**Background Research**

In order to deepen understanding of Hi-C and its analysis methods, a number of papers were read through. Among these was a 2017 paper - [Comparison of Computational Methods for Hi-C Data Analysis](https://www.nature.com/articles/nmeth.4325), written by Mattia Forcato, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari and Silvia Bicciato. The paper stated that HOMER and HiCCUPS “called the highest proportion of promoter–enhancer interactions” at 40 kB resolution; the same resolution as the Hi-C data to be analysed. HOMER was chosen over HiCCUPS as HiCCUPS “aggregates nearby peaks into a single interaction”, thereby identifying fewer peaks.

**Downloading Data**

The Hi-C data to be analysed came from the [UCSD Ren Lab](http://chromosome.sdsc.edu/mouse/hi-c/download.html?fbclid=IwAR3Smoju3sA3YCSe5gAO9CTOCw5Rdoq9Wprb8OsBVPyd70mlGLo9DUNpPHU) on mm9 mouse embryonic stem cells. Although normalised matrices were provided, HOMER required raw data in order to form its unique tag directory so that it could be further analysed. Therefore, the files GSE35156\_GSM862720\_J1\_mESC\_HindIII\_ori\_rep\_HiC.nodup.hic. summary.txt.gz and GSE35156\_GSM862721\_J1\_mESC\_HindIII\_rep\_HiC.nodup.summary.txt.gz were downloaded from the GEO Database with the accession number [GSE35156](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35156).

**Ren mESC and Super- enhancer: Bedtools**

Using first Bedtools then Microsoft Excel, the raw data was sorted by chromosome then the start of the chromatin interactions before doing basic analysis. The command <sortBed -i ~/ HindIII.bed> was used to sort by the above order, outputting the sorted data directly onto the terminal display. The data was then copied into Excel and the following statistics were calculated:

|  |  |
| --- | --- |
| Minimum | 160000 |
| Maximum | 5240000 |
| Mean | 1092792 |
| Median | 880000 |
| Mode | 560000 |
| Standard Deviation | 740266 |

Table 1.0: Basic Statistics for Ren data

All data, excluding the mean, was a multiple of 40 000, due to the 40 kb resolution of the Hi-C data. As can be seen by the table above, the distance between the start and end of chromatin interactions had a wide range with high maximums and low minimums and a relatively large standard deviation.

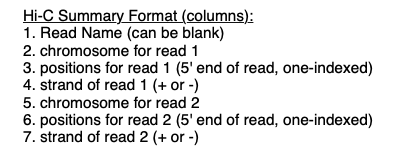
Bedtools was then used again to find the interactions between the sorted Ren data and the 231 super- enhancers provided. This was achieved through the command <bedtools intersect -a ~/ mESC-SE.bed -b ~/HindIII.bed which output the results onto the terminal, displaying the chromosome and the starts and ends of each interaction. The majority of the super- enhancer interactions were found, with only eleven missing. The unmatched super- enhancers were located in chromosomes 5, 8, 9, 10, 11, 15, 16, 17.

|  |  |  |
| --- | --- | --- |
| chr10 | 59420365 | 59437537 |
| chr10 | 62346394 | 62361563 |
| chr11 | 97517673 | 97524159 |
| chr11 | 98823511 | 98826466 |
| chr15 | 88539016 | 88539831 |
| chr16 | 8758173 | 8779472 |
| chr17 | 45593477 | 45596503 |
| chr5 | 123584659 | 123590728 |
| chr8 | 44405736 | 44406755 |
| chr9 | 58119837 | 58128504 |
| chr9 | 58119837 | 58128504 |

Table 2.0: Missing Super- enhancers

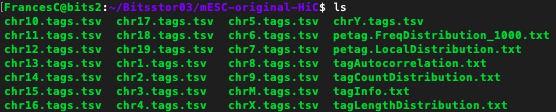
**HOMER: makeTagDirectory**

For its command <[makeTagDirectory](http://homer.ucsd.edu/homer/interactions2/HiCtagDirectory.html)>, HOMER accepted two types of input files, mapped read files and Hi-C summary files. The GEO data was in Hi-C summary format, therefore the command was <makeTagDirectory mESC-ori-HiC -format HiCsummary GSE35156\_GSM862720\_J1\_mESC\_HindIII\_ori\_rep\_ HiC.nodup.hic.summary.txt.gz>. All optional parameters were defaulted. Hi-C summary files are tab delimited text files with the following assignments:



Screenshot 1.0: Explanation of Hi-C format from [HOMER](http://homer.ucsd.edu/homer/interactions2/HiCtagDirectory.html)

Upon completion, HOMER output a directory to terminal, consisting of files of tags sorted by chromosomes and additional information files.

