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: None, you can't have a false positive in a population where everybody is actually positive.

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: H_0: the mean observation from C_1 equals the mean observation from C_2.

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

ANSWER: Yes, if the probability of something is greater than zero then it is possible, however unlikely.

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

Question 5.

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: No. If 20 have the disease and 85 test negative, then at least 5 of the 20 with the disease must test negative. So that's at least five false-negatives, contradiction the assumptoin that the false-negative rate is 00.

Question 6. Suppose you want to do a standard T-test. The individual measurements are assumed to follow what kind of probability distribution?

ANSWER: Normal.

Question 7. 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: False. A p-value tell you nothing about the true-positive rate.

Question 8. Explain why we estimate parameters with confidence intervals instead of exact values.

ANSWER: Because estimates are derived from data and data are variable. The confidence interval allows us to quantify how sure we are that the value is in a certain inverval.

Regression

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

ANSWER: $Y = X^3 + \epsilon$.

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

ANSWER: Yes, the inferred regression line is just an estimate of the true line, estimated from data. As such, it will vary from data set to data set and will rarely if ever be equal to the true line.

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

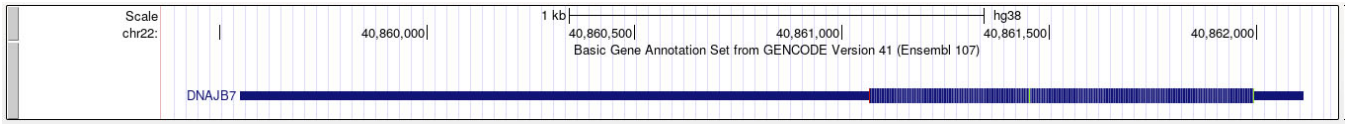
ANSWER: Minimizing. The least squares algorithm involves finding a global minimum of a function of two variables.

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? ⟵ **THIS ONE**
(B) The regression line can be parallel to the Y-axis?
(C) The regression line can be a straight line with a negative slope? ⟵ **THIS ONE**
(D) The regression line can enter the 4th quadrant. ⟵ **THIS ONE**

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. ⟵ **THIS ONE**

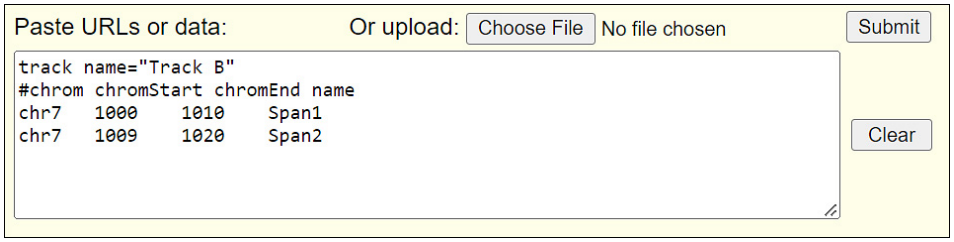
Genome Browser

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: You can click on the gene ID and pull up the gene-info page. Or you can zoom in far enough to see which side has the start codon and which side has the stop codon.

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

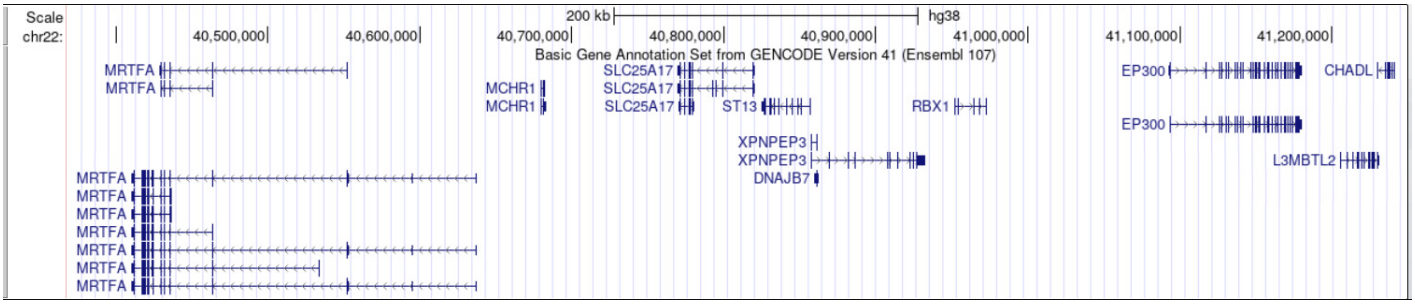


Answer: _____

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.

ANSWER: Go to the table browser and download all transcripts from chr10 to a file. Do the same for chr11. Then do `wc -l` on both files and see which is bigger.

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



ANSWER: Five: ST13 DNAJB7 RBX1 CHADL L3MBTL2

UNIX

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

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

ANSWER: It will take the first 20 lines of `data.txt` and put it in a file named `head`. (You normally would not want to do that).

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:
```
> grep AA data.fa
```

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

ANSWER:
```
cd
```

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

ANSWER:
```
ls -a * | wc -l
```

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

ANSWER: It compresses a text file.

Question 6. What does the following UNIX command do?

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

ANSWER: It counts the number of *distinct* things in column five of the file UCSC˙mouse˙knowngene˙id˙mapping.

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:
```
> grep A.*B data.txt | grep -v B.*A
```

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.

ANSWER:
```
> grep ^[0-9].*[a-z]$ file.txt
```

Question 9. What does the UNIX tar command do?

ANSWER: It packs a number of files into one file, for each transport or archiving.

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.

ANSWER:
```
grep ls * | grep -v t$ | grep -v T$
```

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

ANSWER:
```
> grep A | grep -v A.*A file.txt
```

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

ANSWER: Find all lines in `file.txt` that start with an A, end with a C and have B somewhere in between.

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

ANSWER: True. `B*` matches any number of `B`'s including zero of them.

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

ANSWER: It's the same as:
```
> grep ^A.*C$ file.txt
```

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

ANSWER:
```
> grep ^A.*B.*C$ file.txt
```

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:
```
> grep ^A*C*$ file.txt
```

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:
```
> grep ^AA*C*C$ file.txt
```

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:
```
> grep -w 100 file.txt
```

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:
```
> grep -w 100 file.txt | grep -w 10
```

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:
```
> grep A.*B file.txt | grep -v A.*X.*B
```
Or `> grep A[^X]*B file.txt`

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:
```
> grep -B 1 CCCCC data.fa | grep ">" | grep -c ^E
```

Alignment

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

ANSWER: False

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: Three

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.

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

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

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

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 ⟵ **THIS ONE**

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

ANSWER: $\binom{20}{10}2^{20}$. You first choose the 10 A/T locations, and for those locations there are 2^{20} possible sequences with A/T in those 10 locations (and C/G in the other 10). You can choose 10 locations in $\binom{20}{10}$ ways.

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

ANSWER: We use indels when comparing extant sequences from different species, because we do not know if the ancestral species had the base which was subsequently deleted, or did not have the base and it was subsequently inserted.

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: It means it is bounded in the limit, or more specifically there a constant C and an $N \in \mathbb{N}$ (the natural numbers) such that $f(x) \leq C$ for all $x \geq N$.

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 ⟵ **THIS ONE**
(C) x^2 ⟵ **THIS ONE**
(D) x^3 ⟵ **THIS ONE**