ANALYSIS OF GENOME-WIDE DATA ON TRANSCRIPTION FACTOR BINDING AND THEIR TARGET GENE REGULATION:

A CASE STUDY OF LIVER X RECEPTORS

Petri Pehkonen, Sami Heikkinen and Carsten Carlberg

Department of Biosciences, University of Eastern Finland, FIN-70210 Kuopio, Finland
petri.pehkonen@uef.fi, sami.heikkinen@uef.fi, carsten.carlberg@uef.fi

ABSTRACT

Chromatin immunoprecipitation combined with massive parallel sequencing (ChIP-seq) facilitates a genome-wide detection of transcription factor binding sites. In our study, we have built a pipeline for ChIP-seq data analysis, which is comprised of the recently published programs for sequence read alignment and peak finding and in-house developed data analysis tools supplemented with various visualization schemes for summarization of results. As examples from the pipeline, we present in detail two methods developed for in-house use: 1) detection and comparison of common binding locations across two or more samples and 2) a novel application of segmentation in finding clusters of transcription factor binding locations and regulated target genes. We applied the developed methods to the analysis of ChIP-seq data of the nuclear receptors liver X receptor (LXR) α and β chromatin binding and target gene expression in human macrophage-type cells.

1. INTRODUCTION

ChIP-seq technology facilitates genome-wide detection of transcription factor binding sites, thus providing essential data for understanding the regulation of their target genes. As in regular ChIP, in the ChIP-seq protocol the cells are first treated, cross-linked with formaldehyde and the sonicated chromatin is precipitated with specific antibodies against the transcription factor under investigation. The DNA fragments that are eluted from these precipitates are sequenced in total using sequencers like Illumina Solexa GA II or Applied Biosystems SOLiD. The analysis of raw ChIP-seq data, i.e. sequence reads, often starts with the alignment of the short reads obtained from the sequencer. For this, there are several efficient programs available including Maq [1], Bowtie [2] and BWA [3]. Further analysis includes finding statistically significant pileups of reads, often in terms of comparison against a negative control, such as IgG or input DNA. Subsequently, a variety of procedures will be applied, such as peak comparison across multiple conditions, annotation of peaks with overlapping or neighboring genomic features, integration of binding location information with gene expression data and detection of interesting patterns with various data mining and visualization techniques.

We have built an in-house pipeline of ChIP-seq data analysis tools comprised of read alignment using Bowtie software [2] and peak finding using MACS [4] supplemented with various downstream analysis and visualization techniques implemented in R [5] using partially the Bioconductor libraries [6]. Here, we present the ideas and application of two separate methods developed for the pipeline. First, we present principles of a method for detection and comparison of common binding locations between two or across several samples. This is important as often the impact of the treatment, such as ligand effects on some nuclear receptors, is under investigation. Therefore, often at least a pair-wise comparison between samples is performed. Furthermore, studies including several samples or time points may require to find common and unique binding locations across all samples in parallel. An example case could be a study on transcription factor binding at many points of a time series. Secondly, we show an exploratory approach for studying the changes in transcription factor binding activity across the genome. For this, we present a novel application of dynamic programming based segmentation [7], referred also to as change-point-analysis. This is associated with model selection based on the penalized log-likelihood [8, 9] in finding exclusive clusters of transcription factor binding locations.

Application of ChIP-seq is very informative for the study of the proteins of the nuclear receptor superfamily due to their direct activation by the binding of small lipophilic molecules such as cholesterol. LXRs are oxysterol sensing nuclear receptors that serve as regulators of fatty acid, cholesterol and glucose homeostasis. We applied the developed computational methods to the analysis of ChIP-seq data from LXRA and LXRB protein binding induced by the synthetic ligand T0901317 (T09) in THP-1 human monocytes leukemia cells, which were treated for 3 days with PMA in order to induce differentiation to macrophages (unpublished data). We demonstrate, for example, a novel application of change-point-analysis to the analysis of spatial organization of LXR binding locations. The results indicate genomic regions that contain clusters of LXR binding sites with individual constellations for each region. As an example result, we display chromosome 19 including 11 adjacent regions with different LXR binding location frequencies.
with common peaks between samples. We use the following method steps 1-2 associate all overlapping peak pairs. Steps 3-5 select the representative peak pairs. B: Step 3 of method for comparing three or more samples. Example shows samples $S_1$ with locations $L_1$ and $L_2$, $S_2$ with location $L_2$ and $S_3$ with locations $L_1$ and $L_2$ present.

