SEQUENTIA
DECODE • DISCOVER • DISRUPT

Case Study September 2021
Jr. Bioinformatician candidates

Dear candidate,
Thank you for agreeing to move forward in the selection process.
You have **7 days** to send us the results for this case study. The results of the exercises, the reasoning used and the presentation of the results will be evaluated.

Exercise 1

using the attached file **Variants_file.vcf.gz**

1. Which program was used to generate the VCF file? Describe how you obtained this information.
2. What is the size of the reference genome used for the analysis? Describe how you calculated it.
3. Generate a table with the percentage of missing values for each sample. Describe how you generated it.
4. Show the clustering of the samples based on the variants included in the VCF. Describe the method you have chosen and justify it.
5. Generate a new VCF removing the variants with a MAF value higher than 1%. Describe how you generated it.
6. How many variants are left from the previous step? Describe how you calculated it.

Exercise 2

The **counts.txt** file contains raw expression counts from six Arabidopsis thaliana samples belonging to two experimental groups (stress: STR and control: CTRL). The matrix was generated from the TAIR10 reference genome and annotation. With the provided data:

1. Evaluate and show the clustering of the samples with the method of your choice. Describe how you performed this step and justify your method of choice.
2. Perform a differential expression analysis to identify the genes that are significantly different between the STR and the CTRL groups and generate a table as a result. You can apply filters to genes and/or remove samples as you see fit. Describe how to perform this task and justify your choice of methods.

3. Generate a heatmap to show the expression profile of the differentially expressed genes. Describe how you generated it.
4. Generate a volcano plot to visualize the results of the differential expression analysis, use the color red for up-regulated genes and the color green for the down-regulated ones. Describe how you generated it.
5. Perform a Gene Ontology Enrichment Analysis using the differentially expressed genes, generate the output in tabular form and with a graphical representation.

Exercise 3: open question

A researcher is studying a human disease, for this purpose she produced RNA-seq data from multiple patients. Using her data, she identified candidate genes and then produced cellular lines harbouring mutations in those genes. Now she produced RNA-seq data from these cell lines and she would like to understand which cell lines are more similar to her patients using all the transcriptomics datasets. Which approach would you use to complete this task? Explain it providing also a list of the steps that you would perform starting from the raw sequencing data of both the experiments.

Exercise 4 (optional): open question

You are working with an IBS patient from which you have produced shotgun sequencing data from the human exome, shotgun data from the gut microbiome and shotgun data from mouth microbiome. How would you integrate this information with each other? What information and correlation you might find? What support for precision medicine could you give?