AARHUS UNIVERSITY

Computational Analysis of ChIP-Seq Data

Jinjie Duan

7/10/2010

Table of Contents

Introduction

Understanding of transcriptional regulation mechanisms is of fundamental importance to the study of biological process such as development, drug response and disease pathogenesis [1].Protein-DNA interactions play vital roles in the transcriptional regulation. Therefore, identifying the interaction between transcription factors (TFs) and their binding DNA is essential to understand many biological processes. Several experiments give information on the TF-target gene interactions. One such experiment, chromatin immunoprecipitation (ChIP) followed by massively parallel sequencing (ChIP-Seq) is a new technology to map protein-DNA interaction in genomes [2] and it is based on the enrichment of DNA associated with a protein of interest(Figure 1).This experiment begins with the cross-linking of protein-DNA interaction using formaldehyde. Then, cells are lysed and DNA is fragmented. Fragments bound by protein of interest are then bound by antibody and precipitated. After reversing the cross-links and purifying these DNA fragments, a DNA sample called “ChIP sample” is obtained. In many experiments, a control sample is prepared in parallel using a similar protocol where an aliquot of sheared cell lysate is not immunoprecipitated but is otherwise processed normally. This DNA is termed input DNA. Compared to the input DNA, ChIP sample is enriched in DNA fragments bound by the protein of interest.

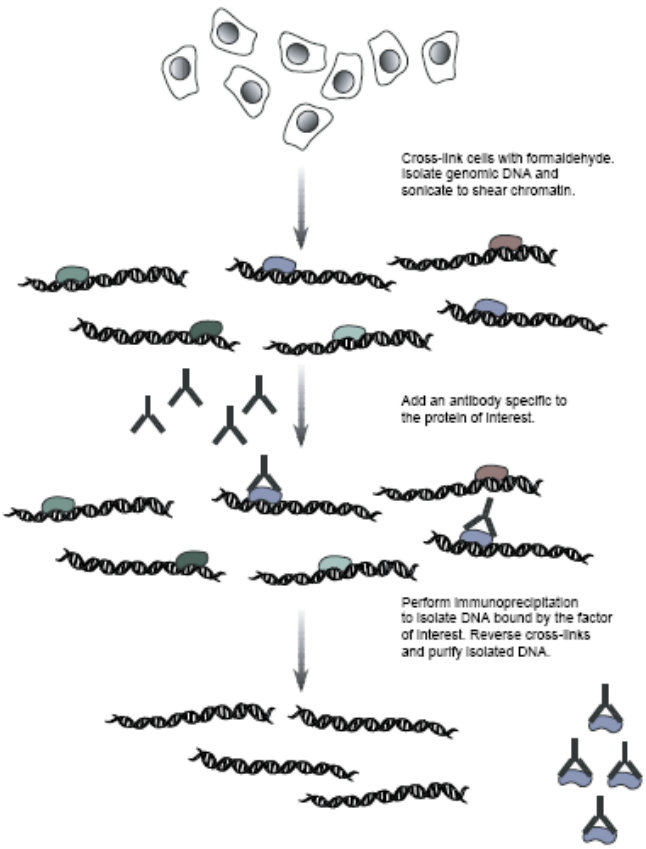


Figure 1 An overview of the chromatin immunoprecipitation (ChIP) procedure [3]. Cells are initially treated with a cross-linking agent that covalently links DNA-interacting proteins to the DNA. The genomic DNA is then isolated and sheared, typically by sonication, into a suitable fragment size distribution (200-600bp is typically used for ChIP-Seq). An antibody that specifically recognizes the protein of interest is then added and immunoprecipitation used to isolate appropriate protein-DNA complexes. The cross-links are then reversed and the DNA fragment purified.

Processing and analyzing ChIP-Seq

In this section, I describe the key issues and concepts involved in ChIP-Seq data analysis. A flowchart of the central steps in the ChIP-Seq procedure is shown in Figure 2.The raw data for chromatin immunoprecipitation followed by sequencing (ChIP-Seq) is generated by the next-generation sequencing (NGS) platform, such as Illumina and ABI SOLiD. These reads are short in length (around 25~30bp; the latest platform yields reads longer up to 50~100bp) and extreme high throughput (around 700MB to 1GB per lane).

