BTK6002 Introduction to high-throughput data analysis, exam 16.10.2019

Juha Kesseli, juha.kesseli@tuni.fi

No calculators are allowed, only normal writing equipment. Notice that there are two pages in this question sheet. You can take this question sheet with you when you leave the exam.

1. For the following statements a-f, indicate whether the statement is true or false. You don't have to write down your justification. Each correct response is worth +1p, each incorrect response is $-\frac{1}{2}$p, and each statement with no response is 0p.
   a) Samtools pileup is commonly used as an input to variant calling.
   b) RNA-seq data analysis always starts with alignment to the reference genome.
   c) In UNIX, each user always belongs to exactly one group determining access rights to files.
   d) A SAM file can be looked at in a text editor.
   e) `/home/studentx/project_work` is an example of a relative path in UNIX.
   f) Each hard drive contains at most one file system.

2. a) You are working on a project analyzing a set of very large BAM files. You have a tool you want to use that takes SAM files as input and writes its output in SAM format as well. How would you use the tool to minimize the disk space you need and the amount of disk access needed in total? (2p)
   b) When doing somatic variant calling from DNA-seq data, using what information can we say that a variant is likely to be somatic? How is your answer affected by the availability of control sample(s)? (4p)

3. a) Which two types of quality scores are commonly used to assess high-throughput sequencing data? (1p) Based on what are the quality scores calculated? (1p) How can the information in the scores be used in data analysis? (2p)
   b) Let's say we have a DNA-seq read that contains three bases that differ from the corresponding bases in the reference genome. What needs to happen in Bowtie2 so that the correct output alignment is reported? (2p)

4. After running DESeq2 on an RNA-seq dataset comparing treated against untreated samples, let's look at the following selected lines from the full results table:

```
log2 fold change (MLE): condition treated vs untreated
Wald test p-value: condition treated vs untreated
DataFrame with 9921 rows and 6 columns
          baseMean log2FoldChange lfcSE        stat        pvalue           padj
         <numeric>      <numeric> <numeric>    <numeric>   <numeric>        <numeric>
A00289 3448.7461824  0.110747348  0.1028871186  1.25670257 0.1789385013      0.548228989
A03360 4342.8325483 -3.179541222  0.1435677118 -22.1466869 1.12413406e-108 2.35540147e-105
A25111 1501.4475742  2.899945154  0.1273575983  22.7707952 9.07163384e-115 3.80147901e-111
A29167 3706.0248547 -2.196911924  0.0979153989 -22.4362856 1.72094428e-111 4.80599633e-108
A31805 3108.8876967 -0.855539896  0.1598068928 -5.35353953 8.62424066e-08  3.61343155e-06
A39140 815.18243968 -0.121151851  0.0984977392 -1.22997196 0.211518555      0.611489459
A39155 730.56767595 -4.618742392  0.1691239519 -27.3097433 3.24789243e-164 2.71917165e-160
A51005 310.91833583 -0.544145221  0.1358684464 -4.00499507 6.20194764e-05  0.001297458
```

a) Out of these eight genes, which one(s) would you call significant? (1p)

b) What is the padj column in the output table? (1p)

c) Why should we look at the padj column if pvalue already gives us the result of the statistical test? (2p)

d) Are there any genes on the list for which the expression in the treated samples is estimated to be less than $\frac{1}{8}$ of that in the untreated ones? If so, which one(s)? (2p)

5. Let's say we are interested in studying how the transcription factor SOX2 is regulating gene expression in a set of brain cancer samples. Which types of measurements can be performed to study the question? What kind of data analysis is needed? How can we assess the results from our analysis i.e. what other information should we compare the results with? How can we combine the information from the different data sources? (6p)