Bioinformatics Lab Rotation

Institute of Molecular Health Sciences - RNA biology:
Group Leader: Assistant Professor Constance Ciado
Supervised by: Dr. Jian Yu

Student: Damianos Melidis, Msc in Computational Biology and Bioinformatics
Student id: 12-945-416, dmelidis@student.ethz.ch

January 4, 2014

Abstract

During this laboratory rotation, I learned how to manipulate ChIP sequencing data using bioinformatics tools, such as alignment of reads to the reference of interest and selection of the most significant alignments. We try to apply this knowledge to contribute in current laboratory project of in-silico identifying the epigenetic pattern of a specific L1 family that is expressed in embryonic stem cells in mouse organism.

1. Introduction

1.1 ChIP sequencing

ChIP sequencing or ChIP seq is a method to identify positions where proteins bind to DNA. The basic technique of chromatin immunoprecipitation is combined with high-throughput parallel sequencing (Next Generation Sequencing, NGS) to find binding sites of proteins in a DNA fragment. The most common uses of ChIP seq consist of specifying how transcription factors and proteins associated with chromatin react on phenotype-affecting mechanisms. An illustrative figure of ChIP seq workflow is shown on the next picture. This brief introduction is based on `http://en.wikipedia.org/wiki/ChIP-sequencing`

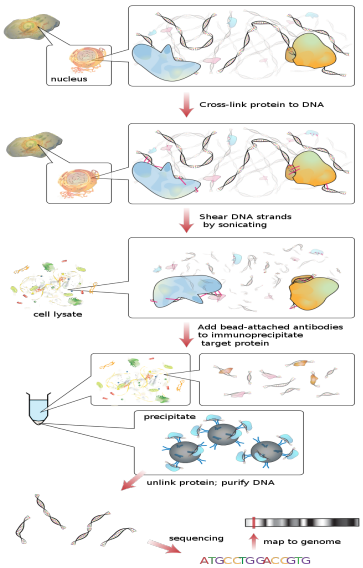

Figure 1: ChIP seq workflow, (`http://en.wikipedia.org/wiki/Polymerase_chain_reaction`)

1.2 L1 elements

In the following section we give a brief introduction to L1 elements ([Leslie A. Pray], [Eric M. Ostertag et al.]). The L1 elements belong to non-LTR transposable elements. The average length of L1 elements is 6 kilobases and they consist approximately the 15-17% of the mouse genome. In figure 2 we show the individual genomic regions of L1 elements. Surprisingly only these elements are the active class of transposons in mice. Studies (such as [Miki, Y., et al.]) have shown that L1 elements transpose in somatic cells in mammalian genome and they may contribute to disease development. duplicate themselves and they are likely responsible for the expansion of nonautonomous retrotransposons, specifically Alu elements, processed pseudogenes and SVA elements in mouse genome. Studies (such as [Miki, Y., et al.]) have shown that L1 elements transpose in somatic cells in mammalian genome and they may contribute to disease development.
In mouse organism the small interfering RNAs (siRNA) known to regulate gene expression, are also regulate transposons movement. The siRNAs that regulate L1 action are derived from the 5' UTR of the L1. Only 100 L1 genes are maintained active in DNA.
Trying to assess the difference between the 3 elements of L1 subfamily Gf, Tf and A genes we perform multiple sequence alignment to their corresponding consensus sequences using the webtool [1], resulting to next figure, where blue and green represent low scores and the more red the more high scores we have for the multiple sequence alignment. Also using the [2] we find the similar sequences between our L1 subfamily members (query) and the mouse genome (database). For Tf genes we had 2390 hits, for A genes 1655 and for Gf we observed 1336 hits. These two tests show us that the three genes are half similar, where difference on nucleotide bases may give a specific function to each one gene.

