Written test: Analysis of data from high-throughput molecular biology experiments BB2490 (BIO) or DD2399 (CSC)

Name: ______________________________Pnr: ____________

Wednesday the 18th of February, 13.00-15.00, Albanova FB52

Instructions: The test consists of 20 multiple choice questions (1 point each). To pass the test you should have the correct answer for 12 of the 20 questions. Write your **answers on this front page**. You are only allowed to bring writing material (pen, pencil, eraser, ruler) to the test.

Question	A	B	C	D
1	□	□	□	□
2	□	□	□	□
3	□	□	□	□
4	□	□	□	□
5	□	□	□	□
6	□	□	□	□
7	□	□	□	□
8	□	□	□	□
9	□	□	□	□
10	□	□	□	□
11	□	□	□	□
12	□	□	□	□
13	□	□	□	□
14	□	□	□	□
15	□	□	□	□
16	□	□	□	□
17	□	□	□	□
18	□	□	□	□
19	□	□	□	□
20	□	□	□	□

Question	**Your score**	**Maximal Score**
Sum of Multiple choice questions 1-20		20 points

Questions

Mark your answer on the front page by putting a cross (X) in **one** alternative. There is only **one** correct alternative.

1. What does it is mean to say a feature is generated under null hypothesis?

   A) It is a probability that the feature stems from a null/alternative hypothesis.

   B) That the feature most likely stems from the conditions in the null hypothesis

   C) The feature was drawn from a probability distribution stipulated by the null hypothesis.

   D) The feature is assigned as incorrect.

2. What is a p-value?

   A) The probability of obtaining a result equal to or more extreme than what was actually observed.

   B) The probability that a the null model is true, given what was actually observed.

   C) The probability that the same result would be observed if the experiment was repeated a second time.

   D) The probability of obtaining a result equal to or more extreme than what was actually observed, given that the null hypothesis is true.

3. "It is said that PEP (local FDR) should be used when looking at the quality of a particular read-out, but why can one not compare this to single hypothesis testing?" Which of the following statements is **not** a correct answer to that question?

   A) As there are more than one hypothesis tested in a high-throughput experiment, we need to make multiple-hypothesis corrections

   B) A posterior probability is more directly interpretable, as it reflects the probability of a particular read-out to be incorrect.

   C) A p-value is hard to interpret when there are more than one hypothesis tested with an experiment.

   D) Single hypothesis tests are harder to calculate, as we then test a larger hypothesis.

4. You want to extract the set of differentially expressed proteins in a high-throughput assay involving 40 000 distinct proteins. You note that at your preselected threshold of a 1% FDR, corresponding to a p-value threshold of 0.001, find 2000 differentially expressed proteins. Which of the following statements is **not** correct?

   A) You expect 20 of the reported differentially expressed proteins to be false positives.

   B) You have estimated $\pi_0 = 0.5$.

   C) You expect 1980 of the proteins reported to be differentially expressed to really be differentially expressed. You just do not know which ones of the 2000 proteins with p-values below the threshold.

   D) You expect 400 of the reported differentially expressed proteins to be false negatives.

5. Which of the following statements about mass spectrometry-based proteomics is **not** correct?

   A) We frequently encounter situations where an uniquely defined peptide does not uniquely identify a protein.

   B) It is more reliable to identify a peptide by its mass than by its fragmentation spectrum

   C) Electrospray is an ionization technique.

   D) An extracted ion chromatogram illustrates the ion current of a certain mass-to-charge rate over time

6. Which one of the following techniques is **not** an example of a *labeling* technique used for protein quantification:

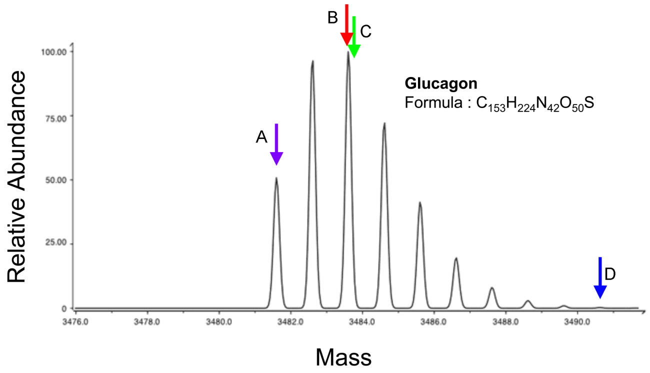
   A) SILAC

   B) TMT

   C) iTRAQ

   D) Spectral counting

7. Which one of the arrows in the figure below points at the monoisotopic mass of the examined analyte?



   A) Arrow A

   B) Arrow B

   C) Arrow C

   D) Arrow D

8. Which of the following techniques is **not** used for targeted proteomics assays?

   A) Selected reaction monitoring

   B) Data-independent acquisition

   C) Shotgun proteomics

   D) SWATH

9. Which one statement is correct?

   A) A gene may give rise to many different transcriptomes

   B) A gene may give rise to many different transcripts

   C) Transcription at a gene always starts at one and the same nucleotide

   D) Non-coding RNA is always translated into protein.

10. In the base quality score system that, e.g., Illumina sequencing technology uses, a base quality is given to every base in every read. Typically, the score is higher in the beginning of the read, and then it goes down towards the end of the read. A score of 30 means that, at that position, the probability of an incorrect base is 1 in 1000. What does a score of 20 mean?