Basic analysis of ChIP-Seq results is relatively straightforward. By mapping all raw reads to the reference genome, the uniquely mapped reads are retained. Using uniquely mapped reads from the ChIP profile and a control profile which is usually created from input DNA, peak calling generates a list of enriched regions (peaks). For ChIP-Seq, the most common follow-up analysis is focusing on discovery of binding sequence motifs which can represent transcription factor binding sites (TFBSs).This *de novo* motif finding corresponds to identifying over-represented subsequence in bound regions to a background frequency model.

Here I introduce some of the popular methods which lead to a better understanding of discovery motif bound with TFBS according to ChIP-Seq bioinformatics analysis.

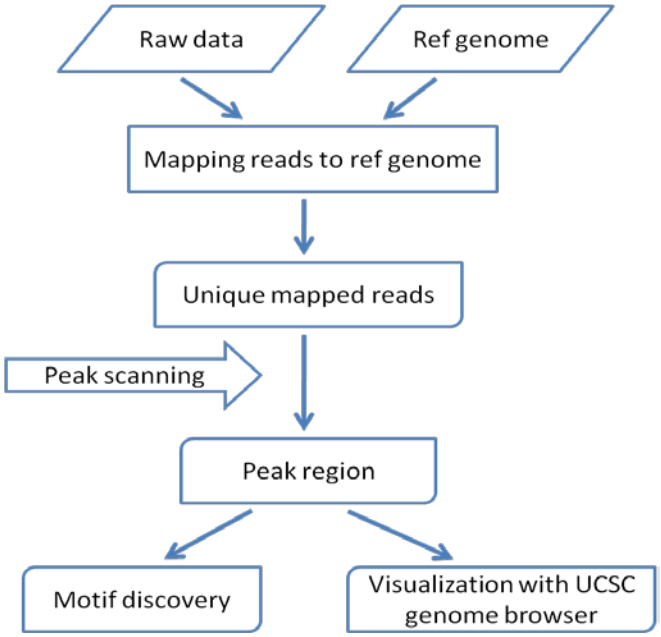


Figure 2 Flow chart of the central steps in the ChIP-seq procedure. By mapping the raw sequence reads to reference genome, the unique mapped reads were remained for further analysis; The significant enriched regions were detected from ChIP data compared to the control data; The binding motifs then were identified by over-represented subsequence in peak regions.

Alignment of Sequence Reads

The first step of ChIP-Seq data analysis is to map reads to a reference genome. Alignment of reads should allow for a small number of mismatches (2~3 mismatches) due to sequencing errors, SNPs and indels or the difference between the genome of interest and the reference genome.

Here I describe an algorithmic approach based on spaced seed alignment. Several associated software programs such as MAQ [4], Eland [5] and RMAP [6] apply spaced seeding techniques, requiring one or several hits per read.

First, all reads are loaded into memory and the first 28 bp of reads were indexed, so it is able to guarantee to find no more than 2-mismatch seed hits.

Second, a mapping quality for each alignment is calculated by measuring the probability of mapping error. Mapping quality is given by the Phred-scaled probability [7] that a read alignment maybe wrong. Given L-long reference x and l-long read z, assume that sequencing errors are independent at different sites of the read, the probability $p(z|x, u)$ of z coming from the position u equals the product of the error probabilities of the mismatched bases at the aligned position. For example, if read z mapped to position u has two mismatches: one with phred base quality 10 and the other with 30, then $p(z|x, u) = 10^{-(10+30)/10} = 0.0001$.

Assuming a uniform prior distribution $p(u|x)$, the Bayesian formula is applied to calculate the posterior probability $P_s(u|x,z)$, and the mapping quality is, $Q_s(u|x,z) = -10log_{10}[1 - P_s(u|x,z)]$

Third, perfect match reads are reported as "unique", and by also placing the repetitive reads randomly amongst equally good alternatives with a low mapping score, instead of discarding them. This avoids any ambiguity and provides more data for the subsequent analysis.

Identification of enriched regions

After mapping sequence reads to the reference genome, the next step is to identify regions that are significantly enriched in the ChIP sample when compared to the control. Here, I describe an algorithmic approach of modelling the shift size of sequence reads and using a dynamic parameter to detect peaks (the signals of putative protein binding) from ChIP-seq data.

ChIP-Seq reads are enriched near TFBSs. Therefore, by using mapped reads, a reasonable model can be built to detect peaks. In this approach [8], the tags belonging to a forward/reverse strand are shifted ½ of fragment size right or left. It also uses a dynamic Poisson distribution to effectively capture local biases in the genome.