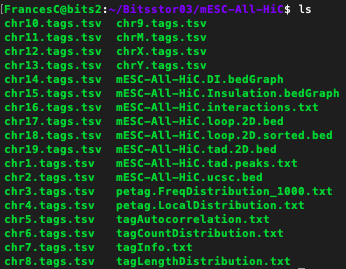


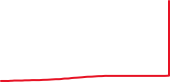
Screenshot 2.0: Tag Directory internal Structure

This command was also repeated for the GSE35156\_GSM862721\_J1\_mESC\_HindIII\_rep\_HiC.nodup. summary.txt.gz file to result in another directory named mESC-rep-HiC . These directories were then combined using the same makeTagDirectory command, <makeTagDirectory mESC-All-HiC -d mESC-ori-HiC/ mESC-rep-HiC/ > to form the combined directory for analysis: mESC-All-HiC.

**HOMER: findTadsAndLoops.pl**

After combining the original and replicate mESC directories, the command <[findTadsAndLoops.pl](http://homer.ucsd.edu/homer/interactions2/HiCTADsAndLoops.html)> was used to locate topologically associating domains (TADs) and chromatin interactions (loops). The following command was used to find all TADs and loops within mESC-All-HiC; <findTadsAndLoops.pl find mESC-All-HiC. As with the makeTagDirectory, optional parameters were defaulted. Another function was also built in the findTadsAndLoops.pl, the filterTadsAndLoops command, though functionality was limited due to the lack of genome file mm9. The output files were placed into the combined directory:





Screenshot 2.1: Added output files from findTadsAndLoops.pl (in red)

**HOMER: findTadsAndLoops.pl (continued)**

While the command was running, TAD and loop details of each chromosome were displayed on screen. These were captured and entered onto Excel to form a table detailing the number of loops and TADs that HOMER found for each chromosome.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Chr. | Avg. interaction | Uncertainty | Without outliers | Average coverage | Coverage ratio | Interaction matrix | TAD | TAD (w/o overlap) | Loop |
| 10 | 5729.1 | 21087.1 | 5887.2 | 5887.2 | 0.547 | 43332 x 1333 | 554 | 313 | 796 |
| 11 | 5999 | 3582.1 | 6099.6 | 6099.6 | 0.567 | 40615 x 1333 | 557 | 337 | 954 |
| 12 | 5630.8 | 2559.1 | 6002.4 | 6002.4 | 0.558 | 40420 x 1333 | 438 | 224 | 850 |
| 13 | 5588.2 | 2276.6 | 5821.3 | 5821.3 | 0.541 | 40095 x 1333 | 512 | 250 | 905 |
| 14 | 5094.4 | 2404.9 | 5500.7 | 5500.7 | 0.511 | 41727 x 1333 | 443 | 249 | 788 |
| 15 | 6372.6 | 2436.8 | 6564.1 | 6564.1 | 0.61 | 34499 x 1333 | 427 | 253 | 661 |
| 16 | 5688.1 | 2171.5 | 5893 | 5893 | 0.548 | 32774 x 1333 | 379 | 225 | 548 |
| 17 | 5168.9 | 2190.1 | 5214.9 | 5412.9 | 0.503 | 31758 x 1333 | 410 | 223 | 759 |
| 18 | 5717.4 | 2155.2 | 5929.5 | 5929.5 | 0.551 | 30258 x 1333 | 371 | 219 | 581 |
| 19 | 5647.2 | 2295.4 | 5968.5 | 5968.5 | 0.555 | 20448 x 1333 | 255 | 152 | 476 |
| 1 | 5538.9 | 2148 | 5700.5 | 5700.5 | 0.53 | 65732 x 1333 | 797 | 499 | 1171 |
| 2 | 5738.4 | 3879.2 | 5887.1 | 5887.1 | 0.547 | 60583 x 1333 | 779 | 439 | 1258 |
| 3 | 5589.1 | 2104.1 | 5713.1 | 5713.1 | 0.531 | 53200 x 1333 | 626 | 395 | 874 |
| 4 | 5240.3 | 2378.5 | 5496 | 5496 | 0.511 | 51877 x 1333 | 692 | 372 | 1227 |
| 5 | 5207.6 | 2150.7 | 5475.9 | 5475.9 | 0.509 | 50846 x 1333 | 621 | 370 | 1201 |
| 6 | 5970.4 | 2297.6 | 6126.4 | 6126.4 | 0.569 | 49840 x 1333 | 636 | 380 | 863 |
| 7 | 4578.6 | 2516.9 | 5205.2 | 5205.4 | 0.484 | 50842 x 1333 | 609 | 323 | 1519 |
| 8 | 5519 | 2403.9 | 5842.1 | 5842.1 | 0.543 | 43913 x 1333 | 509 | 303 | 881 |
| 9 | 5741.7 | 2525.5 | 5871.3 | 5871.3 | 0.546 | 41359 x 1333 | 501 | 306 | 830 |
| M | 986559 | 224524.1 | 986559 | 986559 | 91.676 | 6 x 1333 | 0 | 0 | 0 |
| X | 2383.2 | 1426.4 | 2594.7 | 2594.7 | 0.241 | 55551 x 1333 | 776 | 442 | 4391 |
| Y | 970.4 | 9152.3 | 2664.8 | 2664.8 | 0.248 | 5301 x 1333 | 19 | 8 | 305 |

