BIOL4536, Fall 2023

Exam#2 (Practice) (SOLUTIONS)

Question 1. True or False. BLAST always returns the optimal (highest scoring) local alignment of a query sequence against a database of sequences.

ANSWER: False, it is a heuristic algorithm, so comes with no guarantees.

Question 2. True or False. RNA-Seq reads can align ambiguously to multiple places in the genome but DNA-Seq reads cannot.

ANSWER: False, both can be ambiguous to the genome.

Question 3. In BLAST, why do we need the expected score to be negative? Choose the one correct answer.

(A) Because otherwise the score of random alignments increases as the sequences get longer. ⟵ **THIS ONE**

(B) So that the optimal aignment score could be negative.

(C) So that the BLAST random walk eventually converges to zero.

(D) This is a trick question and we do not actually need the expected score to be negative.

Question 4. (Choose the one correct answer) The null hypothesis for a BLAST p-values is

(A) The database consists of randomly generated sequences. *This one is not the null hypothesis, it is now the null is modeled.*

(B) The expected score of any alignment between the query and a database sequence is zero.

(C) The query sequence is generated according to the background probabilities.

(D) The query sequence is not related to any sequence in the database. ⟵ **THIS ONE**

Question 5. (Choose the one correct answer) The BLAST index we discussed in class:

(A) Maps words to their exact occurrences in sequences in the database.

(B) Maps words to alignments to words of the same length in sequences in the database whose alignment scores exceed a fixed threshold. ⟵ **THIS ONE**

(C) Maps sequences in the database to their alignment scores with the query.

(D) Maps sequences in the database to their reverse complements.

(E) Maps words in the query sequence to the set of all possible words of the same length that they align to with scores exceeding a fixed threshold.

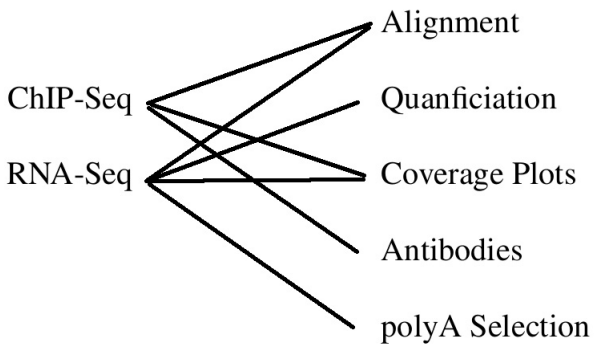
Question 6. Explain how a SAM file bitflag encodes multiple pieces of binary information into one number.

ANSWER: Any number represented in binary is just a sequence of zeros and ones, so each position in the binary expansion can represent a true/false value. The number can then be represented in base-10 for compact represention of the number.

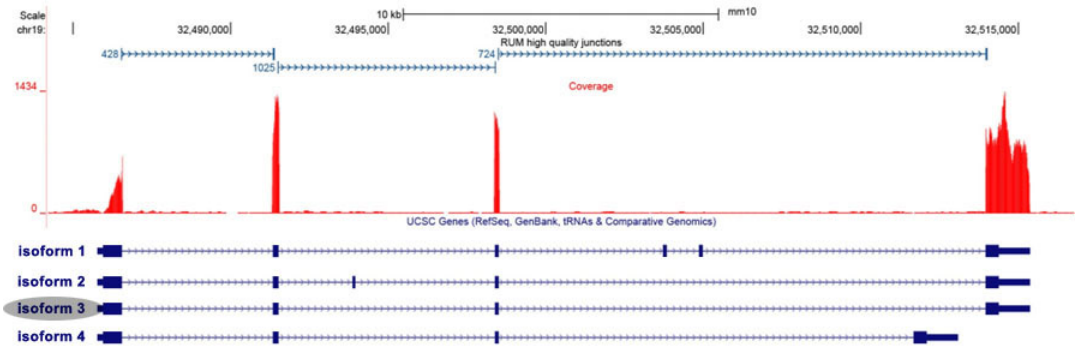
Question 7. (Choose the one correct answer) The FPKM normalization normalizes for:

(A) Depth of sequencing and read length.
(B) Depth of sequencing and gene length. ⟵ **THIS ONE**
(C) Gene length and read length.
(D) Number of replicates and number of conditions.
(E) Length of the genome and depth of sequencing.

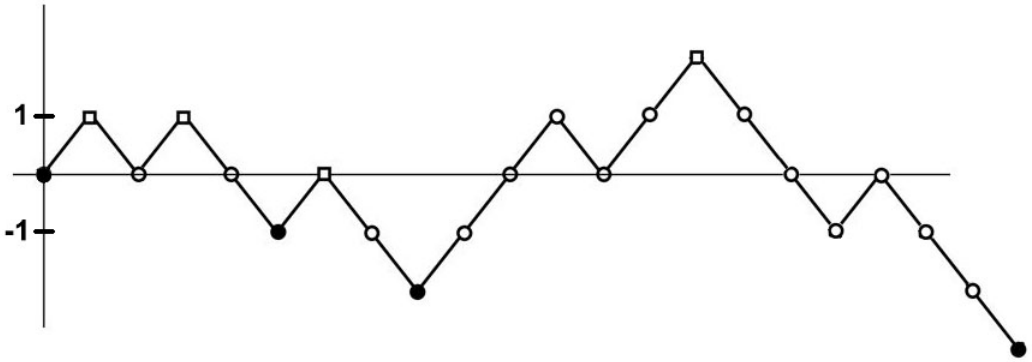
Question 8. Which things on the left go with which things on the right? Draw lines to connect them.



Question 9. Consider the following coverage plot and junctions track. Which of the four isoforms is expressed? Circle all that apply.



Question 10. What is size of the maximal excursion in the following BLAST random walk that has a score of +1 for a match and −1 for a mismatch:



ANSWER: Four.

Question 11. Why do we perform polyA selection in RNA-Seq? And why does polyA selection cause a 3' bias in the signal across the transcripts?

ANSWER: To deplete the ribosomal RNA which is otherwise the majority of the RNA in the sample. It causes 3' bias because the polyA tail is on the 3' end so broken transcripts will capture the 3' end and create preferential signal.

Question 12. Suppose we are using RNA-Seq to perform a differential expression analysis. Suppose there are exactly 30,100 genes and exactly 100 of them are differentially expressed. Suppose we use a p-value cutoff of 0.01 to call a gene significant and suppose all 100 of the truly differential genes have p-value below this cutoff. Calculate what we expect the proportion of false-positives is in the set of all significant genes.

ANSWER: So we expect $30000x0.01 = 300$ false-positives. Thus we'll have 400 DE genes with 100 true-positives. So the proportion of false-positives is 75%.

Question 13. Consider an alignment of a read to the genome with the following start location and CIGAR string.

chr9 1025050 75M1000N25M

Does the read align to the following genome location?

chr9:1025201

ANSWER: No, it alignes up to 1025124 and then skips 1000 bases, so skips over 1025201.

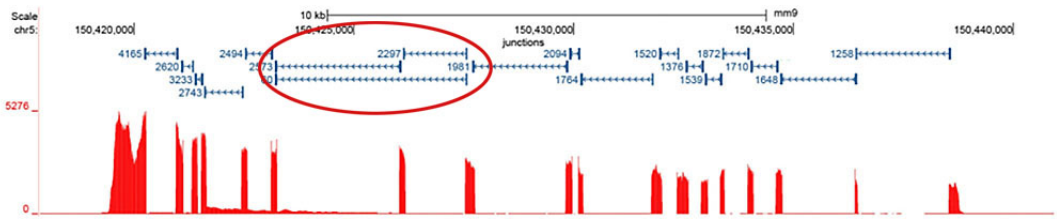
Answer: _____

Question 14. In a fastQ record, like the one shown below, what does the fourth line represent?

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

ANSWER: It represents the quality scores. Each is the quality score for the corresponding base of sequence.