Through modelling the shift size of reads and detecting the significant enriched DNA regions, the peaks can be found.

Peak shift estimation

DNA fragments from a ChIP experiment are sequenced from the 5' end. Therefore, the alignment of these reads to the genome forms two peaks (one on each strand) that flank the binding location of the protein of interest. The binding site can then be interpolated between these peaks (Figure 3).

This algorithm requires a reference genome size (gsize), a sonication size (bandwidth) and a high-confidence fold-enrichment (mfold). The mfold parameter is used to select the regions within MFOLD range of high-confidence enrichment ratio against background to build model. This parameter works by assuming totally N uniquely mapped reads are obtained in a ChIP sample. The reference genome is scanned using a 2*bandwidth window, thus a $\frac{N*(2*\text{bandwidth})}{\text{gsize}}$ number of reads is expected to be seen in a random scan window. If a scan window has T reads inside, then the fold enrichment is computed as $\frac{T}{N*(2*\text{bandwidth})/\text{gsize}}$. If the fold enrichment is larger than the specified mfold, this window will be termed 'significant'.

Through scanning of the whole genome, the high-quality enriched regions can be found. After the scan, 1000 samples of these regions are randomly selected, their forward and reverse strand reads separated, and aligned to the genome (Figure 3). The distance between the summits of the forward strand and reverse strand peaks is defined as 'd'. All the reads are shifted by d/2 toward the 3' ends to the mostly likely transcription factor binding sites.

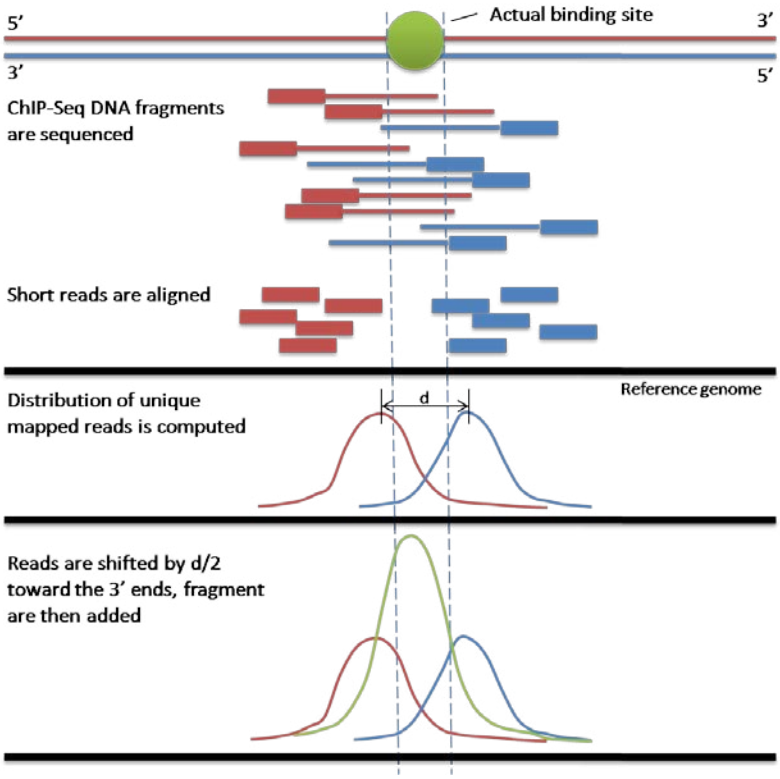


Figure 3 Sequenced short reads from ChIP-Seq experiments are first mapped onto the reference genome. Reads are sequenced from both ends of DNA fragments. Therefore, when surrounding a TFBS, some reads aligned with forward strands form a peak upstream of TFBS, and some reads aligned with reverse strands form a peak downstream of TFBS. The distance between the summits of these two peaks is defined as 'd'. Shifted all reads by d/2, a new peak with potential binding sites was generated.

Peak detection

After shifting all reads by d/2, this algorithm scans 2*bandwidth windows across the genome to find potential peaks with a significant enrichment. With the current genome coverage of most ChIP-Seq experiments, reads distribution along the genome could be modelled by a Poisson distribution. λ_{BG} is a parameter of this model, which is equal to the mean and the variance of the distribution[8]. Therefore, these potential peaks can be detected by Poisson distribution p-value based on λ_{BG}.

