University of Pennsylvania

BIOL4536 Fall 2023

Professor: Gregory R. Grant

Exam#1

Yet More Practice Problems

Prob/Stat

Question 1.

Suppose you have a test for a disease that has a high false-positive rate of 0.3 (so 30%). Suppose you test a population of 100 individuals, all of which have the disease. How many false positives do you expect?

Answer: _____

Question 2. Suppose you want to do a T-test to test whether gene G is differentially expressed between two experimental conditions C_1 and C_2. Write down the null hypothesis.

Answer: _____

Question 3. Is it possible that the null hypothesis is true and the p-value equals 0.00007?

Answer: _____

Question 4. How do you increase the power of a T-test?
(A) Lower the p-value.
(B) Increase the number of replicates.
(C) Decrease the false-positive rate.
(D) Assume normal distributions have mean zero.

Question 5. Write down the equation of a regression model with regression curve equal to a cubic.

Answer: _____

Question 6.

Suppose you have a test for a disease that has a false-negative rate of 0. Suppose you test a population of 100 individuals, 20 of which have the disease. Is it possible that 85 test negative?

Answer: _____

Question 7. Suppose you want to do a standard T-test. The individual measurements are assumed to follow what kind of probability distribution?

Answer: _____

Question 8. Suppose a p-value for a null hypothesis is 0.07 and as a result we do not reject the null hypothesis. True or False, the probability that the null hypothesis is true is 0.93.

Answer: _____

Question 9. Explain why we estimate parameters with confidence intervals instead of exact values.

Regression

Question 1. Write down the equation of a regression model with regression curve equal to a cubic.

Answer: _____

Question 2. Is it possible that the inferred regression line is not equal to the true regression line?

Answer: _____

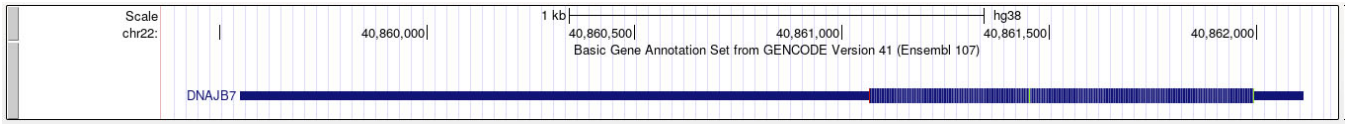
Question 3. Inferring a regression line involves:
(A) Maximizing something
(B) Minimizing something
(Circle the correct answer)

Question 4. Circle the ones that are true (could be one, more than one, or none).
(A) The regression line can be parallel to the X-axis?
(B) The regression line can be parallel to the Y-axis?
(C) The regression line can be a straight line with a negative slope?
(D) The regression line can enter the 4th quadrant.

Question 5. Circle the correct answer. In regression if $\sigma = 0$ (the variance of ϵ is zero) then: (A) All data points are the same.
(B) All data points lie on a vertical line.
(C) All data points lie on a horizontal line.
(D) All data points lie on the regression curve.

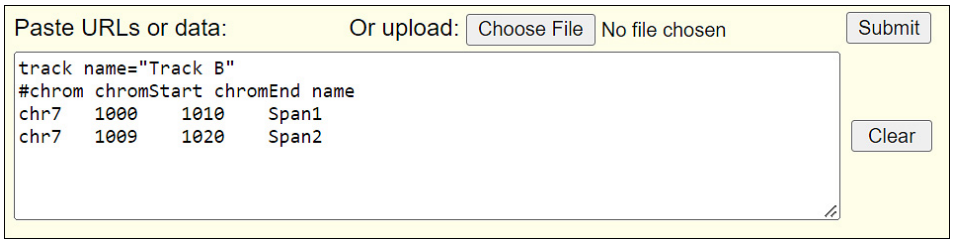
Genome Browser

Question 1. This is a gene with one exon. How can you find out which strand it is on? There are at least three different possible answers.



Answer:

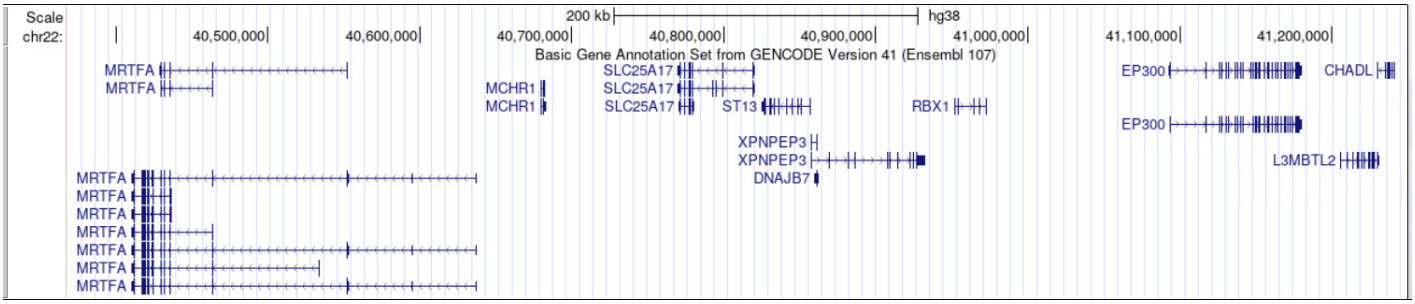
Question 2. This is the configuration of a custom track in the genome browser. How many nucleotides do Span1 and Span2 overlap?