These could possibly represent larger active and passive regions of the chromatin under a control of transcription factories.

2. COMPARISON OF CHIP-SEQ DATA ACROSS SAMPLES

One of the most important tasks in ChIP-seq data analysis is the comparison of different samples in terms of 1) existence of each binding location across the samples and 2) changes at those binding locations that exist in multiple samples. We developed a method that performs a pair-wise comparison of two ChIP-seq samples to detect overlapping binding locations. Additionally, by using combinatorial pair-wise comparisons, the method detects common binding locations across more than two samples.

2.1. Pair-wise comparison of ChIP-seq peaks

Pair-wise comparison is performed either to compare two biological conditions, such as treated versus untreated, or as a step of multi-sample comparison, such as finding common/different peaks throughout the time series. We use following ad hoc approach to detect such common peaks between samples $S_i$ with $n_i$ peaks and $S_j$ with $n_j$ peaks:

1. Construct a list $U$ of all unique peak locations that do not overlap between $S_i$ and $S_j$ and pair each with itself.
2. Construct a list $L$ of all possible peak pairs $S_{i1} \Rightarrow S_{j2}$ and $S_{2m} \Rightarrow S_{im}$, arrow representing the origin and target samples of comparison that overlap between $S_i$ and $S_j$.
3. Reorder each directed peak pair in $L$ so that the shorter peak is set as the target and broader peak as the origin.
4. Delete duplicates from $L$ and combine $U$ with it.

5. Use the target peaks of combined list as the representative peak locations in each sample.

The method is illustrated in Figure 1 by a few schematic peak locations. The goal has been to associate all overlapping peak pairs so that as a result a non-redundant set of pairs is obtained without losing any information. Selection of the smaller of the two overlapping peak regions as a representative location for both samples focuses broad peak regions or replaces such peaks with multiple sub-peaks detected originally only in one of the samples.

2.2. Comparison across multiple samples

Comparison across three or more multiple samples is needed in experimental setups, such as the analysis of transcription factor binding during a time interval consisting of several sequential samples, comparison across multiple treatments etc. We use the following idea to detect overlapping peaks between $n$ samples $S_i, S_2\ldots S_n$:

1. Apply pair-wise sample comparison, presented in 2.1, for all sample pairs $S_i$ and $S_j$, where $i,j=1,\ldots,n$, to detect list $L$ of all peak pairs.
2. Sort list $L$ according to the length of the source peak.
3. Remove duplicate pairs of the sorted list $L$ by preserving the first pair. This will associate all the smallest overlapping peaks with the larger overlapping peaks.

This results to a list of peak locations with target pointer to the smallest overlapping peak as illustrated in Figure 1B.

2.3. Results

Figure 2 shows the results for comparing T09- and vehicle-treated human macrophages for overlapping LXR binding locations detected using the MACS 1.3.7.1 program [4]. Indicated are the proportion of common and
unique peak locations between the samples from total number of peaks (A) and the proportion of unique binding locations per sample from all unique binding locations (B) with changing FDR cutoff. Lower and upper saturation points for the number of remaining peaks, FDR = 0.5% and FDR = 9%, respectively, are indicated in the figure. The peak numbers are invariant with the more extreme FDR-cutoffs. This is due to sample-swap based FDR-calculation procedure used in MACS, which, for example, gives zero FDR values for the all peaks that have their raw P-value smaller than any of the peaks in the negative control. Results shown in panel A indicate clearly that with the stronger peaks the proportion of common peaks is the highest, whereas when loosening the FDR-threshold the proportion of common peaks starts to decrease. This is partially explained by the added noise but also indicates that the strongest peaks exist in both samples. As shown in panel B, the increase of unique peaks is greater in the vehicle than in the T09-treated sample when loosening the FDR cutoff. This may indicate that in T09-treated sample the LXR binding is more focused than in vehicle-treated sample.

The running time of the comparison depends on the used sorting algorithm and the method for detecting duplicated elements in an ordered list.

3. SEGMENTATION FOR FINDING CLUSTERS OF BINDING LOCATIONS

Some transcription factor binding sites are found rather close to the transcription start site (TSS) of their target genes, while others are found in distal enhancer regions from where they contact the TSS via looping. It is assumed that many, if not all transcription factors, chromatin modifiers and co-regulators are combined in some 100-200 transcription factories per nucleus that control the structure of chromatin on a large scale [10]. In order to study the distribution of binding locations, we present novel method/application for finding clusters of transcription factor binding sites.