We often observe that there are tag distributions with biases (Figure 5) in control sample. Several sources of bias can be: local chromatin structure, DNA amplification

and sequencing bias or genome copy number variation [8]. Therefore, a dynamic parameter λ_{local} is used instead of λ_{BG} , λ_{local} = max (λ_{BG}, [λ_{1k},] λ_{5k}, λ_{10k}), where λ_{1k} , λ_{5k} and λ_{10k} are λ estimated from the 1kb, 5kb or 10kb window centred at the peak location in the control sample, or from the ChIP-Seq sample when a control sample is not available (in which case λ_{1k} is not used). When using λ_{local} to compute the p-value of each potential peaks, if the p-value of a candidate peak is above a threshold p-value (default 10^{-5}), the candidate peak should be removed due to local biases; otherwise the candidate peaks are detected and reported.

Motif Finding

For ChIP-Seq data, the most common follow-up analysis is the discovery of binding sequence motifs which can represent TFBSs. Here I describe the expectation maximization algorithm (EM algorithm) for motif finding.

The EM algorithm for identifying motifs (a pattern that recurs in the sequence) in an unaligned biopolymers was introduced by Larence and Reilly [14] as a means of solving a supervised motif learning problem. The EM algorithm takes as input a set of unaligned sequences and a motif length *W* and returns a probabilistic model of shared motif, assuming that each sequence in the dataset contains a single example of motif. Due to this fact, the algorithm is also called “one-occurrence-per-sequence” model or “one-per” model”. Before the iteration begins, the EM presents an initial guess where the motif appears (the starting offset) in each sample.

It then estimates the probability that the shared motif stats in position *j* in sequence *i* in the dataset. These probability estimates Z_{ij}(Z is used to refer to the matrix of offset probabilities Z_{ij}), are then used to re-estimate the probability of letter *l* in column *c* of the motif m_{lc} (*m* refers to the matrix of letter probabilities m_{ij}),for each letter in the alphabet and $1 \leq c \leq W$. The initial motif is represented by the frequencies of each base in each column of the site along with the frequencies of each base outside the site (background).

The expectation and maximization steps are performed consecutively to simultaneously estimate the probability of each possible starting point of the motif(s) in the sequence and discover a model of the motif until the algorithm converges as it is guaranteed to do [15]. The expectation step uses the proposed motif to estimate the probability of finding that motif at any position in each of the sequences. During the second step, maximization, the probabilities (from step one) can then be used to re-estimate the proposed motif. This is done iteratively until the motif model no longer changes.

The pseudo-code for the EM algorithm and the re-estimation details are described in [15], using the following variables

S unaligned set of sequences $S_1, S_2, \ldots, S_i, \ldots, S_n$ each of length L

W width of motif

z matrix of probabilities that the motif starts in position j in S_i

m matrix representing the probability of character c in column k (the character c will be A, C, G, or T for DNA sequences or one of the 20 protein characters)

```
EM (S, W) {
    choose starting point and initial value for m
    repeat until m stops changing {
        compute z from m        //the estimation step
        compute m from z        //the maximization step
    }
    Return m, z
}
```

The Multiple EM for Motif Elicitation (MEME) [13] algorithm enhances the basic EM approach to offer the following novel contributions:

- Increased likelihood of finding globally optimal motifs. By using subsequences of the nucleic or amino acid sequence input as starting points, EM is guaranteed to find an optimal local maximum. And, although it is not guaranteed, EM tends to converge to the global maximum since the subsequences themselves are taken from the data where the similar and optimal value(s) for the shared motif exist.
- Allowance for multiple, different shared motifs to be discovered in the same inputted sequences. This is achieved by probabilistically erasing shared motifs discovered by EM and then repeating EM to find subsequent motifs.

The MEME algorithm was first described in [15], listed below, but has also been parallelized [16].

```
MEME(S, W, COUNT) {

    for i = 1 to COUNT {

        for each subsequence in S {

                run EM for 1 iteration with starting point derived from this
                subsequence.

                choose model of shared motif with highest likelihood

                run EM to convergence from starting point which generated that model

                print converged model of shared motif

                erase appearances of shared motif from dataset

        }

    }

}
```

Results and Discussion

In this report, I have described methods to detect the DNA motif from ChIP-Seq. The most important step in motif finding is to assemble a 'dataset' of DNA sequences in which you will discover motifs. Therefore, it is critical to use all available background information to select the sequences that are likely to contain motifs.