Answer: _____

Question 3. Describe briefly how would you use the genome browser and UNIX to figure out whether chromosome 10 has more gene isoforms (in total) than chromosome 11. Answer in just a few sentences, you don't need to give the details of what fields you have to fill out or the exact UNIX commands, just the general procedure as best as you can describe it concisely.

Question 4. How may one-isoform genes do you see in this genome browser graphic?



Answer: _____

UNIX

Question 1. What would be the consequence of the following UNIX command?

```
ggrant@workstation:~$ tail -20 data.txt > head
```

Answer:

Question 2. Write down a UNIX grep command that will find all lines with two or more *consecutive* A's in a file named data.fa.

Answer: _____

Question 3. Write down the shortest UNIX command that will take you directly to your home directory.

Answer: _____

Question 4. Write down a UNIX command to count the files in the current directory, including the hidden files.

Answer: _____

Question 5. What does the UNIX `gzip` command do?

Answer:

Question 6. What does the following UNIX command do?

```
ggrant@workstation:~$ cut -f 5 UCSC_mouse_knowngene_id_mapping | sort -u | wc -l
```

Answer: _____

Question 7. Write down the grep command to search for sequences in a file named `data.txt` that have an A somewhere before a B but does not have a B anywhere before an A.

Answer: _____

Question 8. Construct a UNIX command to return the lines from `file.txt` that start with a number and end with a lower case letter.

Question 9. What does the UNIX tar command do?
Answer:

Question 10. Write down a UNIX command to count the (non-hidden) files in the current directory that do not end in the letter 't' lower case *or* capital.

Question 11. Write a UNIX command to find lines in file.txt with exactly one A.
Answer: _____

Question 12. What does the following do?
```
> grep ^A.*B.*C$ file.txt
```

Question 13. True or False, the following will match any string: `> grep B* file.txt`

Question 14. Simplify the following grep command
```
> grep ^A.*B*.*C$ file.txt
```

Answer: _____

Question 15. Rewrite the following as one grep command without using pipe
```
> grep ^A.*C$ file.txt | grep B
```

Answer: _____

Question 16. Write the grep command that find the lines of `file.txt` that consist of any number of A's followed by any number of C's, and nothing else.

Answer: _____

Question 17. Write the grep command that finds the lines of `file.txt` that consist of at least one A in a row followed by at least one C in a row, and nothing else.

Answer: _____

Question 18. Find all lines of `file.txt` that have the number 100 in them somewhere, (and not as a substring of a larger number like 1000)

Answer: _____

Question 19. Find all lines of `file.txt` that have the numbers 10 and 100 in them somewhere, (where neither are substring of a larger number)

Answer: _____

Question 20. Find all lines of `file.txt` that have an A followed by any string that does not contain an X, followed by a B
Answer: _____

Question 21. Suppose you have a fastA file data.fa where sequences are always on one line. Count the number sequences that have CCCCC in them that also have IDs that start with the letter E.

Answer: _____

Alignment

Question 1. True or False. If you have two sequences of the same length, then there is an alignment with exactly one indel.

Question 2. How many different cells could be the next one filled in for in the Needleman-Wunch algorithm?

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

Answer: _____

Question 3. Consider two sequences `ATC` and `AGC`. Write down all possible alignments between them, where the `A`'s align with each other and the `C`'s align with each other.

Question 4. The first draft of the human genome was accomplished primarily with: (circle the one most sensible answer)

(A) High-Throughput Sequencing (HTS)
(B) Mass-Spectrometry
(C) Next-Generation Sequencing (NGS)
(D) Gene Expression Microarrays
(E) Sanger Sequencing

Question 5. How many sequences of length 20 can we make that are 50% A/T and 50% C/G?

Answer: _____

Question 6. Explain why we use the term "indel" instead of having "insertions" and "deletions"?

Complexity

Question 1. What does it mean for a function $f(t)$ to be $O(1)$, in other words f is big O of 1.
Answer:

Question 2. Let $f(x) = (x^3 + \sqrt{x})^{1/3}$. Then f is big O of which of the following? Circle all that apply.

(A) $\sqrt{x}$
(B) x
(C) x^2
(D) x^3