Aaron Morales
BME 205
Fall 2025
2074743

Assignment 4

Competency of Modely

In order to assess the efficacy of the genomic overlap script, it's most effective to analyze the results through a number of curated test examples. More specifically subjecting the script to a number of permutation-based overlap tests can help capture the expected statistical behavior and determine whether the results are indicative of a competent model

Sanity Checks

This section will demonstrate the results from a number of simple examples to gauge if the model is working as intended. From a test genome.fai of 1000bp, I created three sets of .bed files, those demonstrating perfect(100bp), partial(50bp), and no(0bp) overlap between the regions.

chrom	start	end	observed_overlap	p_value	bonferroni_p	significant
chr1	100	200	100	0.009900990099009901	0.009900990099009901	True

chrom	start	end	observed_overlap	p_value	bonferroni_p	significant
chr1	150	250	50	0.10891089108910891	0.10891089108910891	False

chrom	start	end	observed_overlap	p_value	bonferroni_p	significant
chr1	900	1000	0	1.0	1.0	False

Above here we see the results per region csv for perfect, partial, and no overlap. These values are well within expected with ~0.01 for total overlap, 0.11 for partial overlap, and 1.00 for no overlap. These values reflect the chances of overlap by random permutation, which is expected for ~100 bp overlap in a 1000bp genome.

P value and Bonferroni values

These tests are well and good to show that the model is competent enough, but these singular region .bed files don't demonstrate the bonferroni p value or the statistical significance in a larger dataset. With another synthetic example consisting of setA.bed and setB.bed of 50 and 5 regions respectively within a genome.fai of 10000bp, when permuted 1000 times, we get the following results.

Aaron Morales
BME 205
Fall 2025
2074743

metric	value
observed_overlap	200.0
global_p_value	0.001
num_permutations	1000.0
setA_regions	50.0
setB_regions	5.0
setA_total_bases	200.0
setB_total_bases	600.0
bonferroni_threshold	0.01
significant_regions_bonferroni	1.0

chrom	start	end	observed_overlap	p_value	bonferroni_p	significant
chr1	1000	1200	200	0.000999000999000999	0.004995004995004995	True
chr1	2000	2100	0	1.0	1.0	False
chr1	3000	3100	0	1.0	1.0	False
chr1	4000	4100	0	1.0	1.0	False
chr1	5000	5100	0	1.0	1.0	False

Above shows the respective results.tsv and results_per_region.tsv files for this synthetic example. With a p_val threshold of 0.05, the bonferroni threshold should be 0.05 / nB(5 regions in this case) = 0.01. We can see that the only example with significant overlap(200) gives a p value of 0.001 and a correct bonferroni value of 0.005, which is below the threshold and is statistically significant in this nB = 5 permutation test.

Conclusion

These permutation based overlap tests are able to correctly capture the expected statistical behavior of our test genomic regions. The sanity checks demonstrate appropriate p values for relative overlap lengths and the synthetic dataset was capable of determining statistical significance via bonferroni correction adjusting the p value relative to the number of set B regions. Alongside the results alone, the per-region permutation framework is robust in its construction of randomized null distribution corrected for chromosome bounds, ensuring that observed significance reflects spatial co-localization opposed to permuted chance proximity. Finally, these results validate both the implementation statistical efficacy of the algorithm, demonstrating its predictable behavior across my variety of tests!