I used ChIP and input DNA datasets of human STAT1 in my project. Using the MAQ program and allowing a maximum of two mis-matches, 33.4% (26.7 million) of the ChIP reads were uniquely mapped back to the human reference genome (hg18), and 46.4% (23.4 million) of the Input DNA reads were uniquely mapped(Table 1).

	# of total ChIP-Seq reads	# of multiple mapped reads	# of unique mapped reads	% of multiple mapped reads in total reads	% of unique mapped reads in total reads	% of unmapped reads in total reads
STAT1 ChIP	80,035,849	7,235,361	26,731,492	9.04%	33.40%	57.56%
Input DNA	50,515,792	7,726,505	23,435,631	15.30%	46.39%	36.84%

Table 1 Here I summarize the results obtained from Illumina sequencing of the STAT1, and input DNA datasets. The total number of reads generated for each sample is divided into those that map uniquely in the human genome (hg18), those that map to multiple locations and those that do not map at all.

The uniquely mapped reads were processed to detect significant peaks as described in 'Data and Methods' section. Following MACS [8], I identified a total of 25,350 significantlySTAT1 enriched peaks. By statistical analysis of the distribution of peaks in the whole genome, over 50% of STAT1 peaks were located within or close to gene body (up to 2kb)(Table 2), and most of them (90.6%) were found in introns.

# total peaks	# peaks in upstream 2k of genes	# peaks in downstream 2k of genes	# peaks in intron	# peaks in exon	# peaks in intergenic
25,350	582	517	11,028	1,148	12,799

Table 2 The distribution of peaks I found in human genome.

In order to observe the signal against background noise, I analyzed the distribution of peak tags in the 2kb region upstream of the transcription start site (TSS) and the 2kb region downstream of the transcription end site (TES) compared to Input DNA data (Figure 4). I defined these regions as target regions. By scanning target regions, the target peaks which overlapped with target regions were marked. Both ChIP data and Input DNA data then were normalized to 4000bp according to mapped reads which were located in target peak. The average depth was then calculated for each position of target regions. To estimate the significant signal against background (Figure 4), a differential analysis was used as below,

$$d_j = \frac{\hat{d}_{j-ChIP} - \hat{d}_{j-control}}{\hat{d}_{j-control}}$$

Where $\hat{d}_{j-ChIP}$ is the value of ChIP data depth in position j (in red); $\hat{d}_{j-control}$ is the value of Input DNA data depth in position j (in blue); d_j is the differential value between ChIP data and Input DNA in position j (in green).

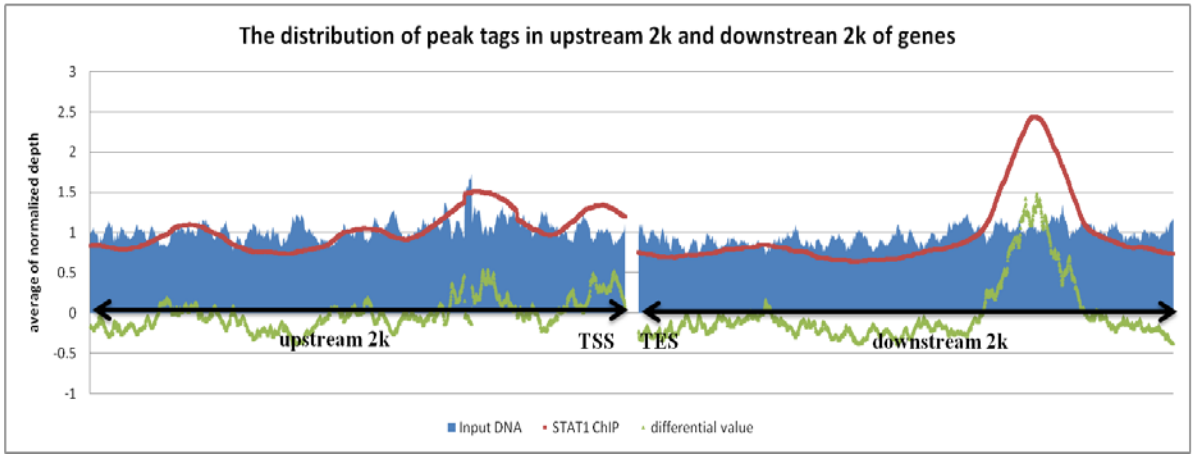


Figure 4 ChIP profile in target regions was calculated by the reads of target peak. The target region was defined as the regions that contained 2k region upstream of the TSS and 2k region downstream of the TES. In target regions, the average alignment depth was calculated for each position. Y-axis represents the average of normalized depth. The distribution of ChIP data is in red; The input DNA data is in blue; The differential value between ChIP data and input DNA is in green.

