Final Report STM

Il Fruitore: **Sole Gatto**

Istituto di afferenza: "Istituto per le Applicazioni del Calcolo" M. Picone (*IAC*)

con qualifica: Associata

Descrizione dettagliata dell'Istituzione ospitante: MRC (Medical Research Council) Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 0QH, UK.

Dipartimento di afferenza: ICT - Tecnologie dell'Informazione e delle Comunicazioni

Titolo del programma: **Statistical algorithms for the analysis of ChIP-sequencing data from SOLiD platform.**

The aim of my stay in Sarah Teichmann's laboratory for three weeks was to collaborate to the set-up of a reliable computational strategy for the analysis of ChIP-seq data and to define a statistical method for their integration with gene expression data.

Once there I illustrated the data collected in several ChIP-seq experiments with 3meK4H3, 3meK27H3 and RNA Pol II on ICF cell lines performed from Maria Matarazzo's group at the IGB-CNR. We started the discussion from the analysis I had already performed in order to improve the computational pipeline. In particular we discussed about EPIChIP (1), a novel software developed by Sarah Teichmann's group, which allows a statistically guided distinction between experimental background and enriched signal from ChIP-seq experiments. The algorithm is based on mixture model consisting of two components. By fitting the observed signal within specifically located windows, it will be possible to classify all annotated genes as modified versus unmodified (with a false discovery rate control). Additionally, this tool permits to integrate the peaks obtained with the gene expression results. After performing such analysis we observed that expression data from the already existing microarray (2) were not correlating properly with the 3meK4H3 peaks like we expected. Since the data in (2) were not obtained on the same cells we used to perform ChIP-seq, we decided to produce expression data from our cells using a novel sequencing approach: RNA-seq. Such approach is in fact comparable in terms of resolution and quality to our ChIP-seq data, and it is in principle unbiased. In order maximize the advantages of the novel technology we decided to prepare paired-end libraries. In this way I prepared the RNA libraries taking advantage of the already set protocols of Dr Teichmann's group and I started the sequencing of the Poly-A transcriptome of the three cell lines we are interested in before coming back to Napoli. The outcome of these experiments will be available very soon.

In the mean time we continued the comparison of the computational pipelines. One of our aims was to compare the enrichment of the peaks among the different samples in order to identify epigenomic signatures that can be associated to the ICF syndrome. We decided to use DEseq (3), an algorithm that models reads over dispersion using a negative binomial model, therefore it allows to compare two samples using the information available in the replicated experiments. To do this we first identified the enriched peaks with SICER (4), and then we counted the number of reads falling in the selected sequences. We required this step because DEseq doesn't use the inputs so it doesn't make a proper background subtraction and subsequently peaks detection. Simultaneously, we performed also ChIP-peak difference using the novel version of SICER. Such version allows the comparisons of two samples and the use of the input for normalization purposes, however it does not allow the use of replicates.

After such discussion we started planning the proper modification that we will make to fit the method to our purposes during the continuation of our collaboration. When all data will be available, we will compare the two approaches to chose the one that perform better for all the signatures under evaluation. Completing such step will provide different sets of enriched regions that can characterize one experimental condition. Once we will have the list of enriched regions we will proceed making pair-wise comparisons and we will obtain a list of differentially enriched peaks. These are the regions where the binding of the protein in exam is altered

between different conditions. In order to connect ChIP enriched regions with expression data and other biological relevant features we will use an R package (ChIPpeakAnno, (5)) that allows the annotation of a wide series of genomic tracks. We will develop a series of scripts for further analysis and data visualization, when required. Then a suitable regression approach will be used to integrate ChIP-seq data with RNA-seq data.

In alternative to this approach we will use EpiChIP. By construction EpiChip does not allow genome wide peak detection, but it mainly performs peak classification around very specific pre-selected regions. On the basis of prior biological knowledge we will select potential peaks around the transcription start site and the transcription termination site of each gene.

In this way we will be able to compare different samples again with DEseq and in both case we will be able to connect such information with RNA-seq data.

Finally, to compare the results from the different pipelines we will perform gene ontology and biological validation.

I can firmly assess that this short term visit has been highly important to me and to my group to start a precious collaboration with Sarah Teichmann's group and to gain more confidence on data analysis. It has been also surprisingly positive to our group the possibility to perform an RNA-seq experiment to make our data more homogeneous and comparable.

1. Hebenstreit D, Gu M, Haider S, Turner DJ, Liò P, Teichmann SA. (2011) EpiChIP: gene-by-gene quantification of epigenetic modification levels. Nucleic Acids Res. 2011 Mar 1;39(5):e27. Epub 2010 Dec 3.

2. Jin, B., Tao, Q., Peng, J., Soo, H. M., Wu, W., Ying, J., Fields, C. R., Delmas, A. L., Liu, X., Qiu, J., and Robertson, K. D. (2008) DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome lead to altered epigenetic modifications and aberrant expression of genes regulating development, neurogenesis and immune function. Hum Mol Genet 17, 690-709.

3. Anders S, Huber W. Differential expression analysis for sequence count data. (2010) Genome Biol.;11(10):R106.

4. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W.(2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics;25(15):1952-8. Epub 2009 Jun 8.

5. Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, Green MR. (2010) ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics.;11:237.