

---

# Laboratory Journal

Rotation 3

---

Harris A. Lazaris

[lazaris@nyu.edu](mailto:lazaris@nyu.edu)

Beginning 10 March 2014



# Contents

<b>Monday, 10 March 2014</b>	<b>1</b>
1 Collect data concerning the Ren “Topological domains” paper. . . . .	1
<b>Friday, 14 March 2014</b>	<b>2</b>
1 Try to calculate Spearman correlation coefficients using HiCNorm . . . .	2
<b>Monday, 17 March 2014</b>	<b>3</b>
1 Run MATLAB code with Aris and Anju for HiC evaluation . . . . .	3
<b>Monday, 24 March 2014</b>	<b>5</b>
1 Discuss with Anju how to create the pipeline for the correction of fil- tered.dat files in order to get the corrected.dat . . . . .	5
<b>March 25 2014</b>	<b>7</b>
1 Write and run script that calculates correlations for Hi-C contact matrices after removing lines and columns that sum to zero. . . . .	7
<b>March 26th 2014</b>	<b>8</b>
1 Create a matrix with all Lieberman-Aiden data . . . . .	8
<b>Tuesday 8 April 2014</b>	<b>9</b>
1 Plan for the next month . . . . .	9



# Monday, 10 March 2014

## 1 Collect data concerning the Ren “Topological domains” paper.

These are the tasks of the day:

- Go to GSE35156 on GSE and try to download the matrix data
- If you do not find any data there, contact the authors.

The data that we are looking for are the matrices that give all the Hi-C interactions for the cell types that the Ren group used in the study: Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376380 (2012). I went on the website <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35156> and found that the matrix files listed there are actually incomplete as they contain only the headers and not the matrix data. I also did a character count to confirm that the size of the file corresponds to the headers only and there is not anything else in there, that I missed.

Having proven that the matrix data are not there I e-mailed Bing Ren (biren@ucsd.edu) and the first author of the paper Jesse R. Dixon (j1dixon@ucsd.edu) and requested the processed data concerning the matrices and also the topological domains. Moreover, I requested detailed description of the methods used (trimming of the data, binning etc). I cc'ed the e-mail to Aris (Tsirigos) as well.

# Friday, 14 March 2014

## 1 Try to calculate Spearman correlation coefficients using HiCNorm

These are the tasks of the day:

- Download the data from <http://www.people.fas.harvard.edu/~junliu/HiCNorm>
- Play with the data to see if you can calculate any coefficient.
- Meet with Aris at 2pm to run the MATLAB code. Anju said that he will be available early next week. So, we can start on our own.
- Discuss with Aris the requirements of the two proposals:

**HPC Project** Bash scripting, high-performance computing, parallel jobs? What should I mention on the proposal? He wants us to write a brief discussion on the question, its importance, our approach and how we are going to implement it.

**Methods Project** We have to start thinking about it too. The code should be placed on `git`. Larry said that it would be great to use some Python (Scipy, Numpy etc). Also he mentioned that we may need to create a database. What kind of database would be appropriate for our data?

# Monday, 17 March 2014

## 1 Run MATLAB code with Aris and Anju for HiC evaluation

These are the tasks of the day:

- Use rep1.mat and rep2.mat as input files for MATLAB.
- Meet with Aris and Anju (Skype) at 2pm to run the MATLAB code.
- Discuss with Aris the requirements of the two proposals:
  - Discuss with Aris the content of the lab meeting presentation to start working on it.
  - Check for R debugger.

Useful command to extract the chromosomes only:

```
more .reg+ | cut -f2 | cut -d' '-f1 | sed 's/^chr //' | \
sed \ 's/X/23/' > ...
```

The presentation is going to be on the Hi-C method in general (Focus on the Lieberman paper (2009)). Then, the second half will be on the latest Ren paper (Nature 2013). Start working on the presentation today.

The data are available in a centralized location and I have created a soft linked to this location (Dekker-Science-2009). In the directory `run_dat_files` the script that I wrote today is placed: `hic_chr_by_chr_matrix_comp.R`. Moreover the `chr_file` which is also in this directory, provides the names for the chromosomes (based on 4096KB resolution). I ran the command:

```
Rscript hic_chr_by_chr_matrix_comp.R
../Dekker-Science-2009/hiclib/hiclib.GM-rep1-HindIII-HiC/
heatmap-res-4096kb.filtered.dat
../Dekker-Science-2009/hiclib/hiclib.GM-rep5-NcoI-HiC/
heatmap-res-4096kb.filtered.dat chr_file spearman false
```

to calculate the Hi-C correlation between the first replicate with HindIII and the replicate 1 with NcoI (for resolution 4096KB):

In the aforementioned example the Spearman correlation is: 0.873 (rounded to three decimal places).

