Examination in MVE510 - Introduction to bioinformatics FACIT
Time: Friday January 17 2018, 08.30-12.30
Examiner: Erik Kristiansson
On call: Henrik Imberg, phone 5325
Aids: None

Maximum is 40 points. Grade 3 requires 18 points, grade 4 requires 26 points and grade 5 requires 34 points.

1. Assignment 1

   (a) Pre-processing aims to remove sequencing errors and other artifacts from the data. These are introduced in the sequencing process, often at a higher rate than traditional Sanger sequencing. Pre-processing will thus to improve the overall quality of the data, which is necessary to avoid errors and misinterpretations in the down-stream analysis.

   (b) One method used for pre-processing is the trimming of reads. Since many sequencing techniques (e.g. Illumina) has a lower quality in the end of the reads, the last nucleotide can be removed ('trimmed') to increase the overall quality of the read. The trimming procedure typically works by removing nucleotides until the quality is sufficiently high, e.g. on the quality of the following nucleotide is above a preset cut-off. A drawback of this approach is that the trimmed reads will be shorter in length.

2. Assignment 2. Two alignments, both with a score of 18

```
TA-TTT
TACTTG

TATTT-
TACTTG
```

3. Assignment 3

   (a) Germline mutations are inherited by the offspring from its parents. These mutations are therefore present in all cells in the human body. Somatic mutations are mutations that appear over time and are only present locally and often in a small subset of the cells. Germline mutations are the main focus in this study.(2p)

(b) As always, the data is first pre-processed. After that, the reads are aligned against a reference sequence, e.g. the full human genome. Since the reads will be relatively close to the reference, the alignment is often done with fast and less sensitive index-based methods such as BWA. After that, the alignment is quality assessed. This can include recalculation of the quality scores and removal of duplicates. Regions with high dissimilarity can also be re-aligned with more sensitive methods. Next, mutations are called. This is typically done using a suitable statistical methods with tests whether the difference in nucleotide distribution between the reference and observed data are caused by sequencing errors or by an underlying mutation. Finally, the list of putative mutations often needs to be quality assessed and artifacts are removed. Sometimes, the list of mutations is also filtered based on their biological function (synonymous/nonsynonymous, location within the gene, the function of the gene etc).

4. Assignment 4

(a) Gene quantification of metagenomic data is done as follows. For guided binning, the sequence reads are aligned to a reference. This can include a collection of microbial genomes, a gene catalogue or contigs that have been assembled from the data. The reference needs to be annotated so the start and end of the genes are known. The alignment is typically done with relatively sensitive aligners to that differences in strains of the organisms in the community are taken into account. Next, the reads are grouped based on what gene they match. This can be done using both at a more higher functional level or at a more specific gene level. The gene abundances is then estimated based on how many reads that ends up in each group. For direct binning, a reference is not used. Instead the genes are identified directly on the reads, which are then binned based on the gene. This requires however reads of sufficient length.

(b) One statistical approach is to model the gene abundance by using discrete overdispersed count models. This approach will describe the nature of the data accurately but a drawback is the complexity of these models and that they need to be fit using numerical methods which can make them slow. Also, modeling the dependence between the expected value and the variance is non-trivial. Another approach is to transform the data to a normal distribu-

tion and use standard methods such as linear models. This can be done using a variance stabilizing transform such as the square-root and the logarithm. An advantage is that methods based on the normal distribution are plentiful, flexible and fast. A drawback is that the transformation will never be perfect - or even good - for all genes. The normal assumption will therefore be not be valid which results in lower statistical performance. A final option is to use non-parametric methods, which does not rely on an underlying distributional assumption. This makes them robust but, as a consequence, they require more sample to reach a sufficiently high statistical power. Non-parametric methods are also sensitivity to many ties, which may be common in low-abundant count data.

5. Unsupervised analysis of data.

   (a) In supervised data analysis, there is a predefined organization of the samples that is utilized in the analysis. For example, if samples are collected from sick and health individual (e.g. samples for RNA-seq) we typically perform a supervised analysis where we compare the genes between the pre-defined groups 'sick' and 'healthy'. In contrast, unsupervised analysis we do not take advantage of any pre-defined structure of the data. Instead, we use analysis methods to see if the samples are grouped or organized according to a specific pattern. One example where unsupervised analysis can be used is to see whether samples taken from the tumour of a set of individuals separate into distinct sub-groups or if they are all from one large group.

   (b) Hierarchical clustering builds a hierarchy between clusters using a pre-defined distance metric. The most common form is the agglomerative clustering where each observation starts in each own clusters, which are then successively merged into larger clusters. The merging of clusters, as a function of the distance metric, can be visualized as a dendrogram. The merging of clusters can be done using approaches such as complete linkage and single linkage. The number of formed cluster will vary depending on the similarity cut-off. k-means clustering aims to divide the observations into k separate clusterings. In contrast to the hierarchical clustering, k is here pre-specified. This is done forming the k-clusters in a way that each observation is a member of the cluster for which it has the shortest distance to its mean.

6. The Burrows-Wheeler transform (BWT) is often used for comparing sequence reads against a reference.

   (a) The BWT enables fast lookup of exact matches. One the BWT is transformed, the complexity does not depend on the length of the reference. This makes it highly efficient when matching multiple reads to the same reference. In comparison to suffix trees, which also have these features, the BWT is of the same size as the reference and can thus be efficiently stored in the memory of a computer. Finally, the BWT only needs to be calculated once for a reference sequence and can then be used for matching all of the reads.

   (b) $GTCAAT

   (c) This can be done by observing that the BTW is the predecessor of the sorted BTW. Thus, by iteratively concatenating the BTW to the sorted BTW the original sequence can be recreated (demonstration of the reconstruction is necessary for full points).

7. RNA-seq is sometimes used to find mutations. The advantage is that this enables both analysis of gene expression and identification of various DNA alterations from the same data. Note that this is a discussion question and the answers below are only suggestions.

   (a) Sequencing can be done similar to an standard RNA-seq experiment. A commonly used platform is Illumina. Note however that the short read length of Illumina will make it harder to identify larger alterations. Also, the sequencing depth will depend on the mRNA abundance of the genes. The coverage, and thus the ability to identify mutations, will vary between genes. If we want to be able to find mutations in the majority of the genes, the sequencing depth must be enough to also result in a high coverage of the low abundant genes.

   (b) The analysis can be done as follows. First the data is pre-processed to remove bad reads. Next, the reads are aligned to a reference. If we are using the human genome, a fast index-based aligner should be satisfactory. Note that the aligner should be spice-aware in order to align reads over exon junctions. Note that we will have low coverage for gene with low expression and quality assessment and removal of artifacts is therefore important in order to avoid false positives. Next, mutations can be called using similar approaches

as used in whole genome sequencing. The identified mutations should be filtered to remove artifacts caused by e.g. duplicates.

(c) This approach will be able to identify SNPs and short indels present in transcribed regions. This includes SNPs that causes a change in the amino acid sequence of the corresponding genes and indels that causes frame-shifts (both have the potential to alter protein function). It is also possible to identify larger structural variations that results in fusion genes. Mutations outside coding regions will however be missed. Also mutations in regions that are transcribed but spliced away (e.g. mutations in intron) will most likely not be identified. Mutations is genes that are low abundant will be harder to find.