Date: April 14th 2014

Title: Creating a script to convert Ren data (from Nature 2013 paper) to the appropriate format to be used by genomic_tools

What to do

1. Install genomic_tools on the cluster
2. Write a script to turn .txt Hi-C files to reg.gz files
3. Think about the next steps for producing matrices and heatmaps using the Ren data.

- I download the source code for genomic_tools from `https://code.google.com/p/ibm-cbc-genomic-tools/source/checkout` and then I installed it locally based on the instructions. I also set up the $PATH to be able to call any of the commands of the package from anywhere.
  The version installed is 2.8.0 which includes the code for hic analysis. Type `gtools_hic -h` any time help for the possible options is required.

- I wrote a script named `txt_to_reg.sh` which reads the .txt files in the current directory and converts them to the .reg format. The script is placed in the **Scripts** directory in my home directory on the cluster.

- In order to perform the binning (split each chromosome in 1MB chunks)
  I will use genomic_tools. The syntax is shown below (with the first of the .gz files as input):

  `gtools_hic bin --bin-size 1000000 -v reg.gz`

  where:

  - gtools_hic bin is the command for the binning
  - –bin-size: is the size of each bin (default 1MB)
  - -v: verbose
- Caution: Use *gunzip* when you want to unzip something and *gzip* when you want to compress it.

Moreover, from now on I will be writing the lab-book in Markdown as it is much easier than LaTeX in terms of placing URLs and code in-line with the text.

For this reason I added the current document to my private repository on GitHub and this is the document I will be modifying every day.

Date: April 16th 2014

Title: Create matrices of different resolutions using Ren's Nature 2013 data

and also test the Python *hic_to_matrix* script when using the right input

(file that corresponds to bins when having 100MB resolution).

What to do

1. Write a bash script to automate matrix creation based on Aris' gtools
2. Check your Python script with the right input to see if it works. Change it to handle bed files properly if you can.

3. You have to start working on the table reviewing all Hi-C papers as well.

- I used `gtools_hic bin --bin-size 100000000 -v GSM1055800_HiC.IMR90.rep1.nodup.summary.reg.gz | head > GSM1055800_HiC.IMR90.rep1.nodup.summary.head.t` in order to create that dummy file that I will used as input to my script. The resolution in this case is correct to compare with Aris .dat file (100MB)
- Indeed, I used the command `cat GSM1055800_HiC.IMR90.rep1.nodup.summary.reg.gz | gunzip | head -10 | gtools_hic bin -v --bin-size 100000000 > 100MB_input` to produce the correct input to use on my script. It is worth mentioning here that head somehow produces only the 5 first lines! When I compared the 100MB_input.dat (matrix generated using my script) with out_100MB.dat (matrix generated using Aris' method) they were identical (see diff below)!
- As Aris said, I generated the matrices using his method for the following resolutions 500KB, 1MB, 5MB, 100MB (I also did 256KB, 1024KB, 4096KB to be able to compare with previous boxplots). All data and the results of this analysis can be found on cluster in `Ren-Nature2013a`.