By uploading my peak results to the University of California-Santa Cruz (UCSC) Genome Browser (Figure 5), the significant enriched regions can be visualized. We can observe that some enriched regions caused by local biases in ChIP sample were removed from the candidate peaks. This filter decreased false positives during peak calling.

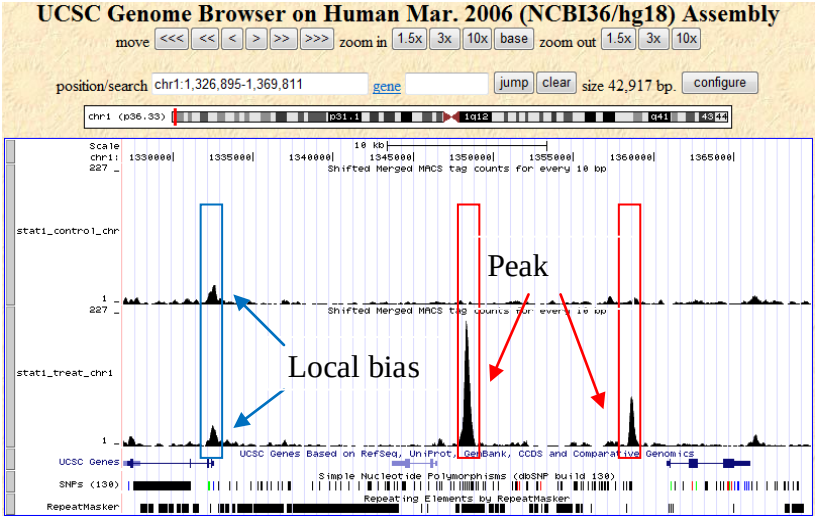
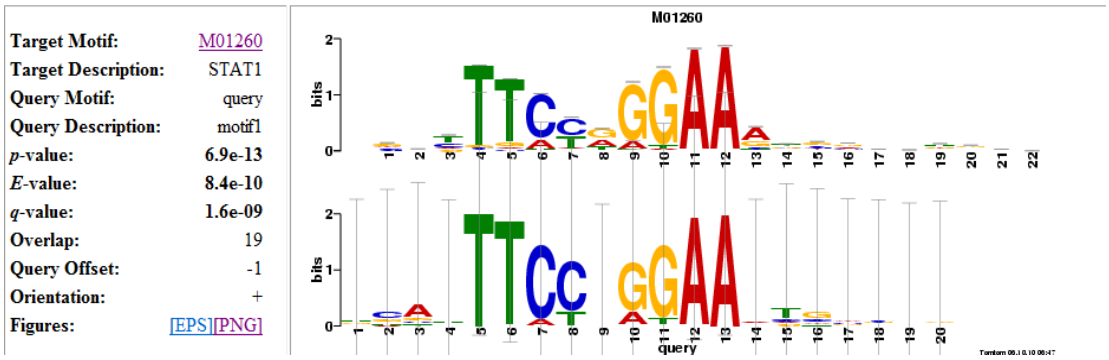


Figure 5 An example of two peaks of STAT1 ChIP-Seq data. The first line data is from control sample, and the second is from ChIP-Seq sample. Compared to the control sample, the region shown with blue box is not a real peak. The enrichment of this region is caused by local biases, whereas the regions shown with red boxes are detected peaks.

As described in 'data and method', 5 motifs were obtained by using MEME (Figure 6). MEME usually finds the most statistically significant (low *E*-value) motifs first. The *E*-value is an estimate of the expected number of motifs that one would find in a similarly sized set of random sequences. The E-value of Motif 1 was far less than other motifs. Therefore, I inferred that Motif 1 maybe the real binding motif of STAT1. For validation, I subsequently compared five motifs I discovered to known motifs contained in TRANSFAC [17] motif databases by using Motif Comparison Tool (TOMTOM). By querying database of know motifs, the results proved that Motif 1 is the known motif of STAT1 (Figure 7).

