MMG1344H: Foundational Computational Biology (FCB) Course I Spring 2025

Assignment 3

Part 1

Here we use data from a 2014 study of the earliest cell-fate decision in mammalian development (Biase, Cao, and Zhong 2014). This study looked at gene expression in sister blastomeres (each cleaved from the same zygote) by using single-cell RNA sequencing (scRNA-seq). They reported highly reproducible between-blastomere differences among 10 samples of 2-cell stage mouse embryos and 5 samples of 4-cell stage embryos. Between-blastomere gene expression differences appeared to dominate between-embryo differences, and these differences were sufficient to cluster sister blastomeres into distinct groups. For numerous protein-coding genes, reproducibly bimodal expression in sister blastomeres was observed, and this could not be explained by random fluctuations.

You will focus on the scRNA-seq gene expression data (in units of “fragments per kilobase of transcript per million mapped reads”; or FPKM) which covers 6812 genes from 4-cell mouse embryos (20 samples of 4-cell blastomeres) and 2-cell mouse embryos (20 samples of 2-cell blastomeres).

1. Implement the k-means algorithm (do NOT use pre-existing Python or R implementations or packages for this) with **k = 4** and Euclidean distance with standardized data. Test and run your code on the scRNA-seq data provided (*Biase_2014.csv*). You will need to “standardize” your data, i.e., rescale the values such that the mean for all genes (over all 40 samples in the data set) is equal (or very close) to 0, and so that the variance for each sample is equal or very close to 1.
2. The k-means algorithm tends to converge quickly but it may be stuck on local optima, so please run your algorithm 10 times with different initialization conditions (controlled by a random seed) to see if this behaviour exists on your data. For each run, please report:
    a. The random seed.
    b. The cluster size and the identity of the objects (cell samples) in every cluster.
3. Using the best clustering run from Step 2, for each cluster, find all the genes that show enriched or depleted expression in that cluster. To estimate significance (using a *P*-value threshold of 0.05), use a two-sided Mann-Whitney *U* test (aka Wilcoxon rank-sum test). For example, to test each gene in Cluster 1, apply the test to examine whether the expression distribution for that gene is different between Cluster 1 samples and samples in all other clusters. Don’t forget to use a Bonferroni approach to correct your P-values for multiple hypothesis testing (and think carefully about how many tests for which you have to correct). Please report significant genes in each cluster with their Bonferroni corrected P-value.
4. Based on the gene lists, can you identify which 4-cell-enriched cluster is similar to which 2-cell-enriched cluster? Please explain why.

Part 2
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.

   Tip: Learn more about the exponential distribution here: https://www.itl.nist.gov/div898/handbook/eda/section3/eda3667.htm.

   Output: 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 nominal p-value for multiple testing using the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR). Compute the FDR-adjusted p-values (sometimes referred to as *q-values*) and find the variants that are considered significant at FDR ≤ 0.01 and FDR ≤ 0.05, respectively.

   Output: A comma-separated file (.csv) listing the variants with FDR ≤ 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 (“FDR”) columns should contain the original (nominal) and FDR-adjusted p-values, respectively. Include a note in your answer indicating how many patient-variant pairs were significant at FDR ≤ 0.01 and at FDR ≤ 0.05.

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. Apply the Benjamini-Hochberg procedure to compute FDR-adjusted p-values for this subset. Report the list of patient-variant pairs that are significantly above background at FDR ≤ 0.01 and FDR ≤ 0.05. 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 an FDR ≤ 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 ("FDR") columns should contain the original (nominal) and FDR-adjusted p-values, respectively. Additionally, report the fraction of diagnostic patient-variant pairs that are no longer significantly above background at each of the two FDR thresholds (1% and 5%).*

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.*

Submission

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.

References

Biase, Fernando H., Xiaoyi Cao, and Sheng Zhong. 2014. "Cell Fate Inclination within 2-Cell and 4-Cell Mouse Embryos Revealed by Single-Cell RNA Sequencing." *Genome Research* 24 (11): 1787–96.

Vendeix, Franck A. P., Antonio M. Munoz, and Paul F. Agris. 2009. "Free Energy Calculation of Modified Base-Pair Formation in Explicit Solvent: A Predictive Model." *RNA* 15 (12): 2278–87.