1.3 Problem Approach

Our goal is to understand the epigenetic pattern of the L1 subfamily (Gf, Tf and A genes) as they are expressed in embryonic stem cells and they can (give birth to) express small RNAs. Recalling figure 3 in this L1 subfamily the 5' UTR region is the only unique genomic region between Gf, Tf and A. A high percentage of this region contains repeated sequences. Investigating the epigenetic factors that regulate the L1 subfamily, we need to *extend* the genomic area before this UTR region. In the end our resulted reference genome is the 5' UTR part and the extended region before it. Afterwards we examine publications with public available ChIP seq data and try to find in-silico the most prominent peaks between the ChIP data and our reference.

[1] `http://www.ebi.ac.uk/Tools/msa/tcoffee/`
[2] `http://blast.ncbi.nlm.nih.gov/Blast.cgi`

Figure 2: MSA for L1 subfamily members

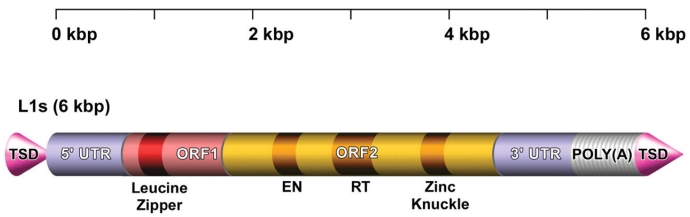

Figure 3: L1 element, (`http://hmg.oxfordjournals.org/content/16/R2/R159/F1.expansion.html`)

2. Results

2.1 Plan of Attack

In order to discover pattern of L1 regulation by epigenetic factors we extract the L1 subfamily (Tf, Gf and A) from mouse genome (mm10) then we extend these genomic regions by 5 kbps using custom python script.[3]Afterwards we examine publications with public available ChIP seq data and mark the most prominent ones in a excel file (also by investigation of the assistant professor). In this point we have the extended genomic regions of L1 subfamily (of interest) and we extract some publicly available ChIP data sets, consequently we need a method (pipeline) to map the ChIP data to our given genomic regions and mine the most significant mappings.
We tried to use *RepeatExplorer* [Novak, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J.] and *Nebula* [Valentina Boeva, Alban Lermine, Camille Barette, Christel Guillouf and Emmanuel Barillot] intergrated into *Galaxy* server [Goecks L. et al.].[4] for the complete ChIP seq data analysis. However we lead to a custom pipeline.
In our custom pipeline firstly we preprocess the data by obtaining the UTR region of L1 subfamily and the extended genomic region upstream/downstream of this family, we unify the ChIP-seq reads, in order to compress the output of the next step. The ChIP-seq data taken from selected publication [Handoko L. et al.]. In the following step we map these unified reads to the UTR region and extended genomic region **separately** using bowtie aligner [Ben Langmead et al.] with parameters discussed in section Methods. Resulting to two different outputs, mapped reads on UTR region and extended genomic region. Then using MACS tool [Zhang et al.] we perform selection of the most significant mappings and manually extract the identified peaks with false discovery rate (FDR) at most 10%. In the next two figures (Figure 4-5) we can see the flowchart of L1 subfamily preprocessing to obtain the UTR region, extended genomic region, the textcolorgreenUTR region and the flowchart for identifying significant matches between these regions and ChIP-seq data.