Motif	E-value	Sequence logo	Sites
Motif 1	2.2e-896		466
Motif 2	2.7e-426		111
Motif 3	6.5e-353		115
Motif 4	8.3e-327		120
Motif 5	2.2e-316		107

Figure 6 The sequence logo of STAT1 binding motifs found through MEME analysis is shown. The sequence LOGO contains stacks of letters at each position in the motif bits. The height of the individual letters in a stack is the probability of the letter at that position multiplied by the total information content of the stack. The *E*-value is an estimate of the expected number of motifs that one would find in a similarly sized set of random sequences. The number of sites is the total number of sites in the training set where a single motif occurs.



Target Motif:	M01260
Target Description:	STAT1
Query Motif:	query
Query Description:	motif1
***p*-value:**	**6.9e-13**
***E*-value:**	**8.4e-10**
***q*-value:**	**1.6e-09**
Overlap:	19
Query Offset:	-1
Orientation:	+
Figures:	[EPS][PNG]

Figure 7 The result of TOMTOM displayed the motif of STAT1 in the TRANSFAC database high similar to the Motif 1 I discovered. Each entry in the table includes the database identifier for matching motif, its description, the p-value of the match, the overlap and offset between my motif and the matching motif, the strand of the matching motif, and a logogram of my motif and the matching motif.

Concerned about the possible roles for STAT1 binding Motif 1, I submitted the probabilities matrix of Motif 1 to Gene Ontology for Motifs (GOMO) [18] and obtained a list of significant GO terms [Figure 8]. The top 5 GO terms are olfactory receptor activity, sensory perception of smell, platelet alpha granule, cytokine activity and response to external stimulus, which are similar with the known gene ontology of STAT1.

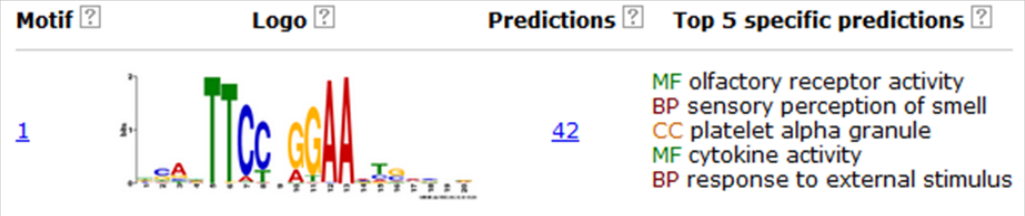


Figure 8 GOMO returns a list of GO-terms that are significantly associated with target genes of the motif, sorted by q-value (minimum false discovery rate).BP stands for biological process, CC stands for cellular component and MF stands for molecular function.

I noticed that the Motif 1 does not match perfectly with STAT1 binding motif in TRANSFAC on the leftmost bases and rightmost bases. As mentioned previously, this can be due to (1) the presence of local bias noise which should be removed; (2) peak calling methods not yet fully developed; therefore, with further analysis, I would like to develop a new method on discovery of DNA motifs, which is to be improved with the following ideas:

1) Keep the sequences which are likely to contain motifs as short as possible. Remove sequences that are unlikely to contain any motifs.

2) Different classes of algorithms have different strengths and weaknesses; therefore it is helpful to run more motif discovery software with different algorithms on the sequence set.

3) Besides extracting the most information from ChIP-Seq data, integrative analysis with RNA-Seq data will be essential. The RNA-Seq transcribed

fragments are identified from RNA-Seq data. The regions can then be analyzed to identify motifs as well as peak regions.

Dataset and Methods

ChIP-Seq dataset

The ChIP-Seq data of STAT1 I used is provided by Rozowsky et al. Nat Biotechnol.2009 [19] and is publicly available (www.camda2009.org). They generated a deeply sequenced ChIP-Seq datasets for antibody against human STAT1 performed in the HeLa S3 cell line. STAT1 is a member of the Signal Transducers and Activators of Transcription family of transcription factors. STAT1 is involved in up-regulating genes due to a signal by either type I or type II interferons.

Mapping sequence reads

For the human ChIP-Seq and control datasets, I aligned the sequence reads against the human genome (version hg18) using MAQ with 2bp mismatches allowed and retained uniquely mapped reads for further analysis. The reference genome obtained from UCSC Genome Browser.

