SEED tool uses similarity based clustering on the given dataset in the fastq format. The basic idea behind clustering is to join sequences with maximum 3 mismatches. The virtual centers of the RNA-seq reads are calculated for each cluster. This is essentially done by finding the sequences that lie between twice the mismatch threshold that we have specified and then the center is calculated with the allowed number of overhanging characters. By default the algorithm uses 3 mismatches and 3 overhanging residues from the virtual centers.

We get the output file which contains the virtual centers for each cluster and the seq ids that lie in the centers. Since we have the correct genes that each read came from we can perform the false positive and false negative analysis by varying the clustering parameters like mismatch edit distance and overhanging character threshold.

Preprocessing: This step involved parsing/correting the data given in the fastq format. The SEED algorithm does not take random length reads. One of the constraint of the algorithm is the maximum and minimum lengths of the read should not differ by more than 5. Therefore, some of the reads were ignored for the final input. Other preprocessing of the data file was done. The python scripts for the preprocessing at given in the seed\_tool folder

Result Analysis: SEED tool gives us the final output txt file which containes the clusters, their virtual centers and corresponding RNA-seq read ids. We can find the the false positives and false negatives from these outputs with the help of the reference genome data that was provided to us.

Following are the plots to compare various parameters and FP and FN fractions. As can be observed from the graph, as we increase the allowed mismatch gaps in the RNA-seq reads for clustering, the false positive and false negative rates increase. This goes with the intuition as well since with higher mismatches allowed we tend to become more lenient about the similarity definition and can consider the reads from different genes to be similar and hence cluster then together. Also, total number of clusters formed with 0 mismatch number are almost double than the ones formed with 3 or more mismatched allowed.

Similar analysis was performed by varying other parameters like the shift. The shift parameter basically represents the maximum shifts allowed from the virtual center in each cluster.