Jason Bedford
Parker Sulkowski
Cong Liang
Gyong Ju (Liz) Min
Nur-Taz Rahman

Team #3: Final Project Proposal

1. What is the question to be addressed?

The Serum Response Factor (SRF) is a ubiquitously expressed transcription factor that serves as a master regulator of cytoskeletal genes. It binds to a consensus DNA sequence, CC(AT)6GG (CArG box), in the promoter of its target genes. Since SRF is widely expressed, it often interacts with cofactors that are tissue-specific or activated during a specific time in development. For example, SRF directly interacts with the Myocardin family of transcriptional coactivators to regulate muscle differentiation genes as well as the E26 transformation-specific (ETS) to regulate growth-induced genes. Recently, SRF was shown to interact with Megakaryoblastic Leukemia 1 (MKL1), a Myocardin family member, to drive megakaryocyte differentiation, but the roles of SRF in cells of the blood lineage are largely unknown. The aim of this project is to use a bioinformatic approach to determine whether SRF regulates gene expression in K562 and GM12878 cells, which are erythroleukemia and lymphoblast cell lines, respectively, and to compare the genes that are regulated by SRF in the two different cell lines. To this end, we will analyze the ChIP-seq and RNA-seq data that are publicly available for each of these cell lines by mapping the SRF binding site, comparing the list of genes that are up- or down-regulated, and utilizing the GO analysis tool to examine the biological pathways in which the genes are involved.

2. Where are the starting data from, what are the sizes of the datasets, how were the datasets generated, and in what format will they be?

K562 and GM12878 are two widely studied cell lines. Both RNA-seq and SRF chip-seq data are available publicly. SRF chip-seq has been done by HAIB (HudsonAlpha Institute for Biotechnology) in a series of cell lines. And RNA-seq data are also available from ENCODE project. CSHL (Cold Spring Harbor Lab) has finished long RNA-seq for the expression profile of different cell lines including data for a SRF shRNA knockdown in K562 cells. The High Throughput RNA sequencing was done on the Illumina Genome Analyzer IIx platform and the CHIP seq data was also generated on Illumina Platform as well specifically the Illumina Genome Analyzer.. We plan to start from the original sequencing data (raw sequence reads in a .fastq format) to make an integrated pipeline of this analysis. The sequencing data for RNA-seq would be around 5GB, and about 1GB for chip-seq.

Here are the accession links for our specific data sets:
ChIP-seq: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32465
RNA-seq: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33816

3. What are the steps in your analysis pipeline? What tools will you be using? How many scripts will you need to write? What is your estimate of the amount of disk space required? (Nur)

For our project, we will have two kinds of data sets: for each cell type, we will have ChIP-seq FASTQ files and RNA-seq FASTQ files. To work with ChIP-seq, we will use **bowtie** to align our reads against the reference genome (hg19). This will create files in the .sam format, which we will convert to the more computer-friendly .bam format using the **samtools**. The next step will be to call peaks using **MACS**. This will create wiggle files and the more important .bed files, which we can do our analyses on. Using the **bedtools** utility, we will see overlaps and non-overlaps in the ChIP-seq data set between K562 and GM12878 cells. The same can also be achieved with the **UCSC genome browser**, with the added

benefit of being able to visualize the called peaks on the genome browser as our own individual "track".

For the RNA-seq data sets, we will use the aligner **TopHat** against the reference genome hg19, which will produce the outputs in the .bam format. We can further our analyses for the differential expression of genes within the two cells by using **CuffDiff**. **RStudio** allows us to nicely visualize differentially expressed genes between different populations, using the **cummeRbund** package. The package can be installed from bioconductor, and we can write an R script to generate a heatmap, which can show us the top 50-80% differentially expressed genes within K562 and GM12878.
In keeping with our goal of finding out how SRF regulates genes in these two cell types, we want to study how binding of SRF alters gene expression. To that end, we propose to write a **Perl** script, which will, broadly speaking, allow us to achieve two things: One, select the regions to which SRF is binding significantly in the two cell types; and two, for those regions, give us an output of the expression level of the neighboring genes (in terms of FPKM). From the .bed files created from the ChIP-seq data sets, we will be able to create a data chart for the genes associated to the regions where SRF is binding for each cell type using the online tool **GREAT**. Using this as an input file for our Perl script, we will be able to obtain the second and final output from our script.

To help us make meaningful observations from our data, we will also obtain **GO term analysis** on the tool **DAVID**. This analysis will give us a global view of the SRF-dependent gene expression in each cell type. Any similarities/differences between the pathways that are regulated by SRF in the cells will be apparent from the GO analysis.

Given that we are a team of 5, and we will handle at least 6 large data sets (two replicates of ChIP-seq data and one RNA-seq data set each cell type) we estimate a total usage of about 80GB for accomplishing our entire project. As detailed above, we will be writing at least 2 scripts for our project: an R script, and a Perl script.

4. What format(s) will the output data be in? How will these data help you answer your proposed question?

We will get a list of differentially regulated genes in the comparison between K562 and GM12878 (which genes are turned on in K562 and which genes are turned off). GO analysis will provide us a list of enriched function pathways which are differentially regulated in those two cell types. This will provide us an idea of the difference between lymphoblastoid cell line and myelogenous leukemia cell line and the role of SRF in these differentially regulated pathways. Some interesting pathways might be cell cycle pathways, apoptosis pathways.