Huiting Su

Stephanie Oprescu

STAT598: Final project

**Proposed project:**

We had proposed using published HiC data (which maps regions of the human genome that are found to interact in a cell) to generate an interactive 3D map of the nucleus, which would use both the interchromosomal data as well as the intrachromosomal data.

In order to do so, we had planned to use the raw data which contains the sequencing reads and map those reads to the genome using Bowtie, filter out reads below a certain depth, filter out repetitive reads, and then generate interaction matrices for all the chromosomes in a single interchromosmal matrix and also for each chromosome to generate intrachromosomal interaction matrices for each chromosome. Subsequent analysis was going to consist of generating heat maps for the interchromosomal interaction matrix and a couple heat maps for the intrachromosomal interaction matrices. We had then planned to use a published algorithm to assemble and 3D structure of the genome and try and generate a score that would reflect the interaction frequency and the similarity of the genes that are found within each interaction region. To further analyze the data, we had also planned to over-lay this score and other genome information in our 3D map of the nucleus.

**Completed project:**

We downloaded the published HiC data from the proposed project, however the information required for analysis that was provided by the paper was unclear. After we attempted to normalize the data based on other previously published methods of HiC data analysis, we found that the normalization process did not yield heat maps that displayed any type of differential intrachromosomal contact frequency.

We therefore decided to analyze another set of published HiC data published by Rao *et al* 2014 (data available at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525>), which provided us with interchromosomal and intrachromosomal contact matrices from mammalian cells. In order to generate the normalized intrachromosomal contact matrices for a respective chromosome, we downloaded the RAW data as well as normalization factors and expected frequencies. Since the dataset for all the chromosomes was very large, we choose to perform the analysis on a subset of chromosomes using the 500kb resolution (the data contains HiC data for 1kb-1Mb resolution, we proceeded with 500kb).

First, we opened the datafile containing the raw observed contact matrix in sparse notation for the assayed chromosome. Since the matrix is symmetric, only the upper half of the matrix was populated with values. In order to perform the first normalization of raw interaction matrix, a file containing the normalization vector was provided by the authors, which was used to divide the corresponding row/matrix value (per published methods). Briefly, the normalization vector file was organized as such: the first normalization vector corresponded to the first row/matrix of the observed interaction matrix (and so on, so that for each row/column there are two corresponding normalization factors, which are multiplied and then used to divide from the corresponding value in the matrix) therefore we were able to generate a matrix of 500kb bins representing the interaction frequency across those bins. However, to normalize for the fact that regions in close proximity are likely to display higher contact frequency due to random polymer interactions during the experimental procedure, the authors provided a vector containing the expected interaction frequency normalization factor to account for this artifact. Thus, to plot a heat map of the interaction frequency of the assayed chromosome, we used the expected and observed normalization factors described above to generate interaction matrices and then divided the observed interaction matrix with the expected interaction matrix. We then used a heat map to plot the interaction frequency and confirmed that our heat map looks as expected per the paper’s figures.

We were then interested in determining which genes are present in the 500kb bins and if there was any functional relationship of the genes present in the 500kb indicated in the interaction matrix for the individual chromosome. This would then allow us to determine if regions that display higher contact frequency are regions that also have genes implicated in similar pathways. To do this, we used Gene-Ontology Analysis, which allows a user to input genes and determine which, if any, functional pathways are over-represented in the queried set. To do this, we used a package for R called EnrichR (details available here: <http://amp.pharm.mssm.edu/Enrichr/>). Briefly, this package allows the user to input a list of genes and the gene-ontology datasets they wanted to analyze the genes against and the returned values contain information about the pathways enriched and the significance of this enrichment, such that the p-value is the probability or chance of seeing the overlapped number of genes out of the total genes queried being annotated to a particular GO term given the proportion of genes in the whole genome that are annotated to that GO term (the GO term being the pathway).