    A) The probability of an incorrect base is 1 in 500

    B) The probability of an incorrect base is 1 in 2000

    C) The probability of an incorrect base is 1 in 10000

    D) The probability of an incorrect base is 1 in 100

11. Dynamic range is the difference in “signal strength” between the highest expressed transcript and the lowest expressed transcript in a sample. The true dynamic range in a cell is typically on the order of 1:1,000,000 which means that the most abundant transcript is present in approximately one million times more copies than the least abundant. Microarrays have the disadvantage that their dynamic range is limited and they cannot represent the full dynamic range present in the sample. A typical RNA-seq experiment has a better dynamic range. But what is the “signal” in an RNA-seq experiment? In other words, what is a possible measure of transcript abundance? (One alternative is correct).

    A) RPKM

    B) PCR duplicates

    C) Reference genome

    D) Number of different k-mers in the transcript

12. In the analysis of differentially expressed genes, the final outcome is, for each gene, an effect size and a p-value. Which one of the following statements is **not** correct regarding differential expression analysis of two samples, X and Y, using RNA-seq?

    A) The p-value is the probability that the observed difference (or larger) in counts or RPKMs would occur by random chance for that gene

    B) Multiple testing correction should be performed

    C) For a gene with two isoforms A and B, it is possible that isoform A is more highly expressed than isoform B in sample X, but that isoform B is more highly expressed than isoform A in sample Y

    D) The effect size is defined by the read length used in the RNA-seq

13. “Epigenetics is generally understood to be the study of heritable regulatory changes that do not involve any changes in the DNA sequence of a cell” is a quote in the lecture notes. There are a number of possible epigenetic changes or modifications, but which one of the following is **not** an example of an epigenetic modification?

    A) Methylation of DNA

    B) Histone acetylation

    C) Transcription factor binding to DNA

    D) Histone methylation

14. In the analysis of a ChIP-seq experiment, one central aim is to find the significant "peaks". Which one statement is true about ChIP-seq "peaks"?

    A) A peak is a significant enrichment of reads in a region

    B) A peak is a transcription factor binding motif

    C) Transcription factor binding site peaks are typically much wider than histone modification peaks

    D) The number of peaks detected is, within +/- 1%, the same in all ChIP-seq experiments.

15. You have produced two genome assemblies, A and B, on the same dataset. Measured on scaffolds, N50 for A is 500 and for B is 550. Which statement is correct, if your goal is to localise genes in the genome?

    A) A is the better assembly and is the one to analyse.

    B) B is the better assembly and is the one to analyse.

    C) Both are good and contigs/scaffolds are likely big enough to contain genes, so it does not matter which one I use.

    D) The difference in N50 is not enough to guarantee that one is a better assembly than the other, but the contigs/scaffolds will probably not be good enough in either assembly.

16. You are frustrated by bad assembly results on your Illumina data (50X coverage of paired end reads, average read length 100 bp) from a 1 Gbp genome when you hear claims that CABOG, a modernisation of the Overlap-Layout-Consensus assembler that Celera used for the human genome, is giving very good results. What is a correct response?

    A) "Ah well, that wont work on my data."

    B) "Hm, that wont work on my data. I am going to order new sequencing with the latest chemistry: that will give me a read length of 150 bp, which should suffice!"

    C) "Allright, I will just concatenate my paired reads and then I should be able to try it out."

    D) "Allright, I should really try it to see how it works on my data!"

17. The quality assurance tool FRC requires you to map reads to each assembly. Both BWA and Bowtie are good choices for the mapping, and both requires you to first build an *index*. How many indices do you need to build if you are analysing the assemblies from question 16?

    A) None. I will just download the standard index for the human genome.

    B) One, because I build the index for the reads.

    C) Two, because I need one for each assembly I want to test.

    D) Three. One for the reads and one each for the assemblies A and B.

18. Suppose you have a reference genome for species S and Illumina reads (single-end, average read length 150 bp) from an individual in the same population as S was assembled from. Which statement on read mapping software is correct?

    A) I cannot expect all my reads to be mapped, even though they are high-quality Illumina reads.

    B) As long as the reads are longer than 100 bp, they should all map.

    C) As long as the reads are shorter than 100 bp, they should all map.

    D) They should all map since the genome and the reads are from the same species.

19. Long DNA reads from the PacBio instrument have quickly become really popular for genome assembly. Why?

    A) Because they are highly suitable for genome assemblers based on de Bruijn-graph technology.

    B) Because the reads have much higher accuracy than Illumina reads.

    C) Because the reads can span through repeat regions and generally provide reliable long-range information.

    D) Because you can get a whole chromosome in one read.

20. What is *scaffolding* in the context of genome assembly?

    A) It is the use of read-pairs or mate-pairs to connect contigs.

    B) It is about selecting primers that extend contigs, with the hope that directed reads will span to another contig.

    C) It refers to picking bacterial artificial chromosomes (BACs) based on your contig content, performing shotgun sequencing on the BAC, and assembly of the reads. This way, you can efficiently "fill" holes in your assembly.

    D) It is about using the Lander-Waterman theory to choose an appropriate read coverage in order to ensure that there are not gaps in the final assembly.