*Monday, 17 March 2014*

### **Future plans**

1. Make the Rscript output chromosomes and correlation (23 chromosomes and the corresponding correlations) and save to file with name determined by the input files.
2. Write an R script that will use the correlations from these files to create boxplots (all boxplots should appear on the same plot).
3. Add the electronic lab-book and the scripts to a private repository on GitHub (you may share with Aris).
4. Ask the guys again for the non-normalized data



# Monday, 24 March 2014

## 1 Discuss with Anju how to create the pipeline for the correction of filtered.dat files in order to get the corrected.dat

These are the tasks of the day:

1. Discuss with Anju about the method for Hi-C correction.
2. Correct the diagram for hiclib 4096kb resolution
3. Update the lab book with all the images that you have already generated from the analysis of hiclib data (namely 128kb, 1024kb and 4096kb).
4. Follow Anju's instructions

Anju's instructions:

1. Start with the matrix (.dat file)
2. Load the matrix to memory
3. Run HiCHarvard.m from MATLAB (ask Aris for the files that HiCHarvard requires (the chrom file, the length file and the GC file).
4. You should get back P which is the normalized (with HiCNorm) version of the matrix that was provided.

Run the following command to create file with chromosome vectors to run with RunExperiment.

```
cat -n chromosome_vectors/hg19.chr.w\=1024kb.vec\  
| sed 's/$/ + 1 1/'> hg19.chr.w\=1024kb.reg+
```

Very important: When you write BASH scripts, do not include the names of the directories there as it carries in the variable the whole path. Go to the directory you have the files you want to process and create a symbolic link to where the bash script is, in order to be able to run it. Create a subdirectory in the directory you are in order to write the output files there.

*Monday, 24 March 2014*

### **Future plans**

1. Meet Anju tomorrow. Ask him about the pipeline and which were the exact values of the parameters they used in the paper. Also which range of values they used when they tested the parameters.
2. Check the Dekker GSE entry. What files they have? Which are biological and which are technical replicates? How did they combine those? In other words how did they end up with the filtered and corrected ones that they finally analysed? Did they combine biological or technical replicates? If yes, how they did it?

# March 25 2014

## 1 Write and run script that calculates correlations for Hi-C contact matrices after removing lines and columns that sum to zero.

Tasks of the day:

1. Send to Anju what time to talk tomorrow. He said we can talk tomorrow but he has not given exact time yet.
2. Sent e-mail to Mirny lab (more specifically to Anton Goloborodko (golobor@mit.edu)) and asked for the raw data and the processed (after the iterative correction) data used for the creation of contact matrices.
3. Modify the R script in order to remove rows and lines that include zeros (the new script is named: `hic_chr_by_chr_nozero_cor.R`).
4. Siddarth (HiCNorm paper) said that we can download the non-processed data from Lieberman-Aiden. Also for the Dixon et al. paper he gives the corresponding website.

The new script was used to produce the summary for the correlation of `HindIII_NcoI_1_data` (for replicate 1 both for HindIII and NcoI) and both for filtered and corrected.

The results are given in the figures that follow. It seems that filtered and corrected have no difference in terms of correlation (when the zeros are removed). There is difference though if we compare the correlation of matrices including the zeros (see previous day, Spearman correlation  $\approx 0.7$ ) with the correlation of matrices without including the zeros (which is actually lower).

### Future plans

1. Discuss with Anju tomorrow (he has not replied about the file requirements though).
2. Discuss with Aris the possibility that Anju writes a manual for the pipeline (reviewers have asked for it anyway and if the software is targeted to biologists, they will certainly need one).
3. Do not forget to update Github repository.

# March 26th 2014

## 1 Create a matrix with all Lieberman-Aiden data

Tasks of the day:

1. Continue to create boxplots where the lines with zeros have been removed from the dat files. We would like to see how this affects the results from the other resolutions as well.
2. Leave at 16:30pm and go home to get stuff for the Ethics class.
3. E-mail was sent to Ming Hu to ask for the 1MB contact matrices that were derived from the Lieberman-Aiden paper and were used for the HiCNorm paper.

### Future plans

1. Meet Anju next week. He says that he wants to write a manual on the new method when we find time. The aim of the meeting is to run the MATLAB HiC pipeline with a part of a file to see what the expected output is and what is the range and meaning of parameters.
2. We have to talk with Aris about the Methods project and the requirements.

# Tuesday 8 April 2014

## 1 Plan for the next month

Tasks of the day:

1. Download and install genlasso. Try to understand the corresponding documentation.
2. In the next month you must complete the introduction for the paper that we will submit to NAR.

Genlasso downloaded and installed. *Caution:* It is possible to install R packages to R on the cluster. A local personal library is automatically created in your local directory and all the packages are installed there (see R directory in home directory)!

### NAR Introduction

I am supposed to write an extensive review (about ten pages) which will consist of:

- 30% Biological question and background information.
- 60% The current methods and their main findings.
- 10% Strengths of this study, unanswered questions and my own ideas.