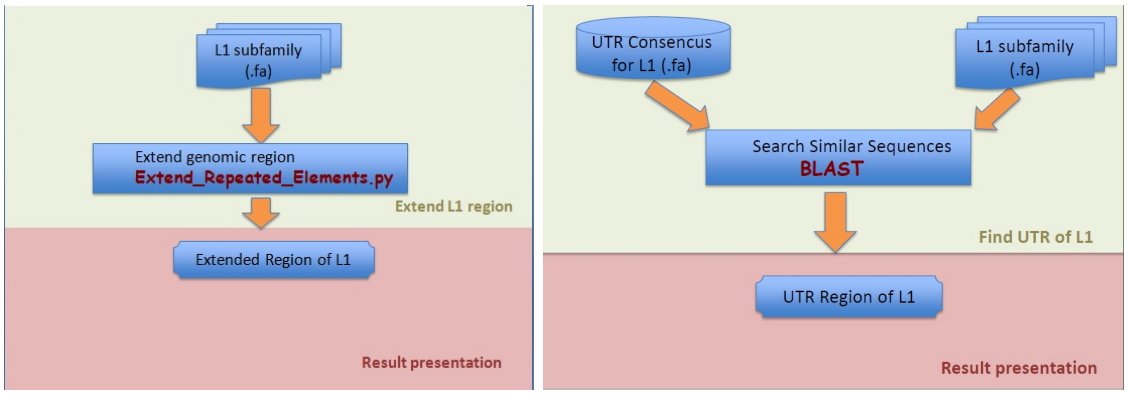

Figure 4: Processing L1 elements to obtain their extended genomic (left) and UTR (right) region.

[3] Please run "python Extend_Repeated_Elements.py" to get help information, tested usage: "python Extend_Repeated_Elements.py mm10 RepeatMasker.gtf 5000". The script is on /home/dmelidis.

[4]The Galaxy server can be started by "bash start_galaxy.sh" (being in /home/jiyu/bin), please run "bash start_galaxy.sh –help" for helpful information.

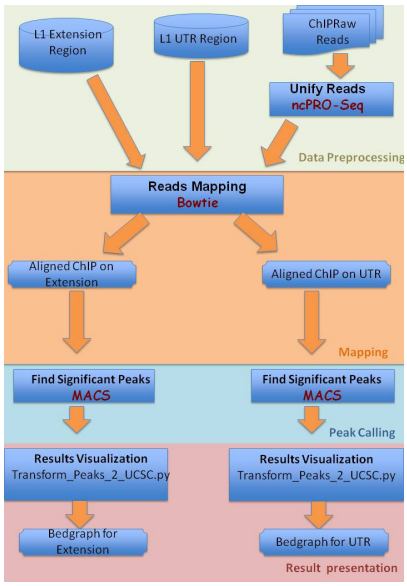

Figure 5: Custom pipeline for discovering epigenetic regulation of L1 subfamily..

2.2 Visualization of Results

The extracted peaks are processed by python script [5] in order to obtain files viewable in the UCSC browser (.bedgraph files). The final peaks for the UTR region and extended genomic region are shown in the next two text boxes. As we can observe the peaks show that a certain peak for one of the region is not in the neighborhood of the other (e.g we cannot find any peak on the UTR region that has in its genomic neighborhood a peak on the extended region or vice versa). Determined by finding miRNA regulation of L1 subfamily we use UCSC browser capability of displaying a genomic region with miRNA and spliced EST. Consequently we upload these two files into UCSC and we are able to produce figures like Figure 6.

File: Selected_Peaks_UTR_UCSC_fold_enrichment.bedgraph

Identified peaks for the UTR region with FDR 10 %
chr num start end % fold enrichment
chr14 44298965 44300354 2.71
chr1 85220575 85223028 2.73
chr3 17185419 17185911 9.68
chr7 20306276 20306787 5.31
chr7 20984588 20985097 5.34
chr7 22301928 22302439 5.31
chr8 59677538 59678213 2.88

File: Selected_Peaks_Extension_UCSC_fold_enrichment.bedgraph

Identified peaks for the extended genomic region with FDR 10 %
chr num start end % fold enrichment
chr11 103116020 103116789 37.08
chr11 103116192 103116789 37.08
chr11 87426529 87427340 20.86
chr13 65703383 65704336 16.76

[5] Script "Transform_Peaks_2_UCSC.py" in /home/dmelidis

chr13 65726447 65727220 40.79
chr13 65817266 65818210 14.41
chr13 66027142 66027999 13.79
chr13 66108301 66109245 14.41
chr13 66199431 66199925 40.79
chr13 66247865 66248713 40.79
chr13 66807664 66808136 31.15
chr17 66364213 66364793 41.53
chr19 11121019 11121543 23.0
chr1 11619353 11622770 4.43
chr1 52233096 52233789 25.21
chr2 34870671 34871588 31.15
chr3 92024486 92024860 35.03
chr3 92024486 92024860 37.08
chr4 126677280 126677902 29.66
chr4 21727378 21727980 26.7
chr6 115676331 115676989 28.92
chr6 57052735 57053215 13.94
chr7 108549246 108551541 4.75
chr7 20344272 20344961 3.9
chr7 39517443 39518187 25.95
chr8 72474805 72475586 34.85
chr8 94036649 94037382 29.66
chrX 99706108 99706929 10.84