Question 15. The following is a coverage plot and junctions track of RNA-Seq data for one gene. The gene annotation is not shown. Explain why there must be at least two isoforms expressed.



ANSWER: The circled part shows junctions that are not consistent with one isoform, because they both include and exlude an exon.

Question 16. Refer to the same graphic as in Question 15. Assume there are exactly two expressed isoforms. Which of the following is true? Choose the one correct answer.

(A) Both isoforms are expressed at the same level.
(B) The shorter isoform is expressed roughly 40 times higher than the longer one.
(C) The longer isoform is expressed roughly 40 times higher than the shorter one. ⟵ **THIS ONE**
(D) It cannot be determined which is higher, even approximately.

NOTE: The junction depth of the shorter one is 60 and the longer one that includes the exon has junction depths 2573 and 2297. So that's about 40 times higher.

Question 17. Give one reason why, when performing RNA-Seq, there can be real signal visible across the introns of genes.

ANSWER: Because a small percent of RNA is still in the pre-RNA state when the sample was taken, and pre-RNA still includes the introns.

Question 18. How many sequences are in the motif described by this sequence frequency logo? Write down one of them that is neither the most nor the least likely.

ANSWER: There are three positions that have two possibilities. So a total of $2^3 = 8$ sequences are described. One that is neither most nor least common is: ACGATTTC because it includes one position that is the more common and another that is the less common.

Question 19. Suppose the raw count for a gene in a sample is 30. Suppose there are 12,000,000 read-pairs and the gene has length 4,000 bases. What is the FPKM normalized value?

ANSWER: 30/12/4 = 0.625.

Question 20. True or False. An RNA-Seq read can span multiple exons.

ANSWER: True.

Question 21. True or False. You can have an RNA-Seq read such that no part of it aligns to any exons of any gene.

ANSWER: True, it could align entirely to an intron.

Question 22. True or False. In the quality string for a read in a fastQ file, a capital letter indicates lower quality than a punctuation mark.

ANSWER: False.

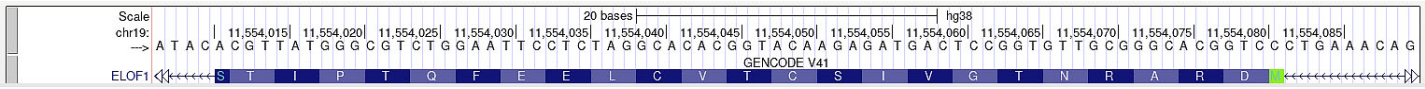
Question 23. Suppose you have a fastQ file of data that you want to Bowtie align it to a specific chromosome of a particular species for which you have the genome sequence. What do you have to do first before you can perform the actual alignment step?

ANSWER: You need to build an index.

Question 24. Consider the alignment of a 100 base read given by this CIGAR string:

chr19 11554011 31M260N69M

One exon of the ELOF1 gene is shown below. True or False. The alignment given by the CIGAR string is consistent with having come from the mature (introns removed) ELOF1 gene? Note, the exon has length exactly 71 bases.



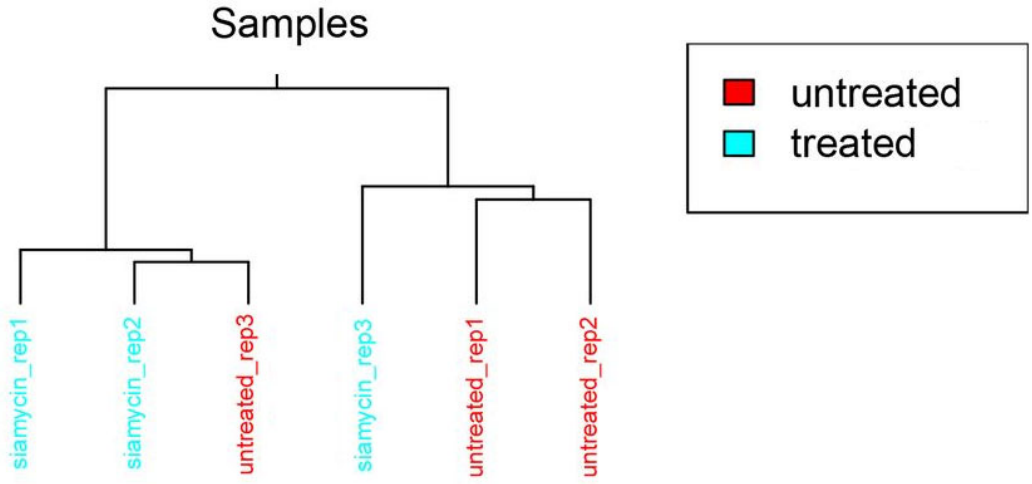
ANSWER: False, it's not consistent, The CIGAR string indicates an intron starting 31 bases from the start position of the exon, not 71 bases.

