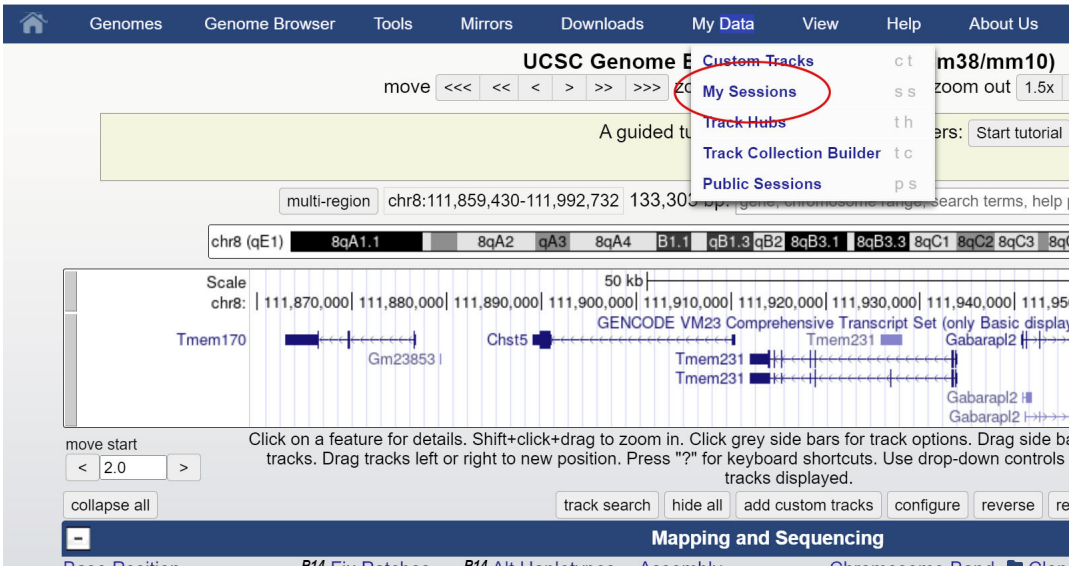
Question 10. Explain why the definition of SNP depends on a reference population.

ANSWER: Because to be a SNP it each version has to occur in at least 1% of the population.

Question 11. Circle the track you would use to find changes in coding sequence of genes that do not affect the amino acid.



Question 12. In the genome browser (as shown), which of the menu items (under "My Data") would you click on to save your configuration and to create a link you can use to share it?



Question 13. An SSH Client is a piece of software that: (circle the one correct answer)
(A) Let's you emulate UNIX on your Mac or Windows machine.
(B) Let's you connect remotely to another UNIX machine. ⟵ **THIS ONE**
(C) Gives you access to your native UNIX operating system.
(D) Layers one operating system on top of another in a secure fashion.
(E) For connecting a UNIX backstation to a web superstation.

Question 14. True of False. If one operating system is layered on top of another, then they both have access to the same file system.

ANSWER: True.

Question 15. Suppose a file `file.txt` has at least 10 lines. Give a sequence of UNIX commands that will make a file `newfile.txt` that has two tab delimited columns, the first of which has the first ten lines of `file.txt` and the second of which has the last ten lines of `file.txt`. *Hint: You may need multiple commands and temp files*

ANSWER:
```
head file.txt > temp.1
tail file.txt > temp.2
paste temp.1 temp.2 > newfile.txt
```

Question 16. What can you conclude if the following two commands return the same answer? (*note: here the dollar sign is just the prompt*)
```
$ sort file.txt | wc -l
$ sort -u file.txt | wc -l
```

ANSWER: It means there are no duplicate rows in the file, they are all different.

Question 17. Suppose `file.fa` is a fastA file of sequences. Construct a grep command to find any sequences in `file.fa` that consist of 100% (capital) A's and no other characters. And there must be at least one A on the row.

ANSWER:
```
grep ^A+$ file.fa
```

Question 18. Suppose `file.fa` is a fastA file of sequences. Construct a UNIX command to return just the sequence IDs of all sequences in `file.fa` that contain five (capital) G's in a row. *hint: you will have to pipe more than one basic command together*

ANSWER:
```
grep -B 1 GGGGG file.fa | grep ">"
```

Question 19. Suppose `file.fa` is a fastA file of sequences. Construct a UNIX command to return just the sequences in `file.fa` that contain five A's upstream of five G's but have no N's or T's between them.

ANSWER:
```
grep AAAAA[^NT]*GGGGG file.fa
```

Question 20. Write down two different optimal alignments corresponding to the trace-back in the filled-in Needleman-Wunch table.

		G	C	A	T	G	C	G
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

ANSWER:

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

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

Question 21. Fill in two squares in the Needleman-Wunch table. Matches score $+1$, mismatches score -1 and indels score -1. Also draw in the appropriate arrows.

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

Question 22. True or False. $(n+2)^3 = O\left(\frac{n^3+8}{n}\right)$.

ANSWER: False. Because

$$\lim_{n\to\infty} \frac{(n+2)^3}{(n^3+8)/n} = \lim_{n\to\infty} n\frac{(n+2)^3}{n^3+8} = \infty \cdot 1 = \infty.$$

Question 23. Suppose you have two DNA sequences each of length ten. How many global alignments can you make between them that have no indels?

ANSWER: One.

Question 24. Suppose the score for a match is +1, the score for a mismatch is X and the score for an indel is Y. Suppose the score for alignment (1) shown below is −2 and the score for alignment (2) shown below is −3. What are X and Y?

$$\begin{array}{ccccc} A & G & G & T & C \\ A & C & G & - & C \end{array} \tag{1}$$

$$\begin{array}{ccccc} A & G & - & T & C \\ A & G & G & - & C \end{array} \tag{2}$$

ANSWER: Write down the two equations:

$$3 + X + Y = -2$$

$$3 + 2Y = -3$$

Solve the second one for $Y = -3$, plug into the first equation to get $X = -2$.

Question 25. How many sequences of length 10 can we make that are 80% A/T and 20% C/G?

ANSWER: $\binom{10}{8}2^{10}$. You first choose the 8 A/T locations, and for those locations there are 2^{10} possible sequences with A/T in those 8 locations (and C/G in the other 2). You can choose 8 locations in $\binom{10}{8}$ ways.