In order to compare the functional relationship of genes to the contact interaction matrix, we binned the chromosome we were analyzing into 500bk bins (which corresponds to the number of rows or column in the interaction matrix). We then used gene coordinates and the gene names (file downloaded from the UCSC genome browser) to create a tibble with three columns, one containing the chromosome coordinate start, the stop (corresponding to a bin) and a third column containing a list of the genes within the region. Then we used the column/row number from the interaction matrix to determine which genes to submit for GO-term analysis using EnrichR. For example, to compare the first 500kb with the last 500 kb (which would correspond to the first row and last column of the interaction matrix), we submitted the genes associated with the first 500kb (first value in tibble) and the genes found in the last 500 kb (last value in the tibble) for GO term analysis and saved the description of the pathway enriched and the p-value associated in separate matrices (which are the same dimension as the interaction matrix). Unfortunately, the method used to compare the genes is slow, and requires that each job be submitted one at a time to the server. Therefore large-scale analysis of all the chromosomes using this method is not feasible. A way to overcome this would be to download the gene-enrichment sets and run the algorithm locally however the algorithm was not written in R and too complicated to parse through in such a short time frame.

For our graphing, we originally wanted to generate an interactive graph which would represent a 3D image of the genome based on both intrachromosomal and interchromosomal data. However, to download all the intrachromosomal data and analyze it proved to be too large, so we focused on chromosome 14 for deeper analysis and generated intrachromosomal heat maps for a couple more chromosomes. We also were unable to use packages to analyze the interchromosomal data since they were developed for MatLab and thus maintained our attention on intrachromosomal data analysis.

First, we generated heat maps of the interaction matrices for chromosomes 14, 5, 9 and 10 (see figures 1-4). These are heat maps representing the observed/expected normalized interaction matrix. As shown in the figures, it is clear that along the chromosome there are regions that are enriched in contact frequencies and these heat map patterns are consistent with published HiC data.

Next, we used ggvis to create interactive plots for the heat maps, which allow the user to hover over a certain point and see the value of the interaction frequency at the location. We also wanted to use ggvis to generate a plot that would allow the user to hover over a point, and if genes in those regions have a GO-term pathway annotated. What we were able to create is a plot which the user puts in the two bins they want to display and then the user can see the interaction frequency and the p-valu that is associated with that based on the GO-term analysis results.

We then plotted a scatter plot of the inverse of the p-values obtained from GO-term analysis using EnrichR vs the interaction frequency for chromosome 14 to see if regions with higher interaction frequency also display more significant enrichment for genes in similarly functional pathways (Figure 5). However, we see no correlation between contact frequency and p-value given the analysis we did.

ow

**Material learned in class:**

We learned R, data analysis via R and learned how to implement algorithms to further our understanding of them. We learned how to manipulate data more efficiently using vectorization as well as produce cleaner code with vectorization. Similarly, we now are more familiar with optimization methods and simulation using Monte Carlo methods.

We also learned how to generate plots using ggplot2, but more importantly learned the graphical power of R through interactive plotting with ggvis. This also provided us with valuable information on how to present data in an efficient and clear manner.

We also learned the malleability of object-oriented programming in R, which will be important for producing code that is applicable to a variety of datasets.

**Material learned via project:**

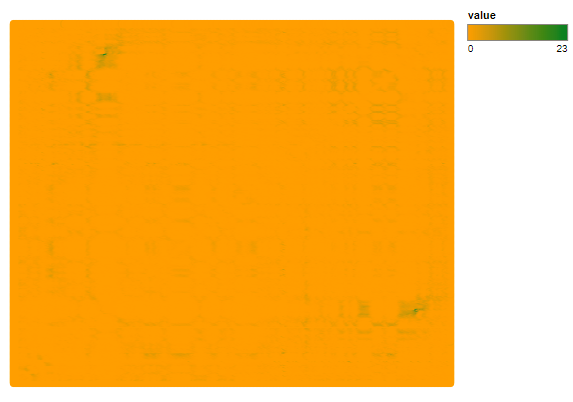
We practiced using tidyverse to clean and process data. We also practiced using ggvis to generate interactive plots, which gave us a better understanding of its functionality. We also were able to explore various packages available for R, such as EnrichR. Because we had such a large dataset to submit through EnrichR, we also learned how to debug our code in a methodical manner. For example, when EnrichR was unable to access the server the code would break and fail and we would loose all of our data about the p-values and terms calculated. Therefore, we had to use system checks in order to have EnrichR attempt the server again in order to minimize the fails and allow us to process all our data.

We also learned how to choose datasets to study because tested a lot of published datasets before finding one that was feasible for us to analyze. Likewise, we learned how to analyze the data accordingly once we found a dataset, furthering underscoring the power of R for data analysis and manipulation.