Table 3.0: findTadsAndLoops.pl command terminal output

In total, HOMER found 10 911 TADs and 21 838 loops, though more than half of the TADs overlapped, with only 6282 isolated TADs. Basic statistics were done on the loops discovered, with one column with all chromosomes included and one column without chromosomes M, X and Y. This was due to the fact that all three deviated greatly from the rest, as can be seen from the large difference in standard deviations below. All three were relative outliers with chromosomes M and Y having minimal to no TADs or loops and X having an abnormal ratio of loops to TADs. Little TAD analysis was done, largely due to the fact that the project focussed on chromatin interactions with super- enhancers.

|  |  |  |
| --- | --- | --- |
|  | All Chromosomes Included | Chromosomes M,X,Y Excluded |
| Average | 992.6364 | 902.2105 |
| Minimum | 0 | 476 |
| Maximum | 4391 | 1519 |
| Median | 856.5 | 863 |
| Standard Deviation | 810.8121 | 268.8198 |

Table 3.1: Loop statistics

**HOMER and Super- enhancer: Bedtools**

Using the loops file output by HOMER, the loops found by HOMER in the processed Ren Lab data was compared to the provided list of 231 super- enhancers with the intersect function of Bedtools. The loops file output was first edited to account for interactions from both sides; meaning that if the file included an interaction from A to B, the reverse was also added in. This essentially resulted in doubling the loop count.

The file: mESC-All-HiC.loop.2D.bed was first sorted using the sortBed function of Bedtools before the results were copied into a text file (mESC-All-HiC.loop.2D.sorted.bed) and used in the command; <bedtools intersect -a ~ /mESC-SE.bed -b ~/mESC-All-HiC. loop.2D.sorted.bed>. Bedtools then output the results onto the terminal, however there were only 28 intersections between the 231 super- enhancers and the sorted file of 40 000 plus HOMER- found interactions.