Question 25. True or False. If you zoom in enough on a coverage plot track in the UCSC Genome Browser then it displays the individual reads.

ANSWER: False, the IGV genome browser can do that but not UCSC.

Question 26. In a ChIP-Seq assay coverage plot, the interesting locations are:
(A) Where the valleys are.
(B) Where the peaks transition into the valleys, and vice versa.
(C) Where the peaks are. ⟵ **THIS ONE**
(D) It depends, it's where the peaks are in TF binding ChIP-Seq and where the valleys are in histone-modification ChIP-Seq.

Question 27. What does the following clustering of RNA-Seq samples indicate?



ANSWER: It indicates two samples may have accidentally been switched.

Question 28. Suppose we're testing 10,000 genes for differential expression and only ten of them are actually differentially expressed. How small does the p-value cutoff for significance have to be in order for us to expect that at least half of the significant genes are true positives? *Note: "significant" here means its p-value is below the cutoff.*

ANSWER: We need there to be no more than 10 false positives, so we need $9990 * p < 10$ So p is approximately 0.001.

Question 29. In the histone mark **H3K79me2** what does the "me2" mean?

ANSWER: it means the 79th kinase on the histone taile is *double* methylated.

Question 30. Explain the FPKM normalization, why we do it and how we do it?

ANSWER: We do it to normalize for read depth and gene length. In each sample, we divide each quantification by the number of millions of reads and by the number of thousands of bases of gene length.

Question 31. True or False. It is not possible to perform an RNA-Seq assay to sequence mRNA and short RNA ($\leq$ 50 bases) at the same time.

ANSWER: True.

Question 32. True or False. PolyA selection in RNA-Seq retains coding RNA (mRNA) and discards everything else.

ANSWER: False, there are plenty of polyadenylated non-coding RNAs.

Question 33. True or False. For DNA-Seq we have to fragment the molecules because chromosomes are so long; while for RNA-Seq fragmentation is not necessary because molecules are so short.

ANSWER: False, even RNA are too long to sequence without fragmentation.

Question 34. True or False. The gapped alignments that BLAST reports in its output are generated by Smith-Waterman.

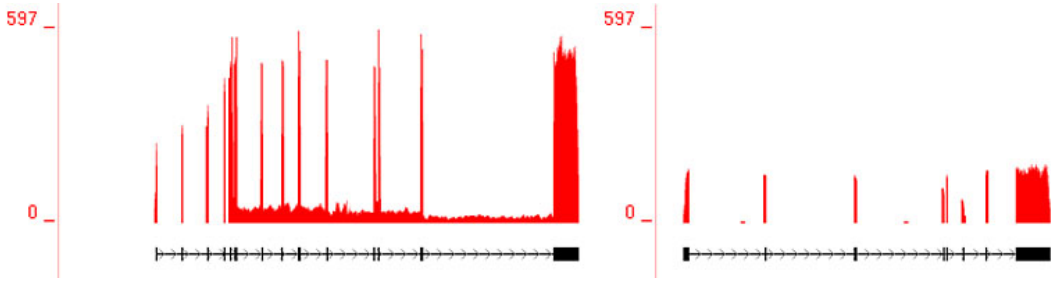
ANSWER: True.

