BISC-577: Project # 3

Due on Tuesday, April 21, 2015

Saket Choudhary
2170058637

Contents

Question # 1

Chip-Seq Experiments: Chip-Seq experiment couples chromatin immunoprecipitation with high throughput DNA sequencing. It is used for identifying binding sites of transcription factors and for idetifying histone related modifications. 'Chip' step involves cross-linking proteins and DNA making the proteins immobile. This is followed by fragmentation through sonication/endonucleases, generatiing 100-300bp fragments. The protein of interest is then enriched using a specific antibody that is know to bind selectively to just this protein. The separated fragments contain(mostly) of the protein bound DNA sequences. DNA can be separated by reversing the cross-links which can then be analyzed for its abundance, computationally. To obtain sufficient signal, $large$(10M+) number of cells are required.
Of the two protocols Nano-Chip-Seq and LinDA, the difference exists at the amplification stage. Nano-Chip-Seq makes use of custom primers during PCR amplification containing a specific restriction site that permits direct addition of illumina sequencing adapters These primers form a hairpin structure preventing self-annealing.
LinDA on the other hand makes use of an RNA polymerase from the T7 bacteriophage.
Since Nano-Chip-Seq relies on PCR amplification and custom primers, these experiments might have technical bias or probably even overrepresentation of primer sequnces which should be checked for in the data analysis stage.

Question # 2

One of the major sources of bias in Chip-Seq studies arises due to the fragmentation step. Fragmentation is necessary to ensure only fragments bound to protein are purified. Sonication is known to be more effective in open chromatin regions and hence regions flainkin euchromatin will shear easily than heterochromatin regions. Transcription factors bind more easily to the open chromatin region which also shears easily and hence gives rise to preferential bias.
"Input" Dna protocol involves isolation of sample that has been crosslinked and sonicated but not immunoprecipitated. The "IgG" control is a "mock" Chip reaction that is guranteed to be random. It works by using a 'control' antibody that will bind to non-nuclear proteins randomly.
Presence of controls("input" or "IgG") can be used to estimate 'background' rate of non-specific binding for transcription factors/histones which then can be used to filter out the false positive peaks from the analysis samples.
The control dataset is an "input" control drawn from the human esophageal epithelial cell line and thus was islotaed post sonification(without immunoprecipitating)

continued on next page...

Question # 3

H3K9Me3: http://www.ncbi.nlm.nih.gov/sra/SRX849433[accn]
Run: SRR1768294
SRA size: 349M
FastQ size: 2.6G
Single End reads of 36bp size.

Total Reads: 18400047

H3K4Me3: http://www.ncbi.nlm.nih.gov/sra/SRX849427[accn]
Run: SRR1768267
SRA size: 439M
FastQ size: 3.4G
Single End reads of 36bp size.
Total Reads: 23514026

Project Name: Conserved epigenomic signatures between mouse and human elucidate immune basis of Alzheimer's disease (house mouse) http://www.ncbi.nlm.nih.gov/bioproject/PRJNA273302
Organism: Mouse

Question # 4

SRR1768267.fastq
Total Reads: 23514026
Unaligned Reads: 1426753 (6.07%)Aligend Reads(Exactly once): 17620103 (74.93%)
aligned exactly 1 time
Aligend Reads(Aligned more than once): 4467170 (19.00%) aligned ¿1 times
Total alignment rate: 93.93% overall alignment rate
Time: 18m1.391s
Sam file size: 4G

SRR1768294.fastq
Total reads: 18400047
Unaligned Reads: 1440302 (7.83%)
Aligend Reads(Exactly once): 9027568 (49.06%) aligned exactly 1 time
Aligend Reads(Aligned more than once): 7932177 (43.11%) aligned ¿1 times
Total alignment rate: 92.17% overall alignment rate
Time 21m3.115s
Sam file size: 3.1G

Mapped to mm10.

continued on next page...

Question # 5

Program used: MACS(v1.4)

Parameters comnfigurable:

1. *gsize*: Genome size of the organism. This is made use in the $p - value$ calculations and hence may impact the number of peaks depending on the threshold. It is more likely to see peaks in a smaller genome than a large one.

2. *pvalue*: $p - value$ cut-off for defining a peak. Default is 10^{-5}, but more stringent cut-offs might be required for noisier datasets

3. *nolambda*: MACS models reads distribution as poisson distribution. A "control" if present can be used to estimate the 'background' λ. If no control is present, the background λ is fixed.

4. *nomodel*: MACS models the shifting size of Chip-Seq tags(which often are shifted to $3'$ end, this size being unknown) to precisely locate the binding sites, which might be difficult to model in case of Chip-Seq's broad peaks. So though a precise location is possible by enabling *nomodel*, the data might not necessarily show bimodal pattern.

Each file tool around one minute to run. The output were bedGraph and bed files.
Bedgraph files are tab delimited files that stores in each row the chromosome number, start position, end position and the read count mapping to these positions. These positions are a superset of the positions appearing in the bed files.
Bed files are also tab delimited with the first column as the chromosome positions the next two as the start and end positions of peak in that chromosome and the fourth column as the $-10\log(10pvalue)$. The summits file has the height information for peaks.

Question # 6

H3K9Me3
Number of peaks: 3338
Mean peak length: 1185.777
Median peak length: 814.5
Max: 46335 (Could be all noise)

H3K4Me3
NMumber of peaks: 28823
Mean peak length: 1981.701
Median peak length: 1594
Max: 54633 (Could be all noise)

The number of peaks detected in H3K9Me3 state are less than that in H3K4Me3, thus pointing that H3K9Me3 impacts the heterochromatin(compact) region while H3K4Me3 must be associated with euchromatin thus indicating H3K4Me should be associated with actication while H3K9Me3 might be asociated with repression. The datasets do not seem to be high quality because of the inherent noise and 'low peaks' present. Mthfs was one of the genes showing a peak for H3K4Me3, pointing that the gene is probably upregulated.

continued on next page...