|  |  |  |  |
| --- | --- | --- | --- |
| Chromosome | Start | End | Length |
| chr1 | 13049615 | 13050000 | 385 |
| chr1 | 36070190 | 36072000 | 1810 |
| chr1 | 84885000 | 84887132 | 2132 |
| chr1 | 137088000 | 137091000 | 3000 |
| chr1 | 137088000 | 137091000 | 3000 |
| chr3 | 34544904 | 34545000 | 96 |
| chr3 | 34544904 | 34545000 | 96 |
| chr4 | 125217000 | 125220000 | 3000 |
| chr4 | 138006000 | 138006368 | 368 |
| chr4 | 138006000 | 138006368 | 368 |
| chr4 | 154537213 | 154538078 | 865 |
| chr5 | 120138000 | 120141000 | 3000 |
| chr5 | 120162000 | 120165000 | 3000 |
| chr6 | 122613000 | 122614260 | 1260 |
| chr7 | 147132000 | 147135000 | 3000 |
| chr9 | 120588000 | 120591000 | 3000 |
| chr12 | 12795000 | 12795881 | 881 |
| chr12 | 12948000 | 12950936 | 2936 |
| chr12 | 111655417 | 111656705 | 1288 |
| chr13 | 98058000 | 98061000 | 3000 |
| chr13 | 98202400 | 98205000 | 2600 |
| chr14 | 71022659 | 71025000 | 2341 |
| chr14 | 76899000 | 76902000 | 3000 |
| chr14 | 76902000 | 76905000 | 3000 |
| chr17 | 31953000 | 31956000 | 3000 |
| chr17 | 47649000 | 47649043 | 43 |
| chr17 | 47649000 | 47649043 | 43 |
| chr19 | 53523440 | 53526000 | 2560 |

Table 4.0: HOMER and Super- enhancer matches

Basic analysis was done on the 28 chromosomes, including mean, median, mode, the minimum and maximum and standard deviations of the lengths between the start and end of each interaction.

|  |  |
| --- | --- |
| Average | 1895.429 |
| Minimum | 43 |
| Maximum | 3000 |
| Median | 2450.5 |
| Mode | 3000 |
| Standard Deviation | 1177.974 |

Table 4.1: HOMER and Super- enhancer interaction statistics

**HOMER and Super- enhancer: Bedtools (continued)**

Though the maximum length was capped at 3000, it is clear that values ranged widely from multiple low values of 43 and 96, and a relatively large standard deviation. However, values clearly were more densely populated toward the higher end, as the median was greater than the average by almost 600.

**HiCCUPS: Juicer**

Since the results from the HOMER and Bedtools method found less than 15% of the 231 super- enhancers, it was decided to try a different approach. As the paper earlier, ‘Comparison of Computational Methods for Hi-C Data Analysis’, had also highlighted HiCCUPS with regards to 40 kb resolution analysis, HiCCUPS became the new method.

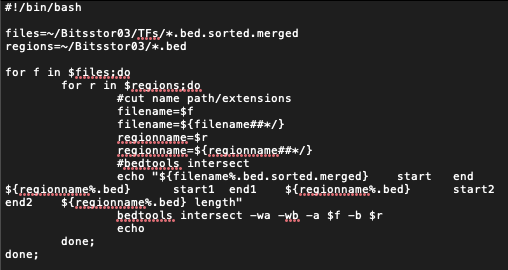
[HiCCUPS](https://github.com/aidenlab/juicer/wiki/HiCCUPS) is an Hi- C analysis algorithm of [Juicer](https://github.com/aidenlab/juicer/wiki), “a platform for analysing kilobase resolution Hi-C data”. It finds chromatin loops with the command <hiccups [-m matrixSize] [-c chromosome(s)] [-r resolution(s)] <HiC file> <outputDirectory>>. In order to use the software, the raw data files from the GEO database had to be reformatted to a [4DN DCIC](https://github.com/4dn-dcic/pairix/blob/master/pairs_format_specification.md) format with seven columns. The ‘pre’ command was then run to convert these 4N DCIC files to .hic files in order for Juicer’s methods to work- <pre <infile> <outfile> <genomeID> >.

Though the reformatting was successful, the actual HiCCUPS command failed, as CUDA was not installed on the system.

**Transcription factors and Regions: Bedtools**

Since HOMER found less than 15% of the 231 super- enhancers and HiCCUPS failed due to the lack of CUDA software, the project shifted its focus onto transcription factors.

To examine where specific transcription factors intersected with the 231 super- enhancers, a script was written using terminal to use Bedtools to find interactions between region 1 and each transcription factor, and region 2 and each transcription factor. Both region files came from one file of chromatin loops (chromosome A, start of A, end of A, chromosome B, start of B, end of B), with one side of the interaction named region 1 and the other region 2.



Screenshot 3.0: Macport1.sh- the script for running transcription factor and region interactions

The above script iterated through each transcription factor file with a .bed.sorted.merged extension, using Bedtools to find interactions between the file and region 1, then region 2. Although there were 56 transcription files to intersect, which should have resulted in 132 files, the end result was 105 files. This was due to the fact that some transcription factors had no intersections with one or both of the region files. Each output file was labelled with the transcription factor name and the region file Bedtools had intersected it with, e.g. Ctcfl-R1.bed- transcription factor Ctcfl and region file R1.