Question 35. The "two-hit" improvement to BLAST was done to make it:
(A) have more accurate p-values.
(B) be less greedy.
(C) run faster and use less resources. ⟵ **ThiS ONE**
(D) have negative expected score.

Question 36. True or False. You need a different BLAST index for each substitution matrix.

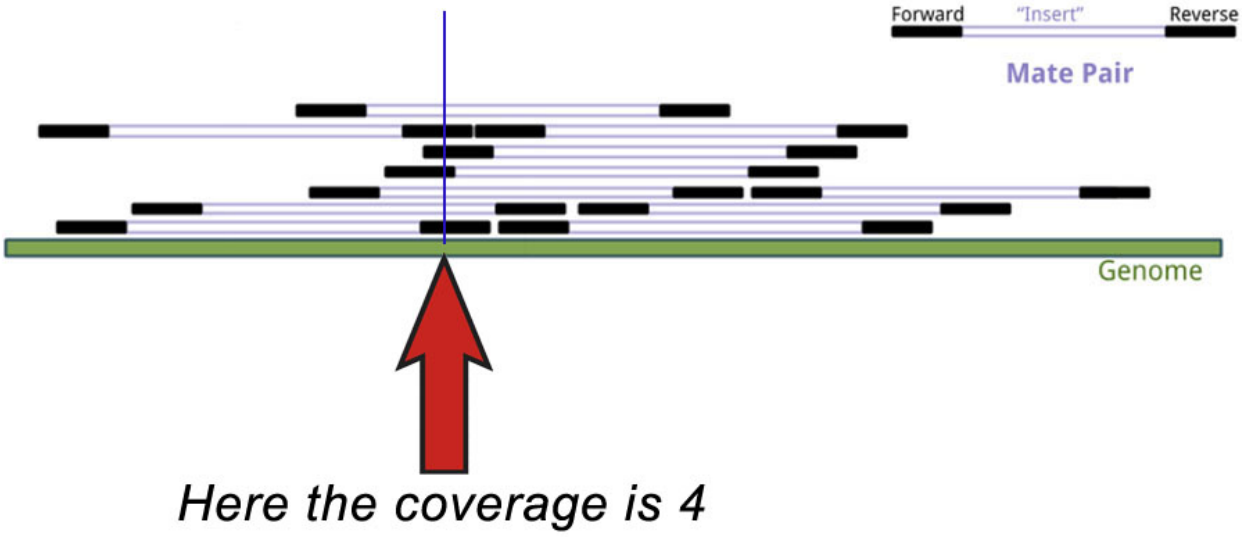
ANSWER: True.

Question 37. Shown below are coverage plots of two different genes from *two different* RNA-Seq assays. Give one reason the gene on the left might not actually be expressed higher than the one on the right.

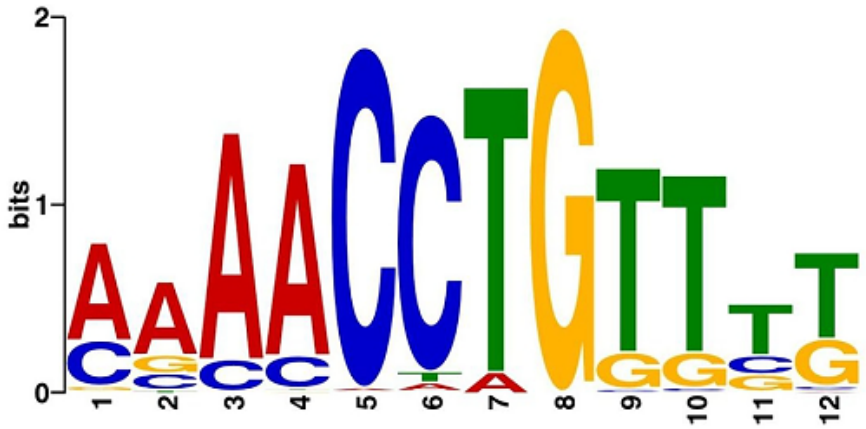


ANSWER: Because the one on the left might be from a sample that was sequenced much deeper than the sample for the gene on the right.