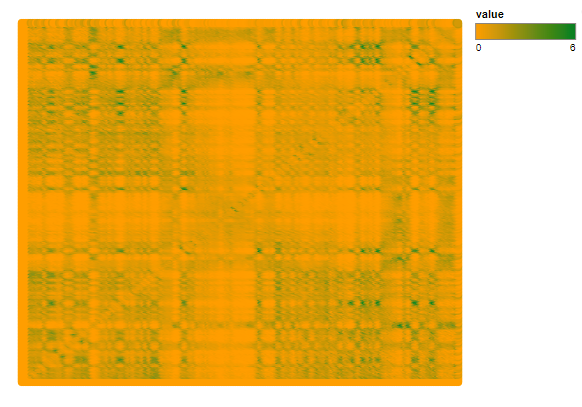
**Limitations/issues:**

One of the largest limitations with our analysis was the time it took to process the gene lists using EnrichR. Since we were not sure how to run the jobs in parallel, we had to submit the job and let it run for over 24 hours; therefore we decided to only analyze a subset of chromosome 14 with this. The estimated time to get through the entire matrix would have taken over 72 hours therefore we only used the subset that was processed during the 24 hours. We were also unable to get through as much of our proposal as we would have liked for multiple reasons; including the amount of data needed to be processed and the availability of R-packages for mapping 3D images of the nucleus using both intra and interchromosomal data.

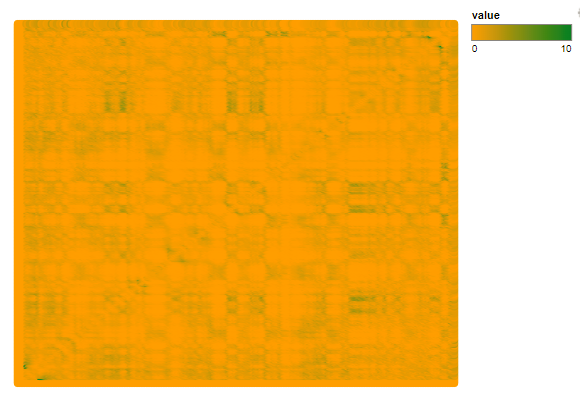
**Figures:**



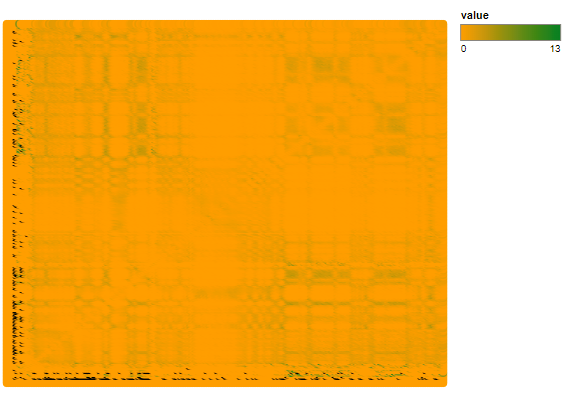
**Figure 1:** Heat map of chromosome 5 contact frequency. The X and Y axes are represent chromosome 14 coordinates, divided up into 500,000bp bins. The value is the normalized contact frequency score, with green being enriched and yellow representing little to no contact



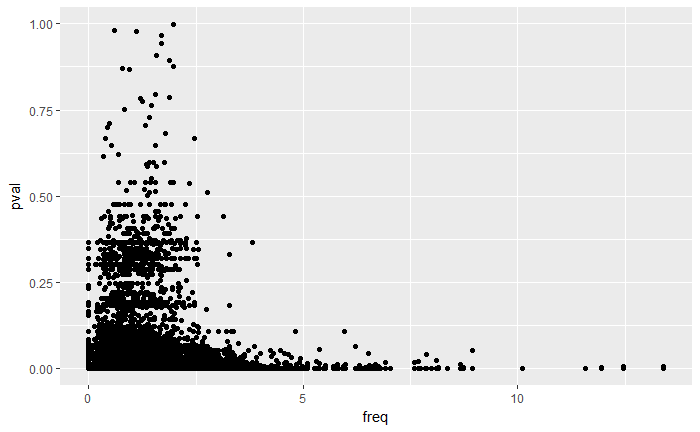
**Figure 2:** Heat map of chromosome 9 contact frequency. The X and Y axes are represent chromosome 14 coordinates, divided up into 500,000bp bins. The value is the normalized contact frequency score, with green being enriched and yellow representing little to no contact