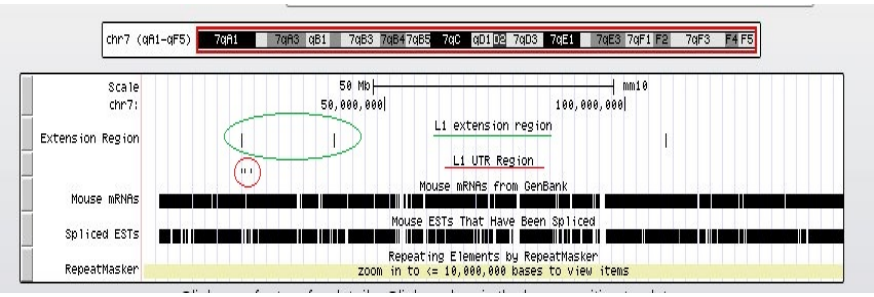

Figure 6: UCSC browser for UTR (red) and extended genomic (green) region in chromosome 7.

3. Discussion

In this section problems during the pipeline design will be discussed in order to consolidate them and not repeat them. Firstly the *RepeatExplorer* program [Novak, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J. does not map regions with highly repeated content into a given reference. In addition the *Nebula* subtools (*FindPeaks* and *MASCS*) cannot be easily used as the first one is not anymore maintained, there is not sufficient documentation for a proper use. For the latter subtool is not straightforward to update it into the installed server *Galaxy* [Goecks L. et al.]. Consequently all the computation was done by custom intallation of MACS program and following the instructive publication [Zhang et al.]. **Another important ascept to consider as future work**, is the estimated fragment length by MACS that is 30 bps, but in the [Zhang et al.] the authors suggest that if the fragment size is less than 100bps the program should be run again with specific parameters (in start of page 1735)
By design and run the pipeline for the sample dataset we show how to de novo *identify epigenetic pattern for regulation of L1 subfamily* in ES mouse cells as supported by example figure 5. After these identification experimental verification is needed to assess the result.
Concluding, I would like to *kindly thank* the assistant professor, the Dr. Jien Yu and in general the lab members *for their support in this rotation and their companionship after the working hours*!

4.Materials and Methods

In this section we will show the proposed pipeline (section 2.1) in more details and also describe the data source for our experiments. As discussed previously our data set is ChIP-seq data from **control** and **RNA Polymerase II variant** input from public available datasets of [Handoko L. et al.].

The commands used for converting SRA files to fasta are:

```
//how to convert the SRA datasets (CEO database) to fasta
./fastq-dump.2.3.3-3 -A SRR172856 ../Paper_DataSets/GSE28247/SRR172856.sra -O ../Paper_DataSets/GSE28247/
./fastq-dump.2.3.3-3 -A SRR172859 ../Paper_DataSets/GSE28247/SRR172859.sra -O ../Paper_DataSets/GSE28247/
./fastq-dump.2.3.3-3 -A SRR172860 ../Paper_DataSets/GSE28247/SRR172860.sra -O ../Paper_DataSets/GSE28247/
./fastq-dump.2.3.3-3 -A SRR172861 ../Paper_DataSets/GSE28247/SRR172861.sra -O ../Paper_DataSets/GSE28247/
./fastq-dump.2.3.3-3 -A SRR172862 ../Paper_DataSets/GSE28247/SRR172862.sra -O ../Paper_DataSets/GSE28247/
```

Trying to explain the extraction of the UTR and extended genomic region in more depth we show the exact tools-commands that were used.