Question 38. When making a coverage plot from the data depicted below. What is the maximum depth of coverage?
ANSWER: Four



Question 39. What is the most frequent motif (or sequence) for a transcription factor with the following sequence frequency logo?



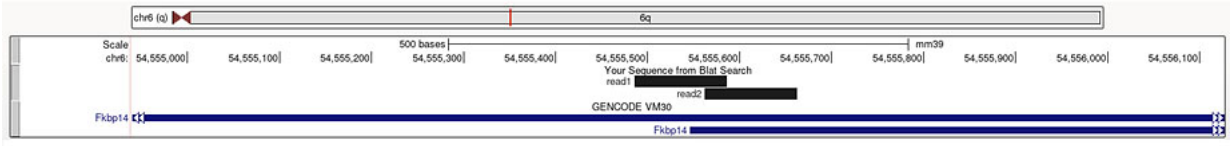
ANSWER: AAAACCTGTTTT

Question 40. True or False. In high-throughput sequencing, generating paired-end data allows us to sequence the fragments in their entirety.

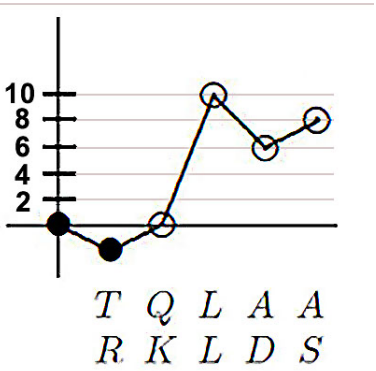
ANSWER: False, we may still miss some part of the fragement between the two reads.

Question 41. This genome browser graphic shows reads aligned to the gene FKBP14. Which the following are true? Circle all that apply.

(A) Both of these reads could align to the top isoform. ⟵ **THIS ONE**
(B) Both of these reads could align to the bottom isoform.
(C) The top read could align to the top isoform while the bottom read could align the bottom isoform. ⟵ **THIS ONE**
(D) The top read could align to the botton isoform while the bottom read could align the top isoform.



Question 42. This is a protein BLAST random walk. What's the score in the substitution matrix for the A/D substitution?



ANSWER: -4

Question 43. True or False. BLAST only reports ungapped alignments (no indels).

ANSWER: False.

Question 44. True or False. BLAST works with any substitution matrix.

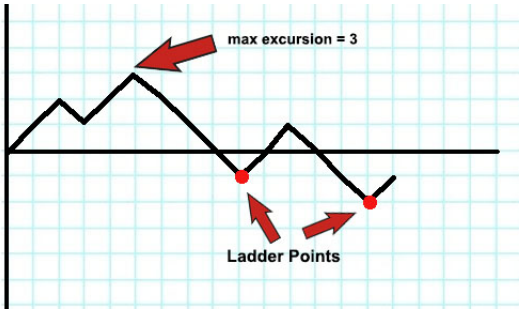
ANSWER: False, the expected score must be negative.

Question 45. Suppose we have an ungapped alignment between two DNA sequences. Suppose the score for any match is $+1$ and any mismatch is -1. Consider the BLAST random walk for this alignment. Suppose the max excursion is 12. True or False, the max excursion can occur at position 10 in the alignment.

ANSWER: False, since matches only score $+1$ there hasn't been enough bases to get to $+12$.

Question 46. Suppose in a random walk heads are one step up and tails are one step down. Steps of the random walk are represented by the horizontal axis and cumulated step is represented on the vertical axis. Starting at the origin, draw in the random walk determined by the sequence of coin flips. Also identify the ladder points and the size of the max excursion.

$$H, H, T, H, H, T, T, T, T, H, H, T, T, T, H$$



Question 47. Why don't we consider an aligment with BLAST p-value of 0.01 as significant. Pick the one correct answer.