Screenshot 3.1: Example of some of the output files

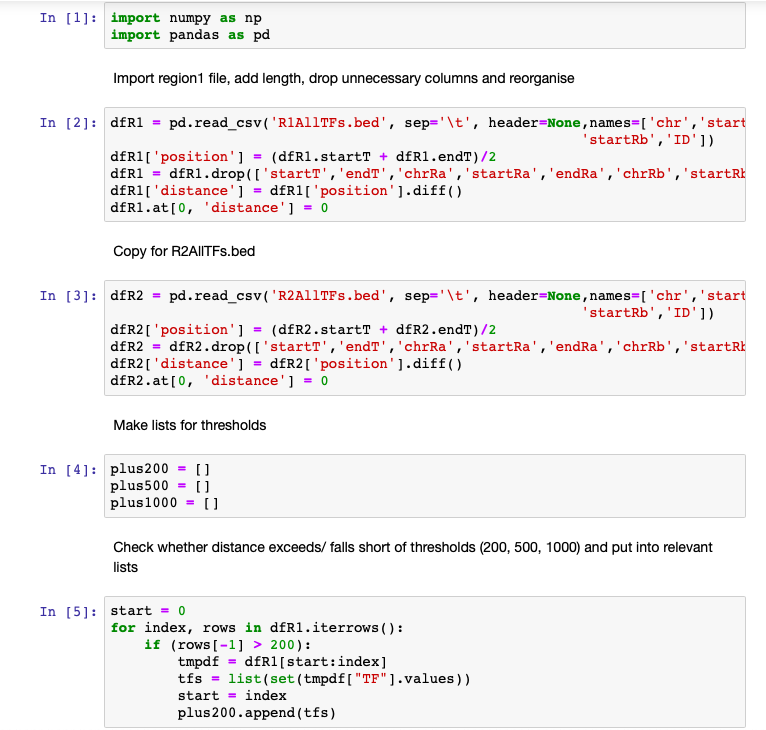
**Jupyter- Clustering: Code**

In order to sort the region files into distance- based clusters of distances greater than 200, 500 and 1000, a Python program was written in [Jupyter](https://jupyter.org/) notebook by accessing the terminal with the command (jupyter-notebook –no-browser –ip=0.0.0.0).

Numpy and pandas were imported first before reading region 1 and region 2 as dataframes. The dataframes had two variables added, with position being the average of the start and end of the interaction, and distance being the difference of the position of the current row to the row above. All unnecessary variables were removed, leaving only the index, chromosome, transcription factor, position and distance.

To account for the three thresholds, empty lists were made for each; plus200, plus500 and plus1000. The list plus200 was for distances greater than 200, etc. Each region dataframe was then iterated over three times, one time per threshold.

For each index and row of the dataframe, if the distance of the row was greater than the threshold (an integer of 200, 500 or 1000), all the values from start (which started at 0) to the current index were added to a temporary dataframe. The start value was then set to be equal to the current index and the transcription factors taken from the temporary dataframe and put in a set to remove any non- unique values. This set (tfs) was then set to a list before being appended to the relevant list.



Screenshot 4.0: Part 1 of the Jupyter program

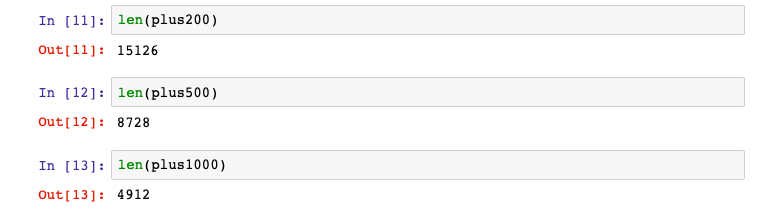
**Jupyter- Clustering: Code**



Screenshot 4.1: Part 2 of the Jupyter program

**Jupyter- Clustering: Results**

The end results of the lists are listed below; with 15126 interactions having a distance of greater than 200, 8728 interactions greater than 500, and 4912 interactions greater than 1000.



Screenshot 4.2: Results of the threshold lists