MMG1344H: Foundational Computational Biology (FCB) I **Winter 2022**

<u>Assignment 3:</u>

Your clinical collaborator has asked for your help in investigating how well their leukemia patients are responding to treatment. The idea is to measure the effectiveness of the treatment by measuring the extent to which it reduces the population of cells bearing variants that were detected at diagnosis.

In the *input.zip* file, you will find a dataset (provided courtesy of Dr. Sagi Abelson) containing base call summaries (reduced to capture only the somatic variants of interest) for 16 control samples and 24 patients:

1. The control data was obtained by sequencing the blood of healthy individuals.
2. The case data was obtained by sequencing the blood of leukemia patients following treatment.
3. The list of somatic variants of interest that were detected when the patients were diagnosed.

To determine how well leukemia patients respond to treatment, you decide to track the variant allele frequency (VAF) for every variant detected at diagnosis. If the VAF of a variant drops to the background level (i.e. is not significantly higher than the VAF in control samples), you consider that the patient has responded to the treatment.

Using the dataset provided:

1. Generate a single table capturing, for every control sample, the VAF for every variant observed in that control sample. If a particular variant was not observed in a particular control sample, set VAF = 0.

2. For each of the variants that were observed at least once in a control sample, fit an exponential distribution to the set of VAF values for that variant. Capture the fitted rate parameter (λ) obtained by the model for each variant in a new "rate" column. Set the location parameter (μ) to 0. You may use any libraries/packages to fit the distribution.
   <u>Tip:</u> Learn more about the exponential distribution here: *https://www.itl.nist.gov/div898/handbook/eda/section3/eda3667.htm.*
   <u>Output:</u> A comma-separated file (.csv) in which the first column ("variant") should contain the genome position and variant (e.g, chr4 106194031 A), the second column ("rate") should be the fitted rate parameter, and the remaining 16 columns (named control1 to control16) should store VAFs for variants observed in every control.

3. Next, we would like to investigate whether the VAF of each variant detected in each patient at diagnosis has fallen to the background after treatment. One might think that, for a given patient, we could ignore the VAF of variants that were not initially detected at diagnosis. However, being an appropriately-paranoid computational biologist, you want

to first make sure that there wasn't a randomly permuted Excel error along the way, so you decide to check whether patient-variant pairs that were detected at diagnosis are at least somewhat enriched for being significantly above background than patient-variant pairs detected after treatment.

Therefore, for each of the investigated variants (the union of variants detected in any patient sample) and for each patient, assess the significance of departure of the patient variant allele frequency (VAF_p) from the null exponential distribution that was modelled from controls. In other words, for each patient/variant combination, calculate the probability (under the null distribution for this variant) of observing a VAF larger than the patient variant frequency VAF_p, i.e. $P(VAF > VAF_p)$. You may use any libraries or packages to calculate "nominal" p-values (p-values that are not yet corrected for multiple testing).

4. Now correct the p-value for multiple testing using a Bonferroni-type correction, and list the variants that received a corrected p-value which is lower or equal to 0.05. *Hint: think carefully about how many tests you need to correct for.*

   Output: A comma-separated file (.csv) listing the variants that received a corrected p-value of less than or equal to 0.05. In the CSV file, the first column ("variant") should contain the genome position and variant (e.g, chr4 106194031 A), the second column ("sample") should contain the patient sample (e.g. Patient_0107), and the third ("p") and fourth ("p.corrected") columns should contain the original (nominal) and corrected p-values, respectively.

5. Determine whether patient-variant pairs that were identified at diagnosis are more likely than patient-variant pairs to be significantly above background after treatment. Fill in a 2-by-2 contingency table based on two binary variables: 1) whether the VAF for a patient-variant pair is significantly above background after treatment and 2) whether the patient-variant pair was in the list of patient-variant pairs observed at diagnosis. Now use the Fisher's exact test (aka one-tailed hypergeometric test), which has the null hypothesis that identification of a patient-variant pair at diagnosis is independent of whether the patient-variant pair is above background after treatment. Can you determine if the sample labels were randomly swapped? How did you make that decision?
   Tip: Learn more about the Fisher's exact test here:
   https://en.wikipedia.org/wiki/Fisher%27s_exact_test
   Output: A 2 x 2 contingency table and the p-value of the Fisher's exact test. Please also provide your answer to whether the sample labels were randomly swapped and how you made the decision.

6. Regardless of your conclusion above, you decide to move ahead with the original analysis and assume that there were no weird permutations and that your data is OK. Therefore, you now revisit the nominal p-values calculated in Step 3. Now we only need

to consider patient-variant pairs where the variant was detected at diagnosis. Please perform the correction for multiple testing with the updated set of tests, again using a Bonferroni-type correction. List the patient-variant pairs that were observed significantly above the background (with a corrected P-value ≤ .05). What fraction of the patient-variant pairs are no longer significantly above the background.
<u>*Output:*</u> *A comma-separated file (.csv) listing the variants that received a corrected p-value of less than or equal to 0.05. In the CSV file, the first column ("variant") should contain the genome position and variant (e.g, chr4 106194031 A), the second column ("sample") should contain the patient sample (e.g. Patient_0107), the third column ("rate") should contain the rate parameters calculated in Step 2, and the third ("p") and fourth ("p.corrected") columns should contain the original (nominal) and corrected p-values, respectively.*

7. Identify the patients who apparently responded to the leukemia treatment. Here, you can define response as having at least one variant detected at diagnosis in that patient which is no longer significantly above background after treatment. For what fraction of patients was this the case?
   <u>*Output:*</u> *A list of patient labels.*

<u>Submission</u>

Please submit a zip file (first_lastname.zip) with

1. Your source code (in .R or .py format).
2. A README.txt file on how to run your program.
3. Two comma-separated (.csv) files.
4. A document (in .doc or .pdf format) with your answers to questions.