Gene Length Technical Exercise – Dashnow Lab

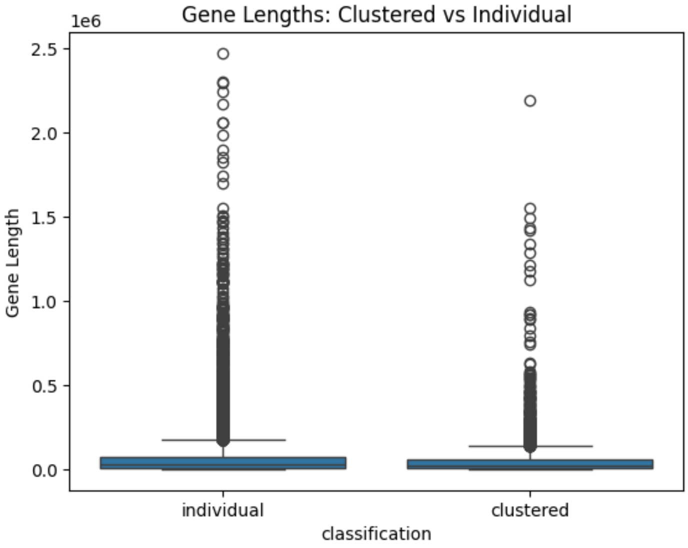
Isha Arora

GitHub Repository: https://github.com/isha-04/TechnicalExercise_DashnowLab

Solution 1

The idea was to consider a gene "clustered" if it is within 1000 bases of the next closest gene and to be "individual" otherwise.

To label each gene as such, I first grouped genes by the seqid (chromosome) and strand. Following this, for each group, I sorted the genes by their start coordinate and then calculated the distance between genes using the end coordinate of one gene and the start coordinate of the next gene. Following this, I checked if the distance between genes was lesser than the threshold and marked both the genes in this pair as "clustered". The others were marked "individual".



To see if length of a gene impacts the likelihood of it being clustered – comparing the differences in lengths between "clustered" and "individual" genes, I ran a non-parametric Mann-Whitney U test where the p-value < 0.001 showing a high significant difference between the median values with the median value of clustered genes being lower.

This led to the assumption that there is a **lower chance of a gene being clustered if the gene is longer**.

Solution 2 *(Can be seen as a Python code in the GitHub Link)*

To identify N closest genes – user input was asked for the specific seqid, start coordinate and end coordinate for the target gene, as well as for the value of N. For all genes with the specific seqid, distances were calculated using the start and end coordinates of the target gene. If genes overlap, the distance between them was considered to be 0. These distances were sorted in ascending order, and the top N values were printed.

Things I would like to do: In Solution 2, I would also like to introduce "strand" as a variable to group the genes by (besides using seqid).