**Figure 3:** Heat map of chromosome 10 contact frequency. The X and Y axes are represent chromosome 14 coordinates, divided up into 500,000bp bins. The value is the normalized contact frequency score, with green being enriched and yellow representing little to no contact



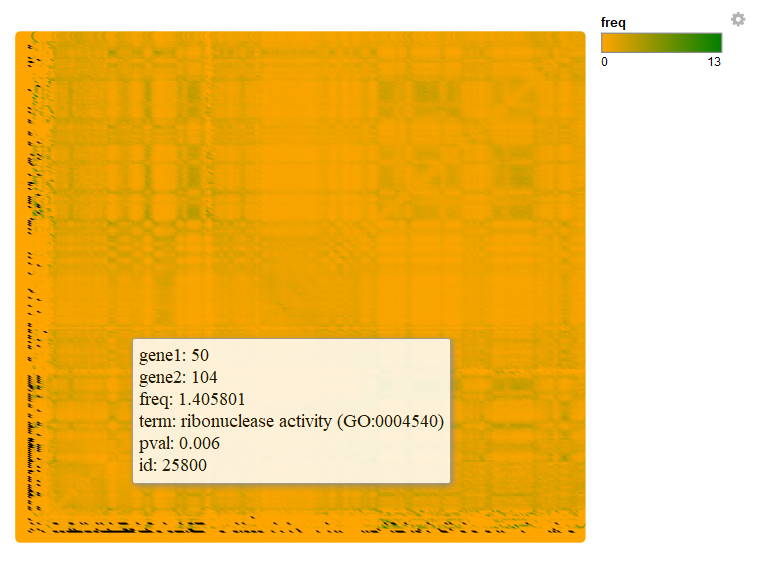
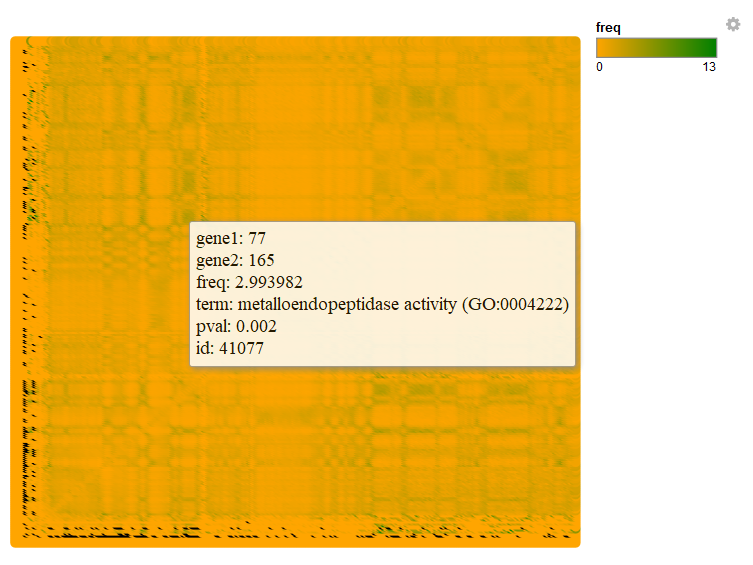
**Figure 4:** Heat map of chromosome 14 contact frequency. The X and Y axes are represent chromosome 14 coordinates, divided up into 500,000bp bins. The value is the normalized contact frequency score, with green being enriched and yellow representing little to no contact



**Figure 5**: p-value and contact frequency for chromosome 14

**Details on how interactive plots work:**

When you are viewing a heatmap of chromosome 14, you can hover over regions to display the gene bin number, the normalized interaction frequency score, the GO-term that is enriched between genes that fall within those two bins as well as the p-val associated with those terms (see screenshots below).



When you are viewing the interactive graph you can select subsets of the heatmap you want to view, allowing you to zoom into certain regions of the graph to visualize it better. The X and Y axis represent the bins on the respective chromosome (for our project this is chromosome 14).