(A) Because of all of the simplifying assumptions used to model alignment. ⟵ **THIS ONE**
(B) Because of multiple testing.
(C) Because we never consider a p-value of 0.01 significant in any situtation.
(D) It's not true, we do consider a BLAST p-value of 0.01 as significant.

Question 48. What does it mean that a histone modification is involved in "activation"?

ANSWER: It means when the mark is present in the locus of the gene then the gene tends to be expressed and when it's not there it tends not to be expressed.

Question 49. What are the two applications of ChIP-Seq that we discussed in class?

ANSWER: Histone Modification and Transcription Factor Binding Site Identification.

Question 50. In the following BLAST alignment, what do the plus signs mean?

RecName: Full=H-2 class II histocompatibility antigen, A-D beta chain; Flags: Precursor [Mus musculus]

Sequence ID: P01921.1 Length: 265 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 41 to 227 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
296 bits(759)	8e-102	Compositional matrix adjust.	140/187(75%)	161/187(86%)	1/187(0%)

```
Query  1    ECYFTNGTERVRLLTKYIYNREEYVRFDSDVGEYRAVTELGRPWAEYWNPQKDEMDRVRA  60
            ECY+TNGT+R+RL+T+YIYNREEYVR+DSDVGEYRAVTELGRP AEYWN Q + ++R RA
Sbjct  41   ECYYTNGTQRIRLVTRYIYNREEYVRYDSDVGEYRAVTELGRPDAEYWNSQPEILERTRA  100

Query  61   ELDTVCRHNY-GLEELTTLQRRVEPTVTISPSRTEVLNHHNLLVCSVTDFYPGQIKVRWF  119
            E+DT CRHNY G E  T+L+R  +P V IS SRTE LNHHN LVCSVTDFYP +IKVRWF
Sbjct  101  EVDTACRHNYEGPETSTSLRRLEQPNVAISLSRTEALNHHNTLVCSVTDFYPAKIKVRWF  160

Query  120  RNDQEQTAGVVSTPLIRNGDWTFQILVMLEMTPQRGDVYTCHVEHASLQSPITVQWWPQS  179
            RN QE+T GV ST LIRNGDWTFQ+LVMLEMTP +G+VYTCHVEH SL+SPITV+W  QS
Sbjct  161  RNGQEETVGVSSTQLIRNGDWTFQVLVMLEMTPHQGEVYTCHVEHPSLKSPITVEWRAQS  220

Query  180  ESAQSKM  186
            ESA+SKM
Sbjct  221  ESARSKM  227
```

ANSWER: Those are the substitutions that have positive scores in the substitution matrix.

Question 51. Suppose the CIGAR string for an alignment of a read to the genome is. Note: Recall that "D" means deletion, "I" means insertion and that is with respect to the genome (in other words deleted from or inserted in the genome). "S" means skip that many bases in the read.

25M475N32M2D4I3S

How long is the read?

ANSWER: 25 + 32 + 4 + 3 = 64 (count everything but N's and D's).

Question 52. Give one reason we might cluster RNA-Seq samples after quantifying.
ANSWER: For quality control purposes to indentify possibly mislabled samples. Or also to see if the experimental conditions do separate in the clustering.

Question 53. Why is gene-level quantification easier than isoform level quantification?

ANSWER: Because different isoforms tend to share many exons, so a read mapping to those exons is ambiguous as to which isoform it came from. But all isoforms are the same geme so it doesn't matter for gene level quantification.

Question 54. Suppose the genotype of a human at a SNP location is heterogenous A/G. Suppose we perform DNA-Seq and the read-depth at the SNP location is 5. Suppose each read has 50% probability of being maternal and 50% probability of being paternal. What is the probability that the reads all indicate the same nucleotide, thus leading to an erroneous (homozygous) genotype call at this location?