3.1. Method

For the analysis of spatial clusters of LXR binding locations we first perform a density analysis of genomic coordinates of peak summit locations. This is made using standard R function “density”, for example, with a 1 Mb size for the sliding window with 0.5 Mb steps over each chromosome using the default Gaussian window kernel function. The density values resulting from each of these windows were weighted using fold enrichment values for ChIP-Seq binding locations to allow focusing on stronger peaks.

In order to detect the exact borders of the hotspot regions, we perform a subsequent clustering of the density data representing the binding locations by using methods developed originally for the analysis of array CGH data implemented in R package SegClust [8, 9]. Firstly, the density data is used for segmentation implemented by the “segmean” function in SegClust. The aim of the segmentation is to find borders (i.e. breakpoints or change-points) between homogeneous local regions in a data stream or sequence. The locations of the change-points and their optimal number are unknown and are detected from the data. “Segmean” performs a dynamic programming (DP) based segmentation algorithm for finding the optimal change-points in terms of changes of mean for a fixed number of change-points K [8, 9]. DP treats a partition of data into m + 1 segments as a union of m and one segments and can therefore reuse the scores calculated in previous steps. Thus, the method is only suitable in the case of additive or multiplicative score functions, and when presuming the statistical independence between the segments. In our application we consider the borders between the chromosomal clusters of
LXR binding locations (represented by the density data) as the change-points and aim to estimate their locations and number. Despite some dependence is allowed in data by using partially overlapping density windows, we assume its impact minimal due to equal sizes of windows and overlapping regions across the genome.

The segmentation algorithm is applied using various $K$ starting from 1 ending to the number of analyzed data points. The results are further used for the “segselect” function, which uses an adaptive model selection [8, 9] between outcomes with different numbers of change-points $K$. The score function used by the model selection is of form $\hat{L}_a = \hat{L}_a - \beta \times 2K$ where $\hat{L}_a$ is the maximum log-likelihood which can be considered as the quality measure of data to fit to the model represented by the segmentation solution with $K$ change-points. However, as such, it does not take into account the number of parameters represented by the $K$ change-points. Therefore, a penalized version of likelihood is commonly used to obtain a trade-off between the fit and number of parameters. The penalty term $2K$, similar as used with Akaike Information Criterion [11], increases with the number of segments. In addition, the term $\beta$ is a constant derived adaptively from the data according to the procedure documented in [8, 9]. This term is small for the change-point numbers, around which the $\hat{L}_a$ is stable, and thus favors such solutions.

A result optimal in terms of this model selection is chosen as representative for each chromosome. The resulting chromosomal regions can be further selected for example according to numbers of binding locations etc.

3.2. Results

The genome-wide segmentation of the 1357 LXR locations with FDR < 1% from MACS indicated genomic hotspots for the actions of LXR. Figure 3 exemplifies the results of the scoring function, the negative penalized log-likelihood, for chromosomes 18 and 19. For both the optimal number of break points is detected as 10 (resulting to 11 distinct segments). Figure 4 shows an example segmentation of density data for LXR binding locations (in either T09 or vehicle-treated sample) for chromosome 19. Interestingly, the figure shows that in addition to separation between completely active and passive regions, the method can also distinguish between genomic areas with differing levels of binding activity.

4. DISCUSSION AND CONCLUSION

We have demonstrated two methods that were developed as a part of our ChIP-seq data analysis pipeline, and their application to the ChIP-seq data from LXR-binding in human macrophage-type cells. The comparison of binding locations is important for detecting the reappearance and vanishing of binding locations due to treatment or experimental condition, and finer induction or repression. The presented ideas related to this are further developed in our group. Our analysis of LXR ChIP-seq data showed that with stronger peaks the overlap between the vehicle and T09-treated samples was higher.

This is not surprising, as with weaker peaks more noise and less real binding locations will be observed.

The second presented application was motivated by the ideas presented before about the transcription factories as the regulators of larger structural parts of chromatin (see for example [10]). Due to previous application in bioinformatics [8, 9], we decided to use segmentation for detecting putative active and passive areas of the genome concerning LXR binding. It should be noted that the similar ideas could be applied in future to analysis of other genomic features and their combinations. Our analysis of these regions reveals hotspots of LXR binding in several chromosomes. As an example we show the chromosome 19 that is known to have the highest gene density of all human chromosome, high G+C content, large amount of CpG-islands and large clustered gene families [12]. Indeed, out results show that LXR binding varies considerably from region to region within chromosome 19. However, further analysis would be needed to show the role of transcription factories in it.

5. REFERENCES