Calling peak

To identify significant STAT1-bound regions in the ChIP data, I used the MACS software with the default parameters. These peaks were reported with p-value and summit of each peak. Since MACS merge overlapping areas of enrichment, the resulting peaks tend to be much larger than the actual binding sites. To enhance the quality of the motif analysis, here I defined the peak sequences as the 200-bp genomic regions centered on the summit of candidate peaks. This is usually a reasonable process for analyzing TFBS, because binding motifs are most likely to appear at or near peak summit regions.

Detecting motif

I used MEME for motif analysis. The input for MEME is a file in FASTA format containing the peak sequences. The MEME algorithm run time is quadratic with respect to the number of characters of input datasets. Due to time limitation for this project, I did not use all peaks sequences as input file. I sorted the sequence peaks according to their ascending p-value. I then chose sequences of top 500 sequence peaks as my input sequences. MEME was parameterized to find 5 motifs of lengths 5-20. Subsequently, I compared the motifs I discovered to known motifs contained in TRANSFAC motif databases by using TOMTOM (Motif Comparison Tool) and identified possible roles for STAT1 binding motif by using GOMO (Gene Ontology for Motifs), a web-server and database provided by MEME suite tool. GOMO used the STAT1 binding motif to find putative target genes and analyzed their associated GO terms. A list of significant GO terms then was reported.

Reference

1. Latchman DS: *Eukaryotic Transcription Factors*. fifth edition. Elsevier Ltd; (2008).

2. Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. *Genome-wide mapping of in vivo protein-DNA interactions*. Science 316, 1497–1502. (2007).

3. Hoffman BG et al. *Genome-wide identification of DNA–protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing*. Journal of Endocrinology 201, 1–13. DOI: 10.1677/JOE-08-0526. (2009).

4. Li, H., J. Ruan, and R. Durbin. *Mapping short DNA sequencing reads and calling variants using mapping quality scores*. Genome Res. 18:1851–1858. (2008).

5. Cox, A. J. *Ultra high throughput alignment of short sequence tags*. Unpublished. (2007).

6. Smith, A. D., Xuan, Z., and Zhang, M. Q. *Using quality scores and longer reads improves accuracy of Solexa read mapping*. BMC Bioinformatics, 9:128. (2008).

7. Ewing, B. and Green, P. *Base-calling of automated sequencer traces using phred. ii. Error probabilities*. Genome Res. 8: 186–194. (1998)

8. Zhang et al. *Model-based Analysis of ChIP-Seq (MACS).* Genome Biology vol. 9 (9) pp. R137. (2008).

9. Sinha, S. and Tompa, M. *Discovery of Novel Transcription Factor Binding Sites by Statistical Overrepresentation*. Nucleic Acids Research, vol. 30, no. 24, December 2002, 5549-5560. (2002).

10. Pesole G, Prunella N, Liuni S, et al. *Wordup: An Efficient Algorithm for Discovering Statistically Significant Patterns in DNA Sequence*. Nucleic Acids Res. 20(11) : 2871- 2875. (1992).

11. MatthieuDefrance, et al. *Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences*. Nature protocols, Vol. 3, No. 10. pp. 1589-1603. (2008).

12. Tompa M: *An exact method for finding short motifs in sequences, with application to the ribosome binding site problem*. Proceedings of the Seventh International Conference on Intelligent Systems on Molecular Biology, 262-271. (1999)

13. Timothy L. Bailey and Charles Elkan.*Fitting a mixture model by expectation maximization to discover motifs in biopolymers*. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California. (1994).

14. Lawrence CE, Reilly AA: *An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences*. Proteins, 7:41-51. (1990).

15. Bailey, T. L, and Elkan, C. *Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization*. Machine Learning Journal, 21, 51-83.(1995).

16. Grundy, W. N., Baily, T. L., and Elkan, C. P.*ParaMEME: A Parallel Implementation and a Web Interface for a DNA and Protein Motif Discovery Tool*. Computer Applications in the Biological Sciences (CABIOS), Vol. 12, pp. 303-310. (1996).

17. Matys V, Fricke E, Geffers R, et al. *TRANSFAC: transcriptional regulation, from patterns to profiles*. Nucleic Acids Res;31:374-8. (2003).

18. Mikael Boden and Timothy L. Bailey, *Associating transcription factor binding site motifs with target Go terms and target genes*. *Nucl. Acids Res*, 36, 4108-4117. (2008).

19. Rozowsky, J. et al. *PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls*. Nat. Biotechnol. 27, 66–75 (2009).