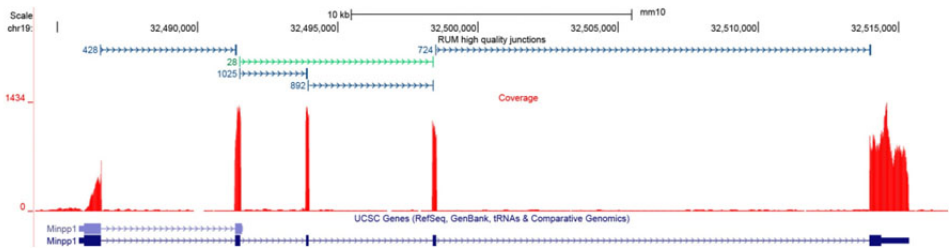
ANSWER: The probability that they're all A's is $\frac{1}{2^5}$ and the same for them being all G's. So the probability that one of those things happened is $\frac{1}{2^5} + \frac{1}{2^5} = \frac{1}{2^4} = \frac{1}{16} = 0.0625$

Question 55. Assume the human genome is 3.6 billion nucleotides. Suppose we are performing 100-base single-end read DNA-Seq. What is the average depth-of-coverage if you have 360 million reads all fully aligned?

ANSWER: 360 million reads is 360,000,000×100 equals 36 billion bases. So that's 10-fold coverage.

Question 56. Suppose a random walk is based on a coin flip, step +1 for a head and −1 for a tail. Suppose the coin is biased with 40% probability of a head. The long term behavior of the random walk is:
(A) It will drift off to negative infinity. ⟵ **THIS ONE**
(B) It will converge to zero.
(C) It will drift off to positive infinity.
(D) None of the above.

Question 57. Consider the following coverage plot and junctions track. Is this data explained completely by the gene annotation that is displayed? Explain why or why not.



ANSWER: No, the green junction indicates an isoform that skips an exon which is not reflected in the annotation.

Question 58. Why do we deplete ribosomal RNA when we perform RNA-Seq?

ANSWER: Because a large percentage of the RNA in a sample is ribosomal RNA, which is just four different trascripts. One does not want to waste all that sequencing real-estate on four transcripts.

Question 59. True or False. An RNA-Seq read can align to the genome in a location where there is no gene annotated.

ANSWER: True. Gene annotation is not complete.

Question 60. Describe what kind of search blastX is for.

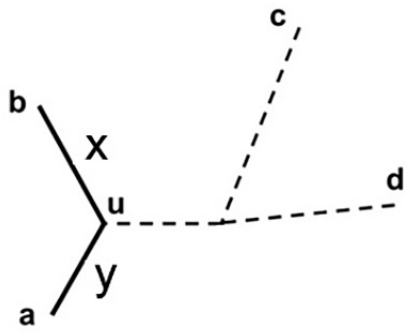
ANSWER: It's for a DNA query that will be translated into protein and aligned against a protein database. It will be aligned in all six possible reading frames.

Question 61. Suppose X is a discrete random variable that takes finitely many values $x = 1, 2, \ldots, n$ and has probability function $p_X(i)$. Show using the formula for expected value that $E[2X] = 2E[X]$.

ANSWER: $E[2X] = \sum_{i=1}^{n} 2ip_X(i) = 2\sum_{i=1}^{n} ip_X(i) = 2E[X]$.

Question 62. What is $X - Y$? *Hint: consider* $d(b, c) + d(b, d) - d(a, c) - d(a, d)$.

	a	b	c	d
a	0			
b	5	0		
c	9	10	0	
d	9	10	8	0



ANSWER: Looking at the figure, $d(b, c) + d(b, d) - d(a, c) - d(a, d) = 2X - 2Y$. From the distance matrix the left hand side is $10 + 10 - 9 - 9 = 2$. So $X - Y = 1$.

Question 63. Why should we not use RNA-Seq for genotyping?

ANSWER: Many reasons. First off we'd only get genotypes mainly at exons. And also all expression could be from one chromosome so the variants on the other chromosome won't produce any reads and homogenoous calls will be made which should be heterogeneous.