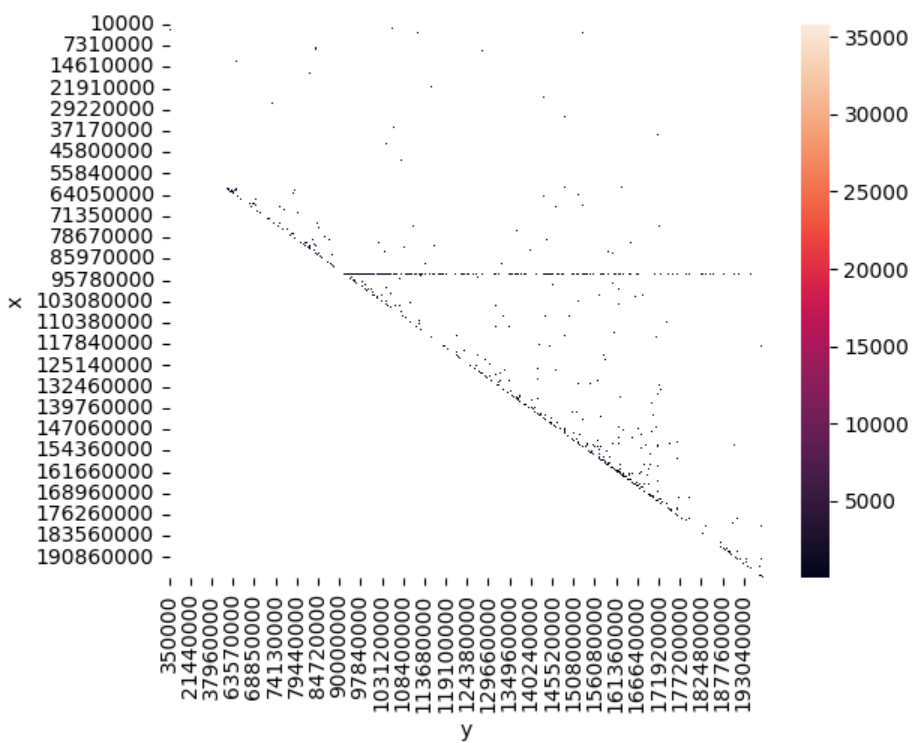
**HIC Analysis Project: Progress Report**

Simon Chow, Gabriel Fonseca, Sumitra Lele

Since we proposed the project, we have been looking into understanding the data better and looking into other packages that do the same kind of analyses. To begin the implementation of the ‘localtadsim’ package, we also started to split the methods into different functions to be able write their pseudocode.



So far, we have downloaded and processed the breast cancer contact domain and HIC data from Gene Expression Omnibus. The data is pulled from the MCF-7 breast cancer cell line. Because our HiC datasets are too large to directly read and work within our Python scripts, we rely on a special tool called ‘Straw’ to handle data extracting and loading. We first tried to use “Juicer” to do this, but ran into a lot of bugs, which is why we switched to ‘Straw’, which is a similar tool made by the same lab. ‘Straw’ enables us to read in specific sections (chromosome number and number of base pairs) of our large HIC datasets, which we can then read in our Python environment for further processing and analysis. So far, we have experimented with reading sections of data from a single chromosome (III) and have created preliminary heatmaps for visualizing the chromosome-chromosome interactions (left).

We found an ‘R’ package - ‘TADCompare’ - which uses a graph representation of the interaction matrix. The contact/interaction matrix shows us the quantified contact between region ‘i’ and region ‘j’ - this can essentially be looked at as a graph adjacency matrix, where the edges are the contacts with weights as the quantified contact and the vertices as the regions on the genome. We wish to use TADCompare to compare the results between ‘localtadsim’ and ‘TADCompare’. 

We have prepared for implementing the methodology in the paper (TAD differential analysis) by writing inputs and outputs into Python function skeletons. The overall workflow for quantifying similarity between clusters is as follows:

* Identify TADs from the interaction matrix by clustering regions on interaction intensity (in the paper, this is done using a package)
* Each TAD within the set represents a distinct genomic region that exhibits a higher level of self-interactions compared to interactions with neighboring regions. The hierarchy of the genomic regions is shown (right)
* Separate TAD sets into control and condition for comparison
* Dynamic Programming for similarity between two TAD sets using conditional entropy scoring
* Calculate p-value for similarity of clusters, saving these results for downstream analyses.