Monday 31st March 2014

I completed the HiC_matrix.R script which accepts as input a series of .txt files (see Lieberman data as provided by Hu Ming)

The Ming data can be found in:

`/ifs/home/cl3011/ROTATION_3/data/contact_matrix`

They consist of files describing the *cis* and *trans* interactions found when using HindIII as the restriction enzyme in the Hi-C experiment or NcoI.

I used HiC_matrix.R in each one of the Hind3_cis_obs and Hind3_trans_obs directories and I created the corresponding genome matrices (containing the *cis* interactions only - the rest of the matrix was filled in with zeros). The problem is though that when I tried to run previous scripts in order to find the Spearman correlation for all chromosome (one-by-one) for the HindIII and NcoI experiments and using one of the chromosome vectors that were generated by HiC_matrix.R, I got back a single correlation for only the first chromosome and then all the other correlations are NA.

I checked the number of lines of the chromosome vector file and that of the input matrices (they all have 3033 which seems to be right given that the resolution is 1MB. However the 1024kb vector that Aris has created has more lines…)

The numbers of the chromosomes appear correct on the matrix that is used as input for the production of the boxplot. However only the first one of the correlations appears. So, there must be something wrong with the input matrices. Is the **write.table** that I use to output the vector and the matrix wrong or I have simply generated the matrices incorrectly in the first place?

Tuesday 1st April 2014

Solving the problem with the code used to analyze the data that Ming Hu provided (HindIII and NcoI Hi-C data (1MB resolution))

I found the problem in the **HiC_matrix.R** script. Instead of looping over all the elements of the chromosome vector (c), I was previously looping over just the first one. I corrected the code and I submitted the revised version to my repository on GitHub (in order to have a safe copy in a location other than the cluster). Before finding this problem, I tested the matrices which had no missing elements using the commands shown below. I was getting NA correlations because the standard deviations were equal to zero (all the matrices except for the one for chromosome 1 were full of zeros).

All the data from Hu Ming (HiC data from the Lieberman-Aiden 2009 paper for both enzymes HindIII and NcoI - 1MB resolution) can be found in the **contact_matrix** directory on my Phoenix account. I analyzed only the *cis* data and the corresponding histogram can be seen below.

The Spearman correlation seems increased compared to what I have found previously (using the individual replicates) but I do not know how they combined the data in only two sets.