Indicate input files with green and output files in orange

```
//Find the extended region of repeated elements (5Mbps)
python Extend_Repeated_Elements.py mm10 RepeatMasker.gtf 5000
//output file RepeatMasker_extended_5000.gtf
//extract the extensions for the L1 genes
less RepeatMasker_extended_5000.gtf — grep -e "L1Md_T" -e "L1Md_A" -e "L1Md_Gf" > L1_Tf_Gf_A_extended_5000.gtf
bedtools getfasta -fi /home/jiyu/Reference/mm10.fa -bed L1_Tf_Gf_A.gtf -fo L1_Tf_Gf_A.fa — fold -w 60
bedtools getfasta -fi /home/jiyu/Reference/mm10.fa -bed L1_Tf_Gf_A_extended_5000.gtf -fo L1_Tf_Gf_A_extended_5000.fa
— fold -w 60

//Find the UTR region
//Create a blast database from UTR consensus
cd /home/jiyu/Degradome/reference/blast
./makeblastdb -in /home/dmelidis/Repbase_5UTR.fa -dbtype nucl
//Use blast to find inexact matches between the consensus database and the L1 sequence
./blastn -db /home/dmelidis/Repbase_5UTR.fa -query /home/dmelidis/L1_Tf_Gf_A.fa -out UTR_L1.txt -outfmt 6

//Get blast results with similarity at least 90.0 %
python FindSimilarFromBLAST.py UTR_L1.txt 90.0
//output file UTR_L1_sim.txt The given file to be processed is: UTR_L1.txt
In file UTR_L1_sim.txt are saved sequences with sim > 90.0
Lines in input: 54261
Lines in output: 35285

//Get the fasta file for the high similar blast results
samtools faidx L1_Tf_Gf_A.fa
xargs samtools faidx L1_Tf_Gf_A.fa < UTR_L1_sim.txt > UTR_L1_blast.fa
```

In the following we also explain the pipeline for aligning ChIP-seq reads to UTR and extended genomic region and extract the more significant ones, showing the exact tools-commands.

Indicate input files with green and output files in orange

<==========Start of pipeline==========>

<==========Unify reads with groupReads.pl==========>
//Find unique reads
groupReads.pl -i Control_28247.fa -f "fa" > Control_28247_red.fa
groupReads.pl -i Pol_II_28247.fa -f "fa" > Pol_II_red.fa

<==========Align reads using bowtie==========>
//Map Control reads to UTR region
//create the BW index for UTR region
bowtie-build UTR_L1_blast.fa UTR_L1_blast

//Map Control to UTR region
nohup bowtie -a –best –strata -S -f -p 14 –chunkmbs 200 UTR_L1_blast Control_28247_red.fa Control_28247_red_UTR_blast.sam
reads processed: 12056881
reads with at least one reported alignment: 476991 (3.96%)
reads that failed to align: 11579890 (96.04%)
Reported 1178662264 alignments to 1 output stream(s)

//Map Pol II reads to UTR region
nohup bowtie -a –best –strata -S -f -p 14 –chunkmbs 200 UTR_L1_blast Pol_II_red.fa Pol_red_UTR_blast.sam
reads processed: 10359301
reads with at least one reported alignment: 206147 (1.99%)
reads that failed to align: 10153154 (98.01%)
Reported 729095155 alignments to 1 output stream(s)

//transform sam to bam and get the bam's index
samtools view -bS Control_28247_red_UTR_blast.sam > Control_28247_red_UTR_blast.bam
samtools sort Control_28247_red_UTR_blast.bam Control_28247_red_UTR_blast_sorted
samtools index Control_28247_red_UTR_blast_sorted.bam Control_28247_red_UTR_blast_sorted.bai

//transform sam to bam and get the bam's index
samtools view -bS Pol_red_UTR_blast.sam > Pol_red_UTR_blast.bam
samtools sort Pol_red_UTR_blast.bam Pol_red_UTR_blast_sorted
samtools index Pol_red_UTR_blast_sorted.bam Pol_red_UTR_blast_sorted.bai

//Map Control reads to extended region
//create the BW index for extended region
bowtie-build L1_Tf_Gf_A_ext_5000.fa L1_Tf_Gf_A_ext_5000
bowtie -a –best –strata -S -f -p 25 –chunkmbs 200 L1_Tf_Gf_A_ext_5000 Control_28247_red.fa Control_28247_red.sam
reads processed: 12056881
reads with at least one reported alignment: 1757173 (14.57%)
reads that failed to align: 10299708 (85.43%)
Reported 374357567 alignments to 1 output stream(s)

//Map Pol II to extension region
nohup bowtie -a –best –strata -S -f -p 25 –chunkmbs 200 L1_Tf_Gf_A_ext_5000 Pol_II_red.fa Pol_II_red.sam
reads processed: 10359301
reads with at least one reported alignment: 977890 (9.44%)
reads that failed to align: 9381411 (90.56%)
Reported 204981294 alignments to 1 output stream(s)

//transform sam to bam and get the bam's index
samtools view -bS Control_28247_red.sam > Control_28247_red.bam
samtools sort greenControl_28247_red.bam Control_28247_red_sorted
samtools index Control_28247_red_sorted.bam Control_28247_red_sorted.bai
//transform sam to bam and get the bam's index
samtools view -bS Pol_II_red.sam > Pol_II_red.bam
samtools sort Pol_II_red.bam Pol_II_red_sorted
samtools index Pol_II_red_sorted.bam Pol_II_red_sorted.bai

<==========Identify significant peaks using MACS==========>
//Peak calling for extended region
nohup macs14 -t Pol_II_red.bam -c Control_28247_red.bam -g mm -n Pol_II_28247_ext_red
//This command also outputs Pol_II_28247_ext_red_peaks.xls
//Peak calling for UTR region
nohup macs14 -t Pol_red_UTR_blast.bam -c Control_28247_red_UTR_blast.bam -g mm -n Pol_UTR_red_blast
//This command also outputs Pol_UTR_red_blast_peaks.xls

//Select manually from .xls the peaks with FRD $\neq$ 100 %
//and save the results into the files Selected_Peaks_Extension.txt and Selected_Peaks_UTR.txt
//Convert extract files with FDR $\leq$ 5.0% with Transform_Peaks_2_UCSC python script
//Extract peaks for UTR region
python Transform_Peaks_2_UCSC.py Selected_Peaks_UTR.txt 5.0
//outputs Selected_Peaks_UTR_UCSC_fold_enrichment.bedgraph
//Extract peaks for extended region
python Transform_Peaks_2_UCSC.py Selected_Peaks_Extension.txt 5.0
//outputs textcolororangeSelected_Peaks_Extension_UCSC_fold_enrichment.bedgraph

//Upload .bedgraph to my data in UCSC

<=================End of pipeline====================>

References

[Leslie A. Pray] Leslie A. Pray (2008) Transposons: The Jumping Genes Nature Education (1):1 2008.

[Eric M. Ostertag et al.] Eric M. Ostertag and Haig H. Kazazian Jr (2001) Biology of Mammalian L1 Retrotransposons Annu. Rev. Genet. 35:501-38 2001.

[Miki, Y., et al.] Miki, Y., et al. (1992) Disruption of the APC gene by a retrotransposal insertion of L1 sequence in colon cancer Cancer Research 52, 643-645 1992.

[Novak, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J.] Novak, P., et al. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. Bioinformatics 2013.

[Valentina Boeva, Alban Lermine, Camille Barette, Christel Guillouf and Emmanuel Barillot] Valentina Boeva, et al. (2012) Nebula–a web-server for advanced ChIP-seq data analysis Bioinformatics 28: 2517-2519 2012.

[Goecks L. et al.] Goecks L. et al. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible and transparent computational research in the life sciences Genome Biol. 11 R 86 2010.

[Handoko L. et al.] Handoko L et al. (2011) CTCF-mediated functional chromatin interactome in pluripotent cells Nat. Genet. 43(8):815 2011.

[Ben Langmead et al.] Ben Langmead et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome Genome Biology 10:25 2009.

[Zhang et al.] Zhang et al. (2008) Model-based Analysis of ChIP-seq data Genome Biol 9 9 R137 2008.