6.3732 PSet 2 Part 1

Stephen Andrews

February 28, 2025

Collaborators: Sophia Chen, Samir Kadariya

2.1:

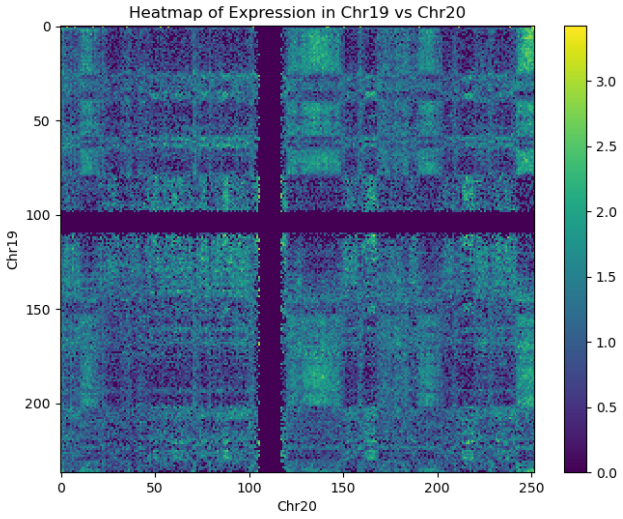
(a)

After iterating through all 253 interchromosome, matrices, we get that the overall mean and standard deviation of the log transformed frequency levels are $\mu = 0.7019$ and $\sigma = 0.6309$.

Note that a few design decisions were made when getting these values. Specifically, whenever our frequency at a specific location was labeled as NaN, we assume that this is a nonzero number whose value was lost in data processing. As a result, we impute its value as the average of the frequency of the other locations in that interchromosome matrix. In addition, the full length of each chromosome was not listed anywhere explicitly in the data, so for chromosome i its length is the maximum location listed across all interchromosome matrices it is apart of.

(b)

After turning our transformed sparse matrix into a dense matrix, we get the following heatmap:



There are 4 main quadrants that are denoted very clearly on this heatmap. Then within each quadrant there are less clearly defined regions with high average interaction value. These less specified regions in each quadrant can be characterized as the rectangles with a high concentration of yellow cells signifying high levels of interaction.

(c)

The number of submatrices, $N_{\text{submatrices}}$, can be calculated by selecting any set of contiguous rows and any set of contiguous columns. Since the 19-20 matrix is 237 by 252, picking a starting point and endpoint for the rows and columns gives us

$$N_{\text{submatrics}} = \binom{238}{2} \cdot \binom{253}{2}.$$

Now we interpret the expression for the p-values provided. Under the iid Gaussian null, the mean of all entries in the $k \times l$ submatrix follows an $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{kl}})$ distribution. Therefore, the probability we observe our mean, m, or a greater value of our submatrix gives us a p-value of

$$1 - \Phi\left(\frac{(m-\mu)\sqrt{kl}}{\sigma}\right).$$

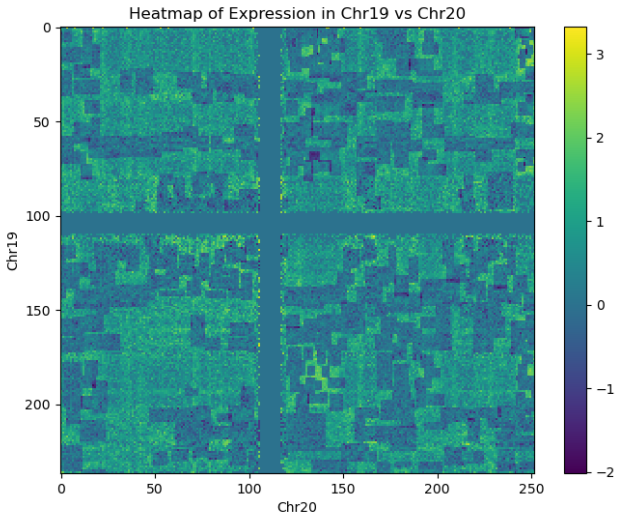
However, since we are in the multiple hypothesis testing case, Bonferroni Correction on the above p-value by the number of tests we are doing gives us the desired expression:

$$N_{\text{submatrics}}\left(1 - \Phi\left(\frac{(m-\mu)\sqrt{kl}}{\sigma}\right)\right).$$

(d)

Greedy Search works by trying to expand our submatrix to incorporate more and more rows/cols of the interaction matrix until we reach a row or column that renders our entire submatrix no longer statistically significant with respect to the Null hypothesis. Intuitively, what is happening on our heatmap from part (b) is we are picking a random starting point and adding adjacent rows or columns until the new larger submatrix is no longer "yellow enough" because the yellow cells denote entries that are uncommon under the null.

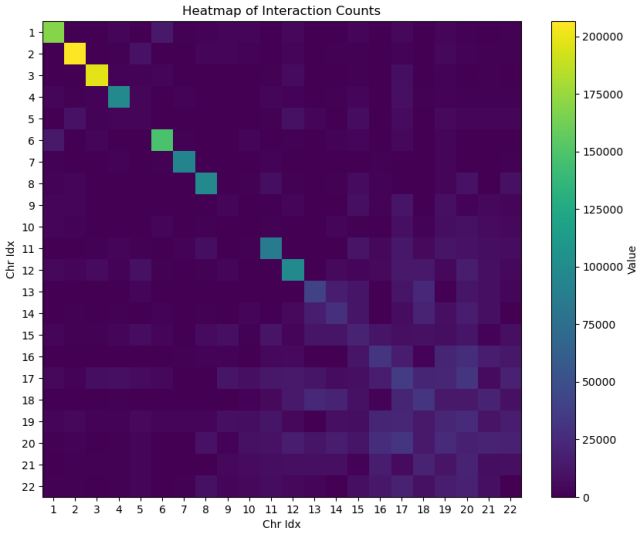
After performing the described procedure to identify interacting regions, we get the following modified heatmap:



The modified heatmap shows how the procedure attempted to put patches over areas of high interactions to allow future iterations to find other areas of high interaction. The interacting regions are where the "patches" are.

(e)

Running the procedure from the previous part on all pairs of chromosomes, we get the matrix below. In addition, note that our stopping condition of 0.01 will be scaled down by the number of times we run greedy search on each iteration as a form of Bonferroni Correction.



As we would expect, there are higher levels of interaction on the diagonal. In addition, as we get to higher chromosome numbers we can see that intra chromosome relationships start to fade, and the inter-chromosome relationships start to shine through.

(f)

To obtain a three-dimensional embedding of the chromosomes, we use MDS because it preserves the pairwise distances from the interaction matrix, ensuring that chromosomes are positioned in 3D space while maintaining their relative distances. To create a logical distance matrix, the first thing we address is that larger interactions signify more similarity between two chromosomes, so we want this to indicate a smaller distance. Therefore we apply the following monotonic transformation to do this mapping:

$$D_{i,j} = \frac{1}{1 + M_{i,j}}$$

where D is our distance matrix and M is our matrix of interaction counts from part (e).

After performing MDS, we